Genomic heterogeneity in advanced colorectal cancer

Timothy George Palmer

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

University of Leeds

School of Medicine

September 2019

Intellectual property and publication statements

The candidate confirms that the work submitted is his own, except for the contributions explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The candidate contributed to study conception and design in collaboration with Professor Philip Quirke (PQ) and Dr Henry Wood (HW). The 'Gift' autopsy project was administered by Mr A Hindley (AH), AH also obtained consent from all donors and their next of kin for involvement in the project. Dr Daniel Swinson assisted in poster design and patient counselling for volunteers to the 'Gift' project from oncology outpatients at St James University Hospital (SJUH), Leeds. Dr E Verghese (EV) and PQ assisted with post mortem sampling in one of the 'Gift' cases (GD2). Immunohistochemical staining for mismatch repair antigens was performed by Miss Gemma Hemmings (GH). Dr Philip Chambers (PC) performed the pyrosequencing assay as described in section 4. Miss Morag Taylor performed the NGS library preparation in the first of the 'Gift' autopsy cases described in section 5. The Next Generation Sequencing (NGS) described in section 5 and 7 was performed by the NGS facility at the Clinical Sciences Building at SJUH (team lead, Dr Sally Fairweather). Alignment of sequencing data and bioinformatic analysis in section 5 and 7 was performed with the assistance of HW. All library preparation and sequencing described in section 6 was performed at the Wellcome Trust Sanger Institute under the supervision of Dr Ultan McDermott (Cancer Genome Project group). All other autopsy examination, clinical data compilation, tissue sectioning and staining, DNA extraction, library preparation, bioinformatic analysis and figure design was performed by the candidate.

Acknowledgements

Firstly, I would like to thank my funders, Cancer Research UK, I am enormously grateful for the opportunity presented to me as part of my PhD Fellowship.

My deepest gratitude is expressed to the ‘Gift’ research autopsy project donors and their families, without their generosity and empathy this work would not have been possible. Mr Aidan Hindley, as the administrator and autopsy technician for the project, was also central to this work. Dr Dan Swinson, who counselled patients and their families regarding the ‘Gift’ project, also deserves heartfelt thanks.

Enormous thanks are also extended to my Supervisors who have been incredibly supportive and patient throughout my PhD, particularly Professor Phil Quirke, who has been a huge source of inspiration and motivation, and Dr Henry Wood, without whom I would still be staring blankly at a computer screen trying to align sequencing data. Further thanks are given to all my colleagues in the Wellcome Trust Brenner Building (especially Miss Morag Taylor, Miss Gemma Hemmings and Dr Phil Chambers) who helped me on the steep learning curve during my time in the laboratory. Dr Ultan McDermott, Dr Patrick Tarpey and the Cancer Genome Project group were of immeasurable assistance during my time at the Wellcome Trust Sanger Institute, I thank them for their time and insight.

Finally, I would like to thank my partner, Anna, and my children, Alder and Etta, for their love and support.

Abstract

Systemic chemotherapy is the primary treatment modality for metastatic colorectal carcinoma (MCRC). However, both conventional and novel chemotherapeutic drugs produce only modest improvements in outcome despite the introduction of predictive biomarkers. It is postulated that genomic heterogeneity within MCRC may be responsible for poor therapeutic response, but a comprehensive analysis of disseminated disease has not been performed.

This thesis provides in-depth analysis of the clinicopathological and genomic features within a cohort of fatal MCRC cases recruited via the 'Gift' research autopsy project. Material sampled at autopsy was analysed at the allelic and chromosomal level using targeted and whole genome sequencing (WGS); the pattern of genomic change within and between deposits was correlated with that of any resected disease and the clinical data for each donor. The insights from this cohort were further explored within a series of locally advanced CRC using micro-dissection of intravascular, intraperitoneal and intranodal tumour deposits.

Initial analysis demonstrated that mutation status at therapeutically predictive genomic loci is virtually homogenous within cases of disseminated MCRC. Phylogenetic analysis of WGS data documented the evolutionary and clonal complexity within MCRC, showing that, whilst clonal distribution may correlate with disease distribution, multiple clones from the same primary tumour may converge on the same metastatic site and demonstrate dissemination via similar metastatic routes. It was also demonstrated that most key genomic events arise early within the development of MCRC and when putative 'driver' events occur within established disease they do not appear to produce dominant metastatic clones. Therefore, it is concluded that, if a tumour has the capacity to metastasise, this characteristic is present across many or all subclones within a tumour and the likely determinant of the pattern of metastatic spread is a combination of the core biology of a tumour and the regional features of the surrounding bowel.

Table of contents

List of tables

List of figures

List of abbreviations

ADPC	androgen deprived prostatic carcinoma
AJCC	American Joint Committee on Cancer
AKT	protein kinase-B
AMP	adenosine monophosphate
APC	adenomatous polyposis coli
ATP	adenosine triphosphate
BCL	B-cell lymphoma
BMPR	bone morphogenic protein receptor
bp	base pair
BRCA1	breast cancer 1, early onset
CGH	comparative genomic hybridization
CIMP	cytosine-phosphate-guanine island methylator phenotype
CIN	chromosomal instability
CMS	consensus molecular subtype
CNA	copy number alteration
CpG	cytosine-phosphate-guanine
CRC	colorectal carcinoma
CRM	circumferential margin of resection
CTDNA	circulating tumour DNA
CTFR	cystic fibrosis transmembrane conductance regulator gene
DCC	deleted in colorectal carcinoma

DFS	disease-free survival
DGGE	denaturing gradient gel electrophoresis
DNA	deoxyribonucleic acid
DPX	Dibutylphthalate Polystyrene Xylene
dsDNA	double stranded deoxyribonucleic acid
E2F-DP	E2 promoter-binding–protein-dimerization partner
EGFR	epithelial growth factor receptor
EMVI	extramural venous invasion
ER	oestrogen receptor
ERBB2	Erb-B2 Receptor Tyrosine Kinase 2
FFPE	formalin fixed paraffin embedded
FOLFIRI	folinic acid, fluorouracil and irinotecan
FOLFOX	folinic acid, fluorouracil and oxaliplatin
FOLFOXIRI	folinic acid, fluorouracil, oxaliplatin and irinotecan
GD	‘Gift’ donor
GDP	guanosine-diphosphate
GEF	guanidine transfer factors
GTP	guanosine-triphosphate
H and E	haematoxylin and eosin
HCC	hepatocellular carcinoma
HRAS	Harvey rat sarcoma viral oncogene homolog
HRMA	high-resolution melting analysis

IGFR2	insulin-like growth factor receptor 2
IHC	immunohistochemistry
IMVI	intramural vascular invasion
IP	intraperitoneal
IV	intravascular
kb	kilobases
KRAS	Kirsten rat sarcoma viral oncogene homolog
LCMD	laser capture microdissection
LS	Lynch syndrome
MAPK	mitogen-activated protein kinase
MCRC	metastatic colorectal carcinoma
MDM2	mouse double minute 2 homolog
Min	minute
MLH	MutL-homolog
MMR	mismatch repair
MP	muscularis propria
MSH	MutS-protein homolog
MSI	microsatellite instability
MSS	microsatellite stable
MTOR	mammalian target of rapamycin
MYC	avian myelocytomatosis viral oncogene homolog
ng	nanograms

NGS	next generation sequencing
NICE	National Institute for Health and Clinical Excellence
NRAS	neuroblastoma rat sarcoma viral oncogene homolog
OC	oesophageal carcinoma
OS	overall survival
p	short arm of chromosome
PC	pancreatic carcinoma
PCR	polymerase chain reaction
PD	programmed cell death protein
PDL-1	programmed death-ligand 1
PEG	polyethylene glycol
PEN	Polyethylene Naphthalate
PFS	progression free survival
PI3K	phosphatidylinositol 3-kinase
PIP2	phosphatidylinositol 4, 5-bisphosphate
PM	post mortem
PMS	post meiotic segregation increased
PNI	perineural invasion
POLD	polymerase delta
POLE	polymerase epsilon
PPi	inorganic pyrophosphate
q	long arm of chromosome

QC	quality control
RAF	RAF proto-oncogene serine/threonine-protein kinase
RAS	rat sarcoma
RCC	renal cell carcinoma
RCT	randomised control trial
RIF	right iliac fossa
SJUH	St James University Hospital
SMAD	small mothers against decapentaplegic homolog
SNP	single nucleotide polymorphism
SSCP	single-strand conformation polymorphism
ssDNA	single stranded deoxyribonucleic acid
TCGA	The Cancer Genome Atlas Network
TGF-B	transforming growth factor-beta
TGFBR	transforming growth factor-beta receptor
TNM	tumour node metastasis classification
TNM5	tumour node metastasis classification 5th edition
TNM8	tumour node metastasis classification 8th edition
UK	United Kingdom
µl	microlitres
VAF	variant allele frequency
VEGF	vascular endothelial growth factor
vHL	von Hippel Lindau

VI	vascular invasion
Wnt	Wingless-int

1 Introduction

Colorectal carcinoma (CRC) is the third most common cancer worldwide (excluding non-melanocytic skin cancer), causing over 880,000 deaths in 2018 [1]. CRC is most prevalent in developed nations (due to low fibre diet, sedentary lifestyle and high alcohol intake), in the United Kingdom (UK) CRC is the second most common cause of cancer related death after lung cancer [2].

Spread of CRC from the bowel to a distant site (metastasis) often represents progression to incurable disease; the most common site of distant spread in colorectal cancer is the liver [3], other common sites of metastatic spread include the peritoneum and lungs. This dissemination of malignant cells may occur directly across the peritoneal cavity or via the portal or lymphatic circulation.

1.1 Genomic Characteristics of Primary CRC

The molecular events and pathways important to the development of CRC are well described; these are mutations or chromosomal aberrations affecting adenomatous polyposis coli (*APC*), rat sarcoma (*RAS*) and *TP53* genes plus deletions of the long arm of chromosome 18 (18q). A model, proposed by Fearon and Vogelstein [4], integrated the most commonly observed key genomic events with the phenotypic stages in the classical 'adenoma-carcinoma sequence' is shown in Figure 1.

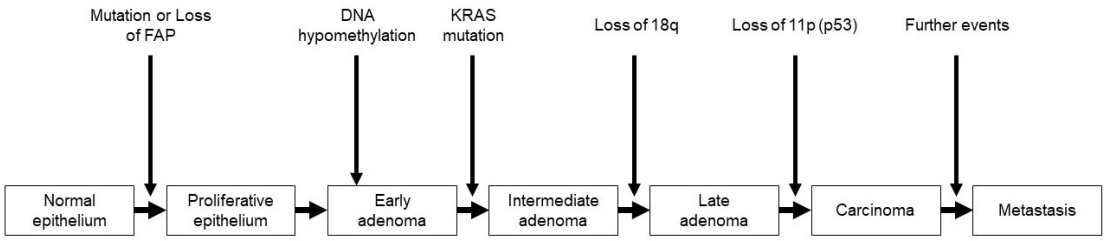


Figure 1: Adenoma-carcinoma sequence (adapted from Fearon and Vogelstein [4])

1.1.1 APC and Wingless-int (Wnt) pathway

Abnormalities of *APC* were first identified in patients with familial adenomatous polyposis (FAP) [5] and subsequently within 80% of sporadic adenomas and carcinomas [6]. *APC* has a tumour suppressor role in the Wnt signalling pathway by binding β catenin; if unbound, β catenin moves into the nucleus and initiates cell division.

Genomic aberrations, which prevent β catenin binding in *APC*, are commonly identified in adenomas and are therefore felt to be an early event. Further mutations in rat sarcoma/mitogen-activated protein kinase (RAS-MAPK) pathway, the phosphatidylinositol 3-kinase/protein kinase-B/mammalian target of rapamycin (PI3K/AKT/mTOR Pathway) and loss of p53 function are observed during transition from adenoma to carcinoma.

1.1.2 RAS-MAPK Pathway

Epithelial growth factor receptor (EGFR) is one of a family of four similar trans-membrane glycoproteins which have intracytoplasmic tyrosine kinase activity, when bound by an appropriate ligand, undergoes dimerisation with internalisation of the binding portion of the receptor. The intracytoplasmic portion of the receptor then activates growth-factor receptor bound proteins which in-turn recruit guanidine transfer factors (GEF). These enzymes catalyse the exchange of guanosine-diphosphate (GDP) to guanosine-triphosphate (GTP) on the surface of Kirsten rat sarcoma viral oncogene homolog (KRAS); this conformational change causes activation of KRAS.

RAS (including KRAS, neuroblastoma RAS (NRAS) and harvey RAS (HRAS)) molecules are GTPase molecules attached to a G protein which act as switch to various down-stream processes; in this instance the activation of KRAS allows binding to rapidly associated RAF proto-oncogene serine/threonine-protein kinases (RAF), this complex propagates various processes including proliferation, cell survival, angiogenesis and differentiation (Figure 2).

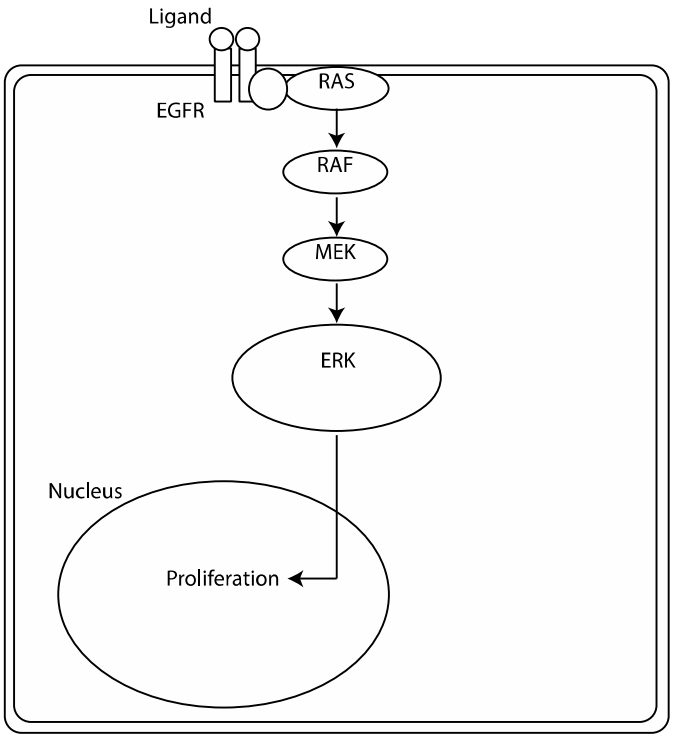


Figure 2: RAS/MAPK Pathway (adapted from Fernandes et al [7])

Within this pathway the most well characterised mutations of prognostic and therapeutic significance are of *KRAS*, *NRAS* and *BRAF*.

Constitutional activation of RAS-MAPK pathway due to *KRAS* mutations at codon 12, 13, 61, 117 and 146 and *NRAS* at codon 12, 13 and 61 are present in approximately 40% of CRC [8]. *RAS* mutations have also been identified in non-neoplastic colonic epithelium, suggesting other events are required for carcinogenesis [9].

Mutation of *BRAF* is an alternate event which may lead to increased activation of the RAS-MAPK pathway. Genomic lesions leading to the constitutional activation of this protein are most commonly observed at codon 600 of the gene and have been identified in approximately 15% of CRC [10]. They have been associated with microsatellite instability (MSI) and poor prognosis in advanced CRC [11].

1.1.3 PI3K/AKT/mTOR Pathway

This is a second proliferative pathway found to constitutionally activated in CRC. Activation of this pathway is initiated via EGFR/RAS signalling, which in turn produces activated PI3K. This molecule then activates AKT, via phosphorylation of phosphatidylinositol 4, 5-bisphosphate (PIP2), which inhibits pro-apoptotic molecules (including B-cell lymphoma-1 (BCL1) and mouse double minute 2 homolog genes (MDM2)) and drives protein synthesis and proliferation via mTOR activation as shown in Figure 3

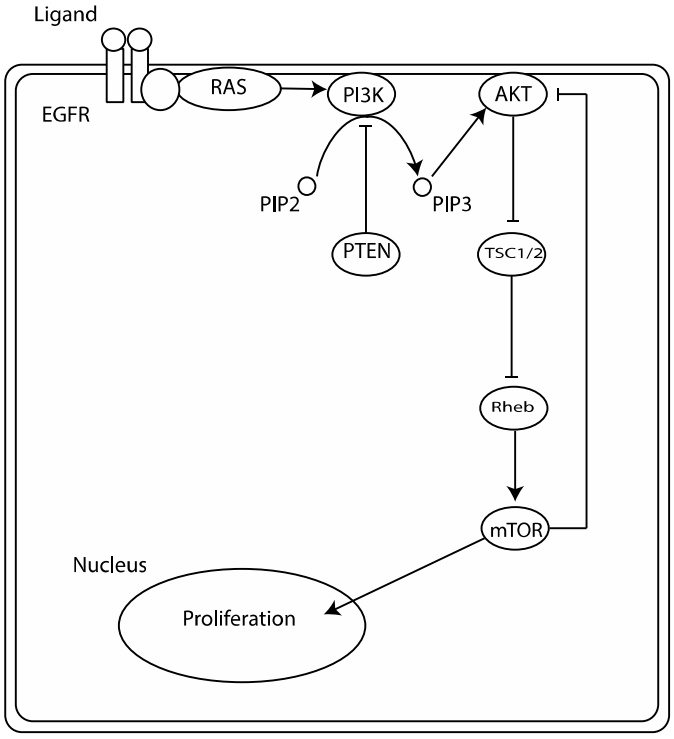


Figure 3: PI3K/AKT/mTOR Pathway (adapted from Janku 2013[12]).

The major regulatory mechanisms in this pathway are phosphatase and tensin homolog (PTEN), which inhibits the phosphorylation of PIP2, and negative feedback of AKT by downstream factors including mTOR [13].

Mutations within this pathway have been identified at several stages, the most common are *PI3K* mutations, present up to 34% of CRC and have been identified almost solely within the catalytic subunit of one isoform of protein PIK3CA [14]. Mutations within this protein either cause constitutional activation or resistance to feedback inhibition by

mTOR [15]. Copy number gains of alternate PI3K activating pathways are also commonly identified in colorectal neoplasia [16].

Genetic and epigenetic aberrations within PTEN itself have also been identified in 10% of CRC [17–19]. The loss of function is deemed to be important to tumourigenesis as it leads to activation of the PIK/AKT/mTOR pathway due loss of the response to inhibitory feedback mechanisms.

Finally, point mutation in the lipid binding pocket of *AKT* has also been identified in CRC; this affects lipid binding so the protein is constitutionally bound to the membrane and active [20]. Up-regulation of AKT activity has been observed in 40% of CRC [18].

1.1.4 TP53

This gene (located on 17p) is an important tumour suppressor as it causes cell-cycle arrest and initiates apoptosis in response to DNA damage and degradation. Mutations of *TP53* are identified in up to half of CRC and felt to occur during the transition from adenoma to carcinoma [4].

The *TP53* gene is transcribed to p53 [21], this DNA-binding protein within normal cells is inhibited by a negative regulator, MDM2. MDM2 produces this inhibitory effect by facilitating p53 degradation [22]. DNA damage, DNA degradation or cell cycle dysfunction induces stabilising phosphorylation of p53. Phosphorylated p53 induces apoptosis via members of the B-cell lymphoma-2 (BCL2) family, which trigger the 'intrinsic' apoptotic pathway, and also stimulates upregulation of 'death receptor' molecules, such as the Fas cell surface death receptor, activating the 'extrinsic' apoptotic pathway [23,24]. Along with this pro-apoptotic role, p53 also causes cell-cycle arrest at both the G1/S and G2/M phase through inhibition of cyclin dependent kinases via p21 and through interaction with E2 promoter-binding–protein-dimerization partner (E2F-DP) dimers which are key to the progression of the cell cycle [25–27].

Therefore, as p53/*TP53* holds such important interactions within the cell cycle, dysfunction of this molecule is seen in a wide variety of malignant neoplasms across the body. [28,29] The frequency and location of *TP53* mutations tends to vary slightly between different anatomical sites and neoplasms, possibly reflecting different modes of carcinogenesis [30], the frequency of *TP53* mutation is higher within distal rather than proximal CRC. In CRC the most common *TP53* mutations are seen in exons 5 to 8 [31]

within portions of the molecule involved with the DNA binding. *TP53* mutation status does not hold prognostic significance although there is the suggestion that mutation position, type and tumour site and stage may provide some prognostic information [31–35].

1.1.5 Deletion of 18q

Loss of a portion of chromosome 18q is observed in 60% of colonic carcinomas; this region of the genome contains three significant genes in the context of CRC carcinogenesis, small mothers against decapentaplegic homolog 2 and 4 (*SMAD2* and *SMAD4*) and deleted in colorectal carcinoma (*DCC*) [36–38].

The SMAD family proteins comprise 8 molecules which are key to signaling within a number of transforming growth factor-beta (TGF-B) superfamily pathways [39–41]. These pathways involve several transmembrane receptors including TGF-B receptor 1 and 2 (TGFBR1 and TGFBR2), bone morphogenic protein receptors (BMPRs). Within the context of neoplasia TGF-B signaling has both tumour suppressor and oncogenic properties [42], however the pathway involving SMAD molecules primarily has a tumour suppressor role. Dysfunction in TGF-B signaling is observed commonly in CRC [43–45] and genetic alterations within TGF-B signaling genes, *SMAD4* or *BMPR1A*, are identified within the majority of individuals with juvenile polyposis syndrome [46,47]. Mutation of *TGFBR2* is also seen in MSI CRC [16].

The ligand bound TGFB superfamily receptors instigate their inhibitory effect of cell proliferation by phosphorylation of either SMAD1, 2, 3, 5 or 8 (the receptor-regulated SMAD functional subgroup). Once phosphorylated these molecules all then bind to SMAD4 which, via complex interaction with pro and anti-proliferative genes, inhibit the cell cycle [39].

Although both SMAD4 and SMAD2 are present on 18q work examining the frequency and timing of genomic events involving these molecules suggests that SMAD4 may be more pivotal in the progression of CRC in the context of 18q loss [48]. This suggestion is logically sound as SMAD2 is only one of several molecules performing a similar role, whereas SMAD4 has a more unique and integral function within TGF-B signaling.

DCC is a tumour suppressor gene that is transcribed to a transmembrane receptor which, when unbound by a ligand (netrin-1), stimulates apoptosis via the intrinsic apoptotic pathway [36]. Loss of *DCC* therefore results in loss of a pro-apoptotic cell

pathway and is characteristically seen in advanced CRC following loss of other tumour suppressor genes such as *TP53* [4].

As both *SMAD4* and *DCC* are often both lost in 18q deletion it is difficult to ascertain the prognostic significance of either genomic event individually, however it has been shown that 18q loss does not hold prognostic significance within microsatellite stable CRC [49].

1.2 The Molecular Classification of CRC

Despite considerable understanding of the molecular pathways involved in tumourigenesis, it has not been possible to classify CRC solely by the presence of individual genomic aberrations. It is, however, recognized that certain groups of tumours show genomic features, which correlate with phenotype and prognosis; these features are the presence of 'chromosomal instability' (CIN), MSI and hyper-methylation. Within these three groups the presence of aberrations in key driver genes has additional prognostic value [50].

1.2.1 CIN CRC

Larger scale changes within the genome characterise almost half of all CRC[51,52]; these changes are described as chromosomal instability and include alterations in the number of chromosomes (changes in 'ploidy'), or loss or gain of large portions or whole arms of chromosomes. A range of mechanisms underlie these largescale changes in the tumour genome including abnormalities of chromosomal cohesion, dysregulation of cell cycle (particularly premature initiation of anaphase), the size and number of centrosomes and microtubule dysfunction (either due to defects in microtubule assembly or microtubule-centromere adhesion).

The presence of CIN and the degree to which CRC display CIN has been shown to be related to patient outcome. Several trials in patients with stage II and III disease have showed that individuals with CIN CRC rather than MSI tumours have poorer disease-free survival (DFS), additionally the degree of CIN (as assessed by image cytometry) was associated with poorer DFS [53,54]. It is also been demonstrated, however, that oestrogen receptor (ER) negative breast tumours with a very high degree of CIN have a better prognosis than those with a more intermediate level of aneuploidy [55] suggesting that there may be a CIN threshold above which a tumour cell population will not survive. This work in breast carcinoma suggested however that this threshold may significantly differ between individuals and between tumour subtypes, therefore stratification of patients for treatment by CIN indices may be problematic.

Due to the prognostic importance of CIN [56] and the possible impact of CIN on chemotherapy response [57,58], two opposing therapeutic approaches are exploited by current drugs or are the focus of future development. These are CIN-reducing (aiming to reduce the acquisition of new chromosomal abnormalities and tumoural heterogeneity)

and CIN-inducing (which aim to drive tumour cells beyond a theorised CIN threshold above which tumour cells cannot survive) [59]. Currently several conventional chemotherapeutic agents are thought to derive their therapeutic effect through CIN-induction (taxanes and vinca alkaloids [60]) and a wide range of molecules are currently being trialled which may alter the rate of CIN or augment the effects of established CIN-inducing therapies [59].

It is now thought that the majority of CIN CRC are of a conventional, 'canonical type' with aberrations in the Wnt pathway and a minority (approximately 15%) show a 'mesenchymal' phenotype with up-regulation of TGF-B; this rarer subgroup displays poorer rates of response to chemotherapy and lower overall survival [52]. The observed poor outcome in patients with upregulation of TGF-B is illustrative of the oncogenic aspects of TGF-B signaling, which are at apparent odds with the apoptotic role played by the SMAD mediated signaling pathway described previously [42].

1.2.2 MSI CRC

MSI colorectal tumours are a subset (15%) of CRC characterised by a failure in mismatch repair (MMR) due to mutation or epigenetic silencing by methylation of MutL-homolog1 (*MLH1*), post meiotic segregation increased 2 (*PMS2*), MutS-protein homolog 2 (*MSH2*) and *MSH6* genes. They may be hereditary ('Lynch syndrome' (LS)) or sporadic and have specific genomic and histological features, which are laid out in the Bethesda guidelines [61].

Normally functioning MLH1 and PMS2 and MSH2 and MSH6 form dimers which isolate and repair single nucleotide errors in DNA replication, so called MMR [62]. Defective or deficient MMR leads to accumulation of mutations, which result in neoplasia. The hereditary form of this tumour type (LS) is seen as part of a congenital syndrome associated with multiple tumours including endometrial, adnexal skin and urothelial carcinomas as well as CRC. In 2017 the National Institute for Health and Clinical Excellence (NICE) recommended that all individuals with CRC should be tested for MSI at diagnosis to identify individuals with LS [63], this testing may also aid treatment stratification [64]. This testing should comprise identification of MSI tumours either by immunohistochemistry (IHC) for the 4 main MMR molecules or by direct MSI testing via polymerase chain reaction (PCR) to demonstrate characteristic DNA replication errors at vulnerable sites in the genome. Tumours found to exhibit MSI are then subject to

further testing to establish the presence of germline MMR gene defects, *BRAF* mutation is particularly characteristic of sporadic MSI tumours and rarely seen in LS patients.

MSI tumours have a 'hypermutator' genotype showing a higher rate of somatic mutation and a lower rate of CIN than microsatellite stable (MSS) tumours. Both the 'hypermutator' MSI and CIN type tumours show genomic abnormalities within same oncogenic pathways although the pattern of specific mutations varies between the two; for example within the RAS-MAPK pathway CIN CRC show a higher frequency of *KRAS* mutation whilst MSI tumours tend to show *BRAF* mutations [65].

Morphologically MSI tumours are typically right sided and are phenotypically characterised by high histological grade and a dense lymphocytic inflammatory response to the tumour [50]. This inflammatory response is thought to arise as a reaction to new antigens within a tumour occurring due to translation of mutated somatic DNA into new foreign epitopes. This brisk immune response is exploited by novel immunomodulatory therapeutic options such as programmed cell death protein -1 (PD-1) blockade [64].

A further, more recently described, subgroup of tumours also displaying a 'hypermutator' genotype are those with mutations in the proof reading regions of the genes encoding polymerase delta 1 (POLD1) and polymerase epsilon catalytic subunit (POLE) proteins [66]. These lesions are less well characterised than those seen in conventional MSI tumours and have been described in a group of phenotypically diverse familial CRC, as some cases resemble a familial polyposis syndrome whilst others resemble LS.

1.2.3 Cytosine-phosphate-guanine island methylator phenotype (CIMP) CRC

Within the human genome regions rich in guanine-cytosine and cytosine-phosphate-guanine (or CpG islands) are often identified as being concerned with the initiation of gene transcription [67], particularly 'house-keeping' genes expressed within normal tissue such as the MMR genes. Methylation of these regions leads to silencing of these essential cellular functions and, as described previously, it is through this epigenetic silencing that the majority MSI CRC arise. It is, however, increasingly recognised that CpG island methylation is present in a group of tumours more diverse than the typical BRAF mutant, poorly differentiated CRC [68]. Work examining the spectrum of CpG island methylation in relation to common driver mutations suggests that although *BRAF* mutation strongly correlates with high level of CpG methylation, a group of *KRAS* mutant

tumours may also show a degree of CpG methylation [52]. This second group of tumours is more varied displaying a degree of MSI and CIN, and it is postulated that this apparently disparate tumour subtype is characterised by the expression of proteins expressed in digestive pathways similar to a subgroup of gastric tumours [69] and confers a similar OS rate to the BRAF mutant CIMP tumours without the poor survival following recurrence in the latter group.

Due to the overlap in genomic and prognostic features within the tumour subtypes described above, several studies have been published using gene expression profiling to refine the classification of CRC [70–73]. In 2015, the most comprehensive study using this approach produced a 'consensus classification' [52] which purported to unify previous gene expression data and separated CRC into four consensus molecular subtypes (CMS). This classification broadly also correlated with the major genomic, epigenetic and phenotypic features described above and with whole genome data published by the Cancer Genome Atlas Network (TCGA) [65]. This classification includes two subgroups showing CIN (one conventional 'canonical' type (CMS 2) and the other 'mesenchymal' type conferring poorer prognosis (CMS4)), one hyper-mutator type (showing MSI, CpG methylation and *BRAF* mutation (CMS1)) and a 'metabolic' type (which contains a mix of MSI and CIN CRC and is associated with *KRAS* mutation (CMS3)). However, whilst the 'consensus classification' system has shown be predictive of treatment response in some analyses [74], it still requires validation within the setting of large prospective clinical trials [75]. Additionally, within the UK, wider availability of gene expression, genome-wide and epigenetic testing would be required to make such a classification system clinically applicable.

1.3 Genomic Characteristics of Metastatic CRC

In contrast to the changes present within primary carcinoma, the evolution of metastatic CRC (MCRC) is less well characterised; an understanding of this process and the heterogeneity between primary tumours and their metastases (intertumoural heterogeneity) as well as intratumoural heterogeneity is essential for effective treatment.

1.3.1 Mutational heterogeneity in CRC

Comparison of mutational status in primary tumours and resected metastases has produced conflicting evidence as to the degree and nature of inter and intratumoural heterogeneity.

Early work focusing upon therapeutically significant mutations, related to the use of EGFR blockade, suggested minimal variation between primary tumours and metastases. Multiple studies found concordance of >90% in *KRAS* codon 12/13 mutation status [76–78]. However further evidence has accumulated as to the influence of other genomic loci in the treatment of CRC [8] and it is increasingly recognised that a broader mutational panel is required to predict response to treatment and biological behaviour of tumours.

The introduction of massively parallel or next generation sequencing (NGS) and array technologies has provided opportunity to interrogate the entire exome or broad mutational spectrum with great sensitivity. The initial work published through the implementation of this technology was in stark contrast the previously cited data; two studies [79,80] showed a much higher degree of discordance between matched CRC primaries and metastases than would have been expected based upon data generated by pyrosequencing. Lee *et al* [79] performed whole exome sequencing of 15 matched primary CRC and liver metastasis showing marked discordance in mutational status; 47% of tumours sampled showed no shared lesions in key CRC pathways between primary and metastases .

It is unclear whether this early data is in fact a reflection of true heterogeneity within the sampled populations or due to technical flaws within the data generation as subsequent studies performed within larger cohorts [81,82] demonstrated much higher concordance with known key CRC-driver mutations. Brannon *et al* [81] sequenced 69 matched primaries and metastases using a panel of 230 “key cancer associated genes” showing complete concordance in *RAS* and *RAF* status. This work did however demonstrate the

emergence of new mutations in *PIK3CA* and *TP53*, this observation was also made by Goswami *et al* [82]; this finding most likely represents the emergence of new tumour cell clones with implications for therapy.

A more recent study examined 26 patients with matched primary and metastatic deposits used ultra-deep sequencing to interrogate a panel of 100 'cancer genes' [83]. This highly sensitive sequencing demonstrated mutational heterogeneity in 4 (12%) cases, this rate of heterogeneity correlated with a Danish study including a similar number of patients using lower depth sequencing [84]. The examples of heterogeneity highlighted by ultra-deep sequencing included new mutations arising in metachronous metastases sampled after resection of the primary tumour and adjuvant chemotherapy, mutations unique to synchronous metastases and a mutation private to the primary tumour, not isolated in the matched metastasis; variation in 'driver' gene copy number was also occasionally identified. Due to the extremely high depth of sequencing this study constitutes high quality evidence that although evidence of mutational heterogeneity exists within genes related to neoplasia it appears to occur in a minority of cases. The use of focused 'cancer gene' panels potentially limits the scope of these studies, as 'cancer gene' panels are focused upon loci involved in the initiation of neoplasia and these panels may not be informative about genomic events which characterise and drive advanced neoplasia. Also, as the mutations within the panels occur early in neoplasia, they are unlikely to differentiate between lesions which have arisen late in the neoplastic process and thereby underestimate the degree of heterogeneity in advanced CRC and not identify novel genomic events.

Naxerova *et al* [85] adopted a novel approach to examine heterogeneity and the evolution of MCRC. This group, rather than examining cancer genes, focused upon the presence of mutations in non-coding polyguanine repeats, regions of the genome that are particularly prone to mutation, within multiple deposits from 19 individuals with MCRC; these deposits were from the primary tumour and metastatic deposits predominantly originating from lymph nodes and the liver. The pattern of mutations i.e. those events which were shared amongst multiple deposits (primary or metastatic), were used to infer the origin and relationship between the primary tumour and metastases and between the metastases themselves. Through bioinformatic interrogation of the data this work asserts that, within the majority (75%) of cases examined, liver and lymph node metastases arose from a different tumour cell clone identified within the primary tumour; this relationship had been suggested by meta-analytical data examining concordance of

KRAS mutation status between primary CRC with liver and lymph node metastases [86]. This would suggest, therefore, that liver metastases do not necessarily arise from lymph node deposits and do not share the same progenitor clone within the primary tumour.

This paper illustrated the relationship between tumours via a phylogenetic tree, with all lesions arising from a common trunk and diverging to form branches, comprising groups or clusters of lesions, which then split further to form the final leaves of the tree which represent single samples. This method of depicting the evolutionary relationship between lesions is widely used across papers within this field, though methodological differences between studies mean that the relationships inferred from such diagrams may differ radically [87–89].

The use of hypermutatable regions of the genome produced a large number of lesions from which to draw comparison between different deposits and therefore the data highlighting the difference between the deposits is very robust, however the information presented has several limitations. Firstly, the data does not give any mechanistic insight as to the metastatic process, it purely documents that liver and lymph node metastases differ genomically within a group of MCRC cases with resectable metastatic disease. These differences may reflect the fact that the entire tumour is capable of seeding metastases and the differences between the lymph node and hepatic metastases are due to a stochastic process or that the genomic divergence between the nodal and hepatic metastases is secondary to a specific characteristic of the tumour cell population which give arise to either type of metastases. As the material used in this study was gathered retrospectively the potential to identify the characteristics of each case in depth is limited. Secondly, once again, the material was taken from resectable disease and the nature of fatal disseminated disease may differ to that identified in this cohort.

Therefore, up to this point work examining mutational heterogeneity in CRC has illustrated the emergence of new tumour cell clones bearing mutations in key oncogenic pathways and has suggested that liver and lymph node metastases may be biologically distinct. Whilst comparison between the cited studies is difficult, as there is significant variation in the sequencing type, breadth and sensitivity, the more recent studies using higher resolution sequencing suggest mutational heterogeneity in ‘cancer genes’ is only identifiable in a minority of patients. Most significantly the cited work has only examined heterogeneity in patients with resectable disease; disseminated CRC is potentially biologically distinct requiring further characterisation.

1.3.2 Copy Number Alteration

As described above the presence of CIN and resultant gain and loss of gene copies (or copy number alteration (CNA)) is associated with a poor prognosis and increasing tumour stage in CRC [53,54]; as such tumoural heterogeneity in CNA is also potentially of prognostic and biological significance. CNA is also a useful tool for exploration of heterogeneity between multiple deposits as the genomic events of this type may be relatively large they can be detected by relatively low resolution sequencing [90]. This allows the entire genome to be examined for CNA at relatively low cost and requires a fraction of the computational input which would be required to analyse a whole genome sequence at single allele level.

Initial work examining CNA within primary and metastatic CRC was aggregated in a meta-analysis performed by Diep *et al* [91]; this analysis included 30 publications analysing a total 859 lesions. Within this large cohort of tumours, the group identified several CNAs which were significantly more common in metastases, these included loss of 8p and gain of 7p and 17q, on the strength of this evidence the authors suggest that these CNAs are important to the transition from primary to metastatic disease. The authors included a large number of lesions which were analysed by the same technique (comparative genomic hybridisation (CGH)), in an attempt to identify genuine differences between primary and metastatic lesions rather than random events or those which reflect different methodologies. However, the lesions compared in the study were not 'matched' i.e. the primary and metastatic lesions compared were not from the same individuals. It is therefore possible that the CNAs identified were indicative of more aggressive primary tumours, which had metastasised, rather than be reflective of evolutionary steps occurring during the development of metastatic potential. A minor further methodological issue is that although all the papers included used the same overall method (CGH), variations in the technique were present between the studies, however as the main CNAs documented were large, the impact of these subtle distinctions is unlikely to be significant.

The largest published comparison of CNA in matched primary and metastatic CRC was performed by Mekenkamp *et al* [92]. It compared 62 matched, resected primary and metastatic CRCs showing striking similarity in CNV between primary and metastatic tumours. According to the percentage of overlap between the copy number plots of the matched primary and metastatic lesions this group concluded that lesions within an

individual shared at least 70% of CNAs. The group excluded recurrent CNA abnormalities identified in previous studies of small cohorts and unmatched samples [91,93]. This study also analysed the CNA data by grouping or 'clustering' all 62 lesions by the overall similarity between the CNA plots for each lesion. This method consistently clustered the lesions from the same individual together demonstrating that a metastasis from one individual resembles the matched primary tumour rather than a lesion from a different patient. In conclusion the paper states that there is evidence of CNA divergence between matched primary and metastatic CRCs, but the degree of divergence was not significant either between matched metastatic and primary tumours or between matched tumours from different metastatic sites. Additionally, the authors suggest, within the context of metastatic disease, treatment decisions may be made from the genome of a single sample originating from the primary tumour. These conclusions may be an oversimplification of the data and the clinical context to which it applies.

Firstly, using the overall overlap between two CNA plots is a coarse method of comparison and may easily overlook quite significant chromosomal lesions, ideally an in-depth CNA analysis for each case would be presented. Secondly, the clustering methodology used was insufficiently sensitive to demonstrate divergence between matched lesions as unmatched tumour was the only comparator. It would appear highly unlikely that two unmatched tumours would converge to resemble each other rather than the primary tumour from which they arose.

Subsequent studies [83,94], performing more in-depth analysis of smaller cohorts of resected matched primary and metastatic disease, also have not shown recurrent specific CNAs which are characteristic of metastases when compared with the corresponding primary tumours. Higher depth analysis has, however, highlighted CNA in the region of specific oncogenes and tumour suppressor genes including *SMAD4* and parkin RBR E3 ubiquitin protein ligase 2 (*PARK2*). The observation that potentially biologically significant CNA lesions emerge within metastases is in contrast to the conclusions draw by Mekenkamp et al [92] and highlights the importance of a more in-depth and nuanced approach to CNA analysis.

In conclusion, as observed in work performing mutational analyses, the published studies examining CNA have examined resected colorectal tumours and have demonstrated a degree of evolution within metastases at the chromosomal level. Genomic alterations in loci containing 'cancer genes' have also been identified between

matched primary and metastatic lesions in several published analyses cases, however these events appear to be individual to the evolution of an individual's disease process and not recurrent across a cohort of cases. It is also not clear whether these CNA events are biologically significant or random, 'passenger', events arising due to the large shifts in genome content occurring due to CIN. Further examination of heterogeneity within disseminated disease is of value, as the spectrum of genomic change within multiple metastatic deposits is probably far greater than that observed in resectable MCRC. With a view to further expansion and exploration of the observations made in the cited literature, examination of genomic heterogeneity within a post mortem setting (plus correlation with any available antemortem samples) may provide the most complete landscape of genomic alteration in MCRC.

1.4 Genomic heterogeneity in carcinoma types

The examination of genomic heterogeneity between matched primary and metastatic tumours has also been performed in other carcinoma types [87,88,95–99]. A range of varying bioinformatic, sampling and sequencing methodologies have identified novel phenomena, some of which are likely to relate solely to a specific tumour type but others many illustrate principals applicable to a range of carcinoma-types in the context of advanced disease.

1.4.1 Renal cell carcinoma

Work in patients with metastatic renal cell carcinoma (RCC), performed by Gerlinger *et al* [88], examined multiple sections from primary tumour and metastatic deposits in 4 patients (all of whom had received the same systemic chemotherapy including mTOR inhibition). Samples were taken before and after chemotherapy. A combination of a mutational panel, generated from whole exome sequencing of primary tumours, and chromosomal analysis was used to demonstrate both intra and intertumoural heterogeneity in the majority of mutations and chromosomal abnormalities identified.

In contrast to the data in CRC, this paper described that the majority of oncogenic events in advanced RCC occur in subpopulations or 'subclones' of cells either within the primary tumour or metastases, although von Hippel Lindau (*vHL*) gene mutation was ubiquitous in all four cases. A novel phenomenon described in this work is the presence genomic events in the same oncogenic pathways in distinct tumour subclones. This apparent convergence of tumour genotype was also demonstrable via IHC for the relevant proteins.

This work is also distinct from that performed in CRC as it has examined a greater number of deposits within an individual than has been examined in CRC, it therefore may represent a more accurate depiction of genomic heterogeneity within advanced malignancy. The possibility that the convergent and divergent evolution documented is as a result of the selective pressure produced by chemotherapy is unlikely as significant heterogeneity was identified in the pre-treatment biopsies, additionally the post treatment samples were obtained shortly after the commencement of chemotherapy. It is of note however that the emergence of new mutations following the introduction of chemotherapy is an established phenomenon within the context of highly specific chemotherapeutic agents such as EGFR inhibitors [100,101].

1.4.2 Prostatic carcinoma

The concept of polyclonality was expanded upon by Gundem *et al* [87] who examined multiple metastatic sites in ten cases of fatal androgen deprived prostatic carcinoma (ADPC). This examination of heterogeneity included a greater number of samples per patient compared to the Gerlinger *et al* [88] and in contrast it found that in ADPC the majority of driver mutations were ubiquitous within individuals, although a broader range of evolutionary overall complexity was identified, particularly the presence of polyclonal metastases. Polyclonal metastases were defined as lesions containing genomic characteristics of more than one branch of the evolutionary or 'phylogenetic' tree of each case. Polyclonal metastases have also been described in animal studies [102,103], suggesting that this phenomenon arises due to distinct lesions (arising from a common primary tumour) that may metastasise to the same anatomical location, so called 'metastasis to metastasis' seeding. This situation possibly arises due to ease of access to a specific anatomical location or biological features of a metastatic site favouring the establishment of tumour deposits, McFadden *et al* [103] suggest that the lymph node micro-environment may provide an ideal environment for the establishment of polyclonality. It has also been shown in breast cancer models that clonal cooperation is advantageous to tumour progression [102].

1.4.3 Breast

Two publications have examined heterogeneity and clonal evolution in the disseminated breast carcinoma. These studies examined multiple deposits from individuals and used a mixture of archived material (usually of the primary tumour and any local metastases) and metastases sampled at post mortem [97,104]. The sequencing methodology used in both studies was similar to that described in Gerlinger *et al* [105], each deposit was characterised by the mutational and CNA profile and the shared and private features of each lesion were used to produce a phylogenetic tree for each case. Although the methods by which sub-clonal inferences were made differed between the two studies, similar patterns of evolution were identified. These were divergence of different tumour cell clones within the primary tumour and metastases with tumour to tumour seeding of metastatic deposits. Both studies also highlighted that within some cases of disseminated disease metastases arise from a single clone within the primary tumour, whereas in others multiple clones of tumour cells within primary tumour appear to spread throughout the body. The sequencing data within these cases was of high quality and

the overall trends between the samples sequenced are likely to be accurate as they broadly correlate with those in other organ systems and appear to be consistent regardless of the method of phylogenetic analysis. As both studies included heterogenous groups of patients many possible drivers of heterogeneity were present, such as the inherent biology of the disease and the effect of different chemotherapeutic regimens [95,96], along with the potentially confounding influence of differing combinations of formalin-fixed paraffin embedded (FFPE) and fresh material used in each case and different sampling protocols used during the diagnostic histopathological sampling of the primary tumours included in the study.

1.4.4 Carcinoma of the hepatobiliary tract

Similar trends to those observed in breast, prostate and renal cell carcinomas have also been made in lesions throughout the hepatobiliary tract. Both hepatocellular carcinoma (HCC) and Pancreatic (adeno)carcinoma (PC) have been shown to undergo significant clonal diversification throughout the evolution of metastatic disease [98,106–108]. Within HCC most work has examined multifocal disease arising on a background of Hepatitis B within partial hepatectomies or explant specimens; this provides an excellent opportunity to examine multiple matched deposits. The observed clonal diversification in HCC is unsurprising as Hepatitis B has many oncogenic properties including an effect on genomic stability and repair [109].

In disseminated PC it has been shown that metastases usually arise from multiple regions within the primary tumour (in contrast to the more restricted clonality observed in CRC where a significant number of cases demonstrated metastases arising from a specific subgroup of clones within the tumour [85]). This observation was made in a post mortem setting and therefore may reflect more in-depth sampling than was performed in CRC. However, PC has a particularly dire prognosis [110] and clonal diversity may be in part responsible for the diseases lethality and therefore similar in depth post mortem examination of MCRC is required to establish if this distinction exists.

A further interesting observation which has been made in PC is that this neoplasm does not evolve in a gradual ‘stepwise’ fashion but rather in evolutionary jumps. The idea of punctuated equilibrium is established theory in evolutionary biology [111] and describes the emergence of new characteristics by “rapid and episodic events in speciation” rather than a “stately unfolding” of events. Notta *et al* [99] used a combination of mutational and CNA data to demonstrate that CNA in PC often arises due to a single event causing

multiple CNAs and chromosomal rearrangements, so called “chromothripsis”. Such a catastrophic genomic event may account for the observed clonal diversity and aggressive behaviour of PC, although similar genomic events have been inferred in breast, prostatic and primary colorectal carcinoma [96,102,112].

1.4.5 Oesophageal carcinoma

An in-depth study of disseminated oesophageal carcinoma (OC) has not been performed, however Muragushu [113] *et al* sequenced material from primary oesophageal tumours before and after neoadjuvent chemotherapy and made an interesting observation pertinent to MCRC. This group used a targeted mutational panel and low coverage CNA sequencing to perform phylogenetic analysis; they demonstrated a similar branching evolution to other neoplasms discussed. In addition, it was demonstrated that the pattern of CNA in OC was relatively preserved, as compared to the mutational changes within the tumour, following platinum-based chemotherapy. This observation once more underlines the potential utility of CNA as a target for therapy.

These papers represent the most in-depth comparison of primary carcinoma and their metastases within an in-vivo non-xenograft setting and suggest that the limited sampling employed in previous studies may have underestimated the degree of heterogeneity present in MCRC. A more in-depth approach to the analysis of MCRC should establish the ubiquity of driver mutations and identify whether CRC share the genomic complexity described in ADPC with multiple tumour cell clones occupying common metastatic sites and if specific molecular subtypes converge upon key pathways vulnerable to targeted therapy.

1.5 The principals of determining phylogeny in metastatic carcinoma

As has been highlighted by the work across several different organ systems cited above, the concept of linear evolution from pre-malignant in-situ lesions progressing to invasive and subsequent metastatic disease have been expanded upon suggest a divergent pattern of evolution involving multiple subclones [114].

To summarise, all of the cited studies have demonstrated the presence of tumour heterogeneity, and it is an established concept that this arise due to an evolutionary process [115] with the accumulation of new genetic events either by a random ('stochastic') process as a result of genomic instability [116] or possibly driven by therapy [100]. These new events produce different populations or 'clones' within a tumour cell population which may be identifiable within the primary tumour and metastases. As has been demonstrated within the cited works (see sections 1.3 and 1.4) the pattern of genomic events within different clones identified throughout the body allow identification of lesions which are the precursor to others. The process by which new tumour cell clones arise and how they relate to one another is referred to as phylogenetic modelling. This process has been performed within clinical samples by both manual [117,118] and computational methods [85,87,105]. The traditional determination of phylogeny operates according to two basic 'Dirichletian' principles, firstly that no genomic event occurs twice within a tumour cell population and secondly that no genomic event is lost. These two rules dictate that if the combined frequency of two distinct events (producing two new clones) is greater than 100% then a proportion of cells must contain both events (the 'pigeon hole' principal) and because events cannot be lost the more prevalent clone must be the ancestor of the other. Although simplistic, these two central tenets of tumour phylogenesis have been demonstrated to be robust in cancer cell line studies [119]. These principals are illustrated in Figure 4.

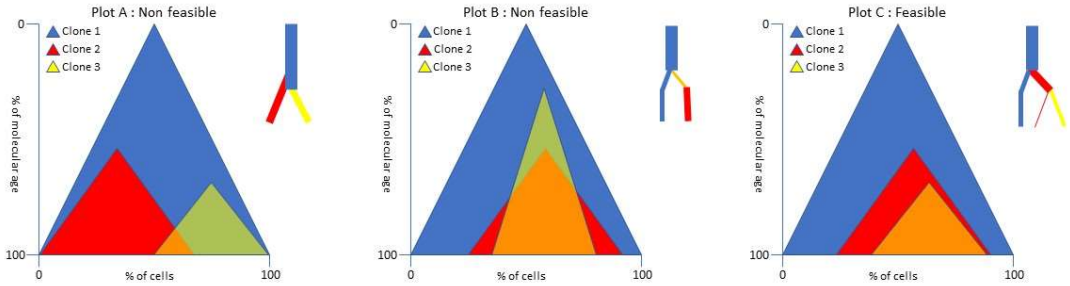


Figure 4: Figure demonstrating the two principals of Dirichletian phylogenetic inference in tumour cell populations. Plot A illustrates the 'pigeon hole principal', it shows a common ancestral clone (blue triangle) with two smaller subsequent clonal populations (which have arisen due to a new mutation or CNA) which have a combined frequency greater than 100%. This necessitates an overlap between the two populations, meaning that, if one population is not ancestral of the other, two identical genetic events must have occurred independently within both clones. This is considered highly unlikely; therefore, Plot A is not feasible. Plot B illustrates that the more prevalent event within a population must be the ancestor, as if it is not, then some cells must have lost a genomic event, which is considered improbable. Plot C demonstrates the most probable sequence of events with the two triangles overlapping with the largest clones pre-dating the smaller ones and no event occurring twice. Figure modified from Beerenwinkel *et al* [120].

The use of these principals is straight forward in the context of monoclonal populations; however, the inference of clonal phylogeny may be complicated by factors including polyclonal tumour samples, changes in ploidy (including loss of heterozygosity), variation in tumour cellularity and sequencing quality. In the context of mutational data, a range of computational approaches have been developed [121–123] which use different statistical models to produce the most probable relationship between lesions (and may incorporate CNA data). These methods of analysis require high depth sequencing and high computational power, the cost of which may be prohibitive in the context of disseminated disease with hundreds of samples. Additionally, these tools have only been developed to compare relatively small numbers of samples [121] (or indeed clonal variation within single samples [123]), therefore it is uncertain whether they can be applied to a large number (>100) of samples. Tracing phylogeny by CNA lessens the financial and computational demand of phylogenetic inference as lower resolution sequencing maybe adequate to identify clonal evolution, however some methods of automated phylogenetic analysis examine loci individually [124] and therefore inappropriately assign a greater importance to large CNA events when in fact, as markers of clonality, every CNA should be considered equally regardless of size. Several tools [117,125,126] have addressed this issue by identifying the breakpoints (sites of

change in copy number) within CNA data and using these to infer relationships between samples and tumour cell clones. However, these tools are only able to incorporate a limited number of samples and evolutionary steps or are limited to examining CNA data within specified portions of the genome.

In summary, methods for inferring the phylogenetic relationship between tumour samples may adopt a manual or automated approach. The current automated approaches require considerable computational and financial power and the capacity of these tools is still limited to a relatively small number of samples (in the context of disseminated disease) or restricted to specific data types.

1.6 Treatment

Over the past 30 years the prognosis of patients with limited metastatic disease has been transformed by technical advances allowing for curative resection [127]; the overall survival in this patient group is comparable to patients with nodal metastases [128]. Unfortunately surgical resection of distant metastases is only possible in 20% of patients [129], as metastases are either too extensive for resection or the patient is not considered a good surgical candidate for metastatectomy; for these patients with advanced disease the main therapeutic option is systemic chemotherapy.

Conventional systemic treatment for CRC in the United Kingdom consists of folinic acid, fluorouracil and oxaliplatin (FOLFOX) followed by the addition of irinotecan (FOLFOXIRI) or, folinic acid fluorouracil and irinotecan (FOLFIRI) [130]. This combination of chemotherapeutic agents has improved progression free (PFS) and overall survival (OS) in patients with inoperable CRC; however national and international drug trials have shown the majority of patients do not survive more than 2 years and although generally well-tolerated, grade 3-4 toxicity is observed in most patients treated with these regimens [131–133].

Highly specific monoclonal antibodies have been introduced to therapy; these new drugs aim to target specific growth pathways central to tumour growth thereby effecting tumour cell death with a more favourable side effect profile. Attempts to target tumour angiogenesis with Bevacizumab (a neutralising antibody against vascular endothelial growth factor (VEGF)), in combination with conventional chemotherapy show some small improvements in PFS and OS [134], however this outcome benefit has to be balanced against an observed increase in toxicity observed with the drug [135]. A further problem observed with the use of the drug is, as yet, no useful biomarker has been devised to predict response to therapy. As a result of the expense of this agent and the clinical issues raised, bevacizumab is not currently recommended for use in the UK by NICE [136]. The range of drugs targeting tumour angiogenesis has been expanded to include three further agents, ramucirumab, aflibercept and regorafinib, which bind to either VEGF or target one or more VEGF receptors (regorafinib also binds targets in several other pathways including RAS [137]). These drugs have shown some improvement in OS in patients with MCRC [138–140], however due to the considerable expense of these agents and small observed improvements in survival, they do not form part of standard care in the UK [141–143].

Constitutional activation of EGFR and associated pathways is a common event in CRC and other malignant neoplasms [144]. As such this receptor and the associated signalling pathway are a second target for monoclonal antibody drugs. This class of drugs, including cetuximab and panitumumab, (in combination FOLFOX) are recommended as first line therapy for attempted down staging of inoperable CRC metastases with a view to subsequent resection by NICE [130].

The evidence as to the efficacy of EGFR antagonists as palliative therapy has not been entirely consistent. Douillard *et al* [8] performed a retrospective analysis of individuals who had received panitumumab plus FOLFOX and demonstrated that the inclusion of panitumumab improved PFS and OS by 2.1 and 5.8 months in tumours which did not contain activating mutations in the downstream RAS-MAPK pathway. However a subsequent prospective randomised control trial (RCT) [145] examining the effect of EGFR antagonists failed to demonstrate any benefit in OS with the addition of panitumumab to irinotecan in patients WT for activating mutations in the RAS-MAPK pathway.

Currently the RAS molecules are of principal interest in predicting non-response in CRC treatment and activating mutations are identified in approximately 40% of tumours. Initially in *KRAS* mutations at codons 12, 13 [146] and subsequently at codons 61, 117 and 146 and codons 12, 13 and 61 in *NRAS* have been associated with treatment failure with EGFR antagonists [8]. In these early trials the small numbers of *BRAF* mutant cases failed to establish *BRAF* status as predictive of therapeutic response, but recent meta-analytic data is, however, supportive of the use of *BRAF* status as a predictive biomarker in EGFR blockade [147].

However, as is shown by the relatively modest improvements in outcome produced by EGFR antagonists, disease progression is almost inevitable despite improved molecular testing. Diaz *et al*.[100] demonstrated the emergence of *KRAS* mutations within circulating tumour DNA (CTDNA) in patients treated with cetuximab within six months of starting therapy; the patients' primary tumours were *KRAS* wild-type (WT) on pre-treatment testing. The emergence of *KRAS* mutant cells in these patients was either due to new mutation arising or a pre-existing *KRAS* mutant clone not sampled in the original assay; the authors of this paper performed mathematical analysis incorporating predictions regarding tumour doubling time, cellularity and mutation rates which suggested that the latter conclusion was most probable. The use of CTDNA is an emerging field in the treatment of CRC and has repeatedly demonstrated the capacity to demonstrate the

emergence of treatment resistant clones in individuals disease progression following EGFR blockade [148–155]. Within this body of literature there are examples of the emergence of new mutational events conferring treatment resistance [148,152] along with CNAs involving the regions encoding *RAS, Erb-B2 Receptor Tyrosine Kinase 2 (ERBB2)* and *MET* [154,155] in the absence of allelic changes. The identification of novel events outwith the spectrum of standard extended *RAS* testing suggests that a broader approach to the examination of treatment refractory and advanced disease may improve the power of pre-treatment testing.

As a result of the marginal benefits and high cost of EGFR blocking therapy further research is ongoing to determine novel predictive biomarkers such testing for the molecules activating EGFR signalling ('ligands'). Ampiregulin and epiregulin are the two principal molecules currently identified as ligands of EGFR and the expression of these molecules has been shown to be a prognostic biomarker along with a prognostic tool predictive of response to conventional chemotherapy [156] and EGFR blockade [157].

Mutations in BRAF status are thought to be of therapeutic significance to the use of EGFR antagonists [147] but BRAF is also exploited as an effective drug target in the treatment of malignant melanoma. Results in early trials using vemurafenib as monotherapy in MCRC were not promising [158], as a markedly heterogeneous response pattern was observed, but case study evidence has been published suggesting that combination therapy with EGFR antagonists may have some efficacy [159]. This combination therapy approach is currently the focus of an ongoing RCT [160].

PIK3CA mutation has not yet been established as biomarker in EGFR therapy [161,162] . However, mutations at this locus has become of clinical significance as patients with this abnormality have been observed to experience improved survival when taking low dose aspirin [163,164]. The mechanism by which this occurs it not entirely understood, but it is thought this effect is due to aspirin reducing prostaglandin-dependent stimulation of the PIK pathway.

Further drugs targeting the PIK and the molecules associated pathways, including AKT, MET and mTOR inhibitors, are in development [165–167] (*mTOR* inhibitors are already established in the treatment of RCC [88]) and combination therapy with drugs acting on the RAS-MEK pathway is also being examined [168–170]. It has been suggested that expression of PTEN may also predict response to EGFR antagonists [17,18], this biomarker is undergoing further validation as part of the FOCUS4 study, along with BRAF [171].

As previously described, a subgroup of MSI CRC are characterised by a brisk immune response which is open to exploitation by immunotherapy [50]. The most established, current therapeutic approach is to augment the patient's immune response to the antigenicity of MSI tumours by blocking the inhibitory 'immune checkpoint' molecules. These molecules are expressed within normal homeostasis to prevent immune response to autoantigens but within neoplasia allow antigenic cancer cells to avoid immune surveillance and cytotoxic T-cell response; the two most well characterised molecules within this context are PD-1 and cytotoxic T-cell associated antigen 4 (CTLA-4). however PD-1 inhibition has produced some dramatic responses in MMR deficient tumours [172,173] and large RCTs are currently evaluating these agents [174]. PD-1 is a particularly attractive therapeutic target as the expression of one of the activating ligands of this receptor (programmed death-ligand 1 (PDL-1)) has been shown to be a predictive of drug response [175]. Agents targeting CTLA-4 are yet to produce clinically significant results as monotherapy [176] although early results from a trial examining combination PD-1 and CTLA-4 in the neoadjuvant setting are promising in the context of dMMR CRC [177].

Other immunotherapies (including vaccine based immunotherapy and transduced T cells) are also in development, however few have progressed beyond phase-1 trials, instances of marked toxicity and mixed therapeutic results have been observed [178,179].

Therefore, up to this point, the use of targeted therapy has produced small improvements in PFS and OS and work in EGFR blockade has shown that treatment resistance is associated with the presence and emergence of clonal diversity in MCRC. As described in section 1.3, the current literature only contains analysis of clonal diversity and heterogeneity within resectable disease whereas targeted and conventional systemic therapies are primarily used in the setting of disseminated MCRC. Exhaustive characterisation of the genomic diversity within cases of disseminated MCRC will provide a more complete picture of the cancer genome in MCRC and clarify the role of tumour heterogeneity in treatment resistance.

2 Aims and Objectives

The aim of this project is to identify and characterise the observable heterogeneity within a group of individuals with mCRC.

These observations will primarily be made in the post mortem setting and from material sampled during these examinations.

The two principal levels of analysis will be phenotypic and genomic.

The phenotypic analysis will comprise documentation of the macroscopic patterns of disease and the microscopic features of each tumour deposit. The microscopic features will be examined on haematoxylin and eosin, immunohistochemical and tinctorial staining.

The genomic study of the tumours will be performed at the allelic and chromosomal level.

The allelic or mutational analysis will examine both an extended *RAS* testing mutational panel and WGS. The chromosomal analysis will be performed using low resolution WGS. Phylogenetic analysis will be performed on the genomic data.

3 The 'Gift' autopsy project

3.1 Introduction

3.1.1 The decline in the hospital autopsy

Hospital or 'consent' autopsy is a post mortem (PM) examination performed with the consent of the next of kin, not to establish the circumstances or cause of death, but to gain additional information about a death for clinical audit or research purposes (which may involve tissue sampling) [180]. Of the three categories of autopsy, forensic, coronial and consent or 'hospital' examination, it is this, latter type, which has seen the most considerable decline in numbers. Whereas approximately 22% of deaths in the UK undergo autopsy examination as a result of referral of a death to the coroner, under 1% of deaths underwent consent/hospital examination [181], this is compared to a rate approaching 20% in 1990. The reasons for this stark decline are not entirely clear, although it is speculated that a combination of public and clinical attitudes to autopsy [182], medicolegal [183] (particularly in the aftermath of the organ retention scandal [184] and the implementation of the 2006 Human Tissue Act [185]), and infrastructural (due to the decline in numbers of centres and pathologists training to performing autopsy [186]) issues are responsible for the observed trend.

It would appear that the concerns held by clinicians regarding perceptions of autopsy are misplaced as the evidence examining public attitudes toward autopsies suggest that patients and their families are likely to consent to research or to coronial autopsy when correctly counselled [187,188]; amongst those individuals studied who have not given consent, the stated reason for refusal include concerns regarding disfigurement, delay in funeral arrangements and religious beliefs. Although the literature is not extensive in this area, what evidence that does exist, suggests that when approached, the families of the deceased are highly receptive to the retention of tissues sampled at coronial autopsy for research [189]. Additionally qualitative data also exists that the families of individuals who have undergone consent post mortem deem the examinations to be important for the improvements in medical care[188] and find the experience beneficial as reassurance is gained regarding the quality of medical care received by family members and through contributing to the advancement of medical knowledge [187,188].

The reduction in the numbers of consent autopsies performed for research purposes is problematic as they represent a rare opportunity to study disease which involves parts of the body essential to life e.g. the brain or diseases in an advanced state *in-vivo* such as disseminated carcinomatosis. To this end, there is some increased interest in establishing research autopsy programs for degenerative central nervous system diseases [190,191] and in those disease states when ante mortem sampling is inappropriate on clinical grounds [192], for instance individuals with inoperable malignancy receiving palliative care. It is within the context of advanced malignancy that research post mortem provides an opportunity to gain a more complete picture of the advanced cancer genome; the current alternative approaches would include multiple biopsies of metastatic sites [88] (which would be aversive for use in standard clinical practice and may still underestimate the heterogeneity of the cancer burden within an individual) or 'liquid-biopsy' measuring for the presence of circulating tumour cells or DNA within the patient serum [193]. The latter of these two methods is an evolving technology which has been proven to demonstrate the emergence of treatment resistant tumour cell clones [100,194], but it is yet to be demonstrated whether analyses of this sort are capable of demonstrating the full range genomic aberrations present in an individual suffering from disseminated carcinomatosis [152].

The University of Leeds established the 'Gift' research autopsy project [195] with a view to performing consent autopsies to provide tissue for research, either as control material or for specific projects with ethical approval. 174 examinations have been performed since 2009, mainly on individuals who have died with advanced malignancy. This project presented an opportunity to perform a series of examinations on individuals who have MCRC, to access the distribution of disease and extensively sample as many lesions as practically possible for phenotypic and genotypic analyses.

As histopathological assessment usually forms part of an autopsy examination, this section will also include assessment of the significant morphological characteristics of each case.

3.1.2 Histopathological assessment of CRC

Patients with CRC are usually diagnosed by histological examination of tissue sampled as part of a diagnostic or therapeutic procedure; this procedure is performed either due to patient symptoms or following referral as part of a screening program. The most common diagnostic procedure for CRC is an endoscopic examination of the large bowel, during which biopsies may be taken from the primary tumour. The samples taken at endoscopy are often small and superficial, therefore the information provided to the clinician is usually limited to the tumour subtype and histological grade; the presence of vascular invasion within the bowel mucosa may be commented upon if present. An excision of the primary tumour will then be performed provided that the lesion is resectable and the procedure is appropriate within the clinical context; traditionally resection of colorectal carcinoma involves removal of a portion of bowel and the regional lymph nodes, however small, superficial tumours may also be excised endoscopically. The resection 'specimen' is examined both macroscopically and microscopically to give a more comprehensive assessment of the features predictive of outcome in CRC.

There is a broad range of literature regarding the histopathological parameters with a bearing on prognosis but the most well established factors are published as part of The Royal College of Pathologists (RCPath) Dataset for histopathological reporting of colorectal carcinoma [196]; this and the significant surrounding literature are summarised below. The items from the RCPath guidance which refer only to local resections of early tumours are not included as they are not relevant to the group of patients recruited to this work.

3.1.2.1 Position and size of the tumour

Tumours of the large bowel may be divided broadly into those which arise within the colon and those arising within the rectum. The rectum is the most common site for colorectal carcinoma followed by the sigmoid colon and then the caecum [197]. Although the anatomical boundaries which delineate these two regions are not completely clear, data from large prospective studies suggests that survival amongst rectal tumours is equal or slightly superior to that of colon cancer and this trend is a reversal of that observed in the 1980s [198–200]. This trend is most likely a result of improvements in treatment (i.e. the quality of surgery and use neoadjuvant therapy [201,202]), which have ameliorated any adverse biological or anatomical factors which historically produced poorer prognosis in rectal cancer.

If the disease within the colon is divided into that which arises proximal ('right colon') or distal ('left') to the splenic flexure overall survival is higher amongst individuals with left-sided tumours. This trend is present once data is adjusted for disease stage and adjuvant treatment effect [203] and therefore these differences most likely reflect the differing biology of disease at different sites within the bowel, which has been explored in the previously cited classification studies [16,50].

3.1.2.2 Margin status

For major resections of CRC tumours (i.e. those which involve removal a section of bowel rather than an endoscopic of a tumour) the specimen is assessed for involvement of the longitudinal and circumferential margins of resection. The longitudinal margin (LM) refers to the cut edge of the bowel at either end of a resection specimen. According to national guidelines the LM should be assessed initially macroscopically and if the LM is greater than 30mm from the tumour it is not sampled [204] for microscopic examination. If the tumour exhibits high risk features on microscopic examination such as high-risk histological subtype (see below) or infiltrative growth pattern the resection specimen may be revisited for sampling of this margin. Increasing tumour distance to the longitudinal margin is predictive of lower rates of recurrence [205] and higher lymph node yields from the surgical specimen [206] which may allow more accurate staging of a tumour.

The circumferential margin of resection (CRM) describes the radial, non-peritonealised, cut aspect of a CRC resection specimen. Within the colon, the CRM lies posterior to the bowel and is usually separated from the colon (and tumours arising within it) by a portion of mesentery which contains fat plus the lymphatic and blood supply for the resected portion of the bowel; the CRM within the colon is therefore rarely directly involved by tumour, although tumour extension to this margin is associated with local recurrence [207]. The anterior aspect of the colon is surrounded by the peritoneum and does not constitute a 'true' margin. The CRM within rectal cancer specimens is non-peritonealised throughout 360 degrees and is particularly vulnerable to involvement by tumour, which is associated with poor prognosis [208]. Surgical technique and the pathological assessment of the CRM has therefore been the focus of extensive investigation.

3.1.2.2.1 The rectal CRM and quality of resection

The quality of resection is a well-established prognostic factor in rectal carcinoma and provides a useful tool for feedback of surgical quality. The assessment of surgical quality in rectal carcinoma involves assessment of the circumferential surgical margin; as the

rectum lies within the pelvis this margin is non-peritonealised and is a 'true margin' comprising connective tissue as opposed to the more proximal bowel which lies within the peritoneum and is predominantly surrounded by serosa. Extension of tumour to the non-peritonealised margin is associated with high rates of recurrence [208] and it has been demonstrated that surgery within the appropriate plane significantly reduces rates of recurrence [202]. The desired plane of resection usually is at the level of the 'mesorectal' fascia which encloses the rectum within a layer of fat but may include a portion of the pelvic floor muscles in low rectal tumours. In recognition of the importance of this ideal surgical plane, standard histopathological reports within the UK should include a comment upon the plane of excision and any defects in rectal tumours to provide feedback to the surgical team [196]. A similar protocol for the assessment of colectomy specimens exists, in which the quality of mesenteric excision is the quality parameter, but this has not yet been validated as part of large prospective trials and therefore does not form part of the core dataset for the reporting of colectomy specimens in the UK [209,210].

3.1.2.3 Tumour stage and size

CRC is staged as per the parameters within the The American Joint Committee on Cancer (AJCC) tumor node metastasis (TNM) staging system and assigned numerical values according to the local extend of the tumour infiltration ('T'), regional lymph node involvement ('N') and the presence of regional metastases ('M'). As the TNM criteria are used for other means of tumour staging e.g. radiological, the TNM stage of a tumour provided by pathological examination is preceded by a 'p'. The most recent (eighth) edition of the TNM classification (TNM8) [211] was issued in 2016 and was recommended for use by the RCPath in 2018. Prior to January 2018, all CRC specimens reported in the UK were staged according to the fifth edition of the TNM classification (TNM5) [212], the subsequent two editions contained definitions which were felt to be insufficiently supported by clinical data and were therefore not adopted for use in the UK. This change in reporting guidelines is of note as the primary tumours described in subsequent sections were reported by the original pathologist according to the TNM5. The differences between the two editions lie in the subgrouping of tumours which have extended out with the bowel serosa ('pT4' tumours), the definition and classification of tumour nodules or deposits within the pericolic fat which are not surrounded by a lymph node or vascular structure and the grouping of metastatic disease; these alterations are shown in Table 1.

Despite these subtle variations and controversies regarding pathological definitions, the pathological TNM stage of a tumour has proven prognostic power and is central to the multidisciplinary management of CRC.

It is of note that tumour size does not form part of the TNM criteria, this parameter was felt to lack prognostic significance according to several small prospective trials summarised in the AJCC guidance published in 2000 [213]. Several subsequent large retrospective analyses have concluded that increasing tumour size does correlate with a reduction in survival although the exact tumour size cut-offs holding maximum prognostic significance are yet to be defined [214–216].

Table 1: Staging criteria for colorectal carcinoma according to the 5th and 8th editions of the TNM classification of malignant tumours [211,212]

	TNM edition 5			TNM edition 8		
pT	1		Tumour confined to submucosa	1		Tumour confined to submucosa
	2		Tumour extends into muscularis propria	2		Tumour extends into muscularis propria
	3		Tumour extends into pericolorectal tissue	3		Tumour extends into pericolorectal tissue
	4	a	Tumour extends into adjacent organs	4	a	Tumour invades visceral peritoneum
		b	Tumour invades visceral peritoneum		b	Tumour extends into adjacent organs
pN	0		0 lymph nodes contain tumour	0		0 lymph nodes contain tumour
	1		up to 3 lymph nodes contain tumour *	1	a	1 lymph node contains tumour
					b	2-3 lymph nodes contain tumour
					C	0 lymph nodes contain tumour but tumour deposits † present in pericolorectal tissues
	2		≥4 lymph nodes contain tumour	2	a	4-6 lymph nodes contain tumour
					b	≥7 lymph nodes contain tumour
pM	x		Presence of distant metastasis not assessable			
	1		Metastasis to distant organs	1	a	1 site or organ without peritoneal metastasis
					b	≥2 site or organ without peritoneal metastasis
					c	Peritoneal metastasis with or without other metastases

* According to TNM edition 5, a tumour nodule greater than 3mm in maximum dimension in the pericolorectal tissue without evidence of a lymph node or blood vessel was regarded as a lymph node metastasis. If a tumour nodule was less than 3mm it was regarded as discontinuous extension of primary tumour and not included in the 'N' stage.

† According to TNM edition 8, a tumour deposit is defined as a discrete nodule of cancer within the peritumoural soft tissue without histological evidence of a lymph node, blood vessel or nerve.

3.1.2.4 Histological types

The majority of carcinomas within the large bowel are adenocarcinoma, the other histological carcinoma-types are adenosquamous, squamous, spindle, undifferentiated and neuroendocrine; these are rarer than adenocarcinoma and have distinct clinicopathological characteristics [217] and the term 'colorectal carcinoma' (CRC) refers to adenocarcinoma if not otherwise specified. Adenocarcinoma of the colorectum encompasses conventional adenocarcinoma (characterised by the formation of glandular lumen) along with histological subtypes including signet ring, mucinous, medullary, serrated, cribriform comedo-type and micropapillary carcinoma. Signet ring and mucinous carcinoma are the most common and well characterised of the histological subtypes; signet ring carcinoma is the rarer of the two (accounting for 1% of CRC) and carries a poorer prognosis as compared to conventional CRC [218]. Mucinous carcinoma is associated with MSI (as are medullary carcinoma and a small proportion of signet ring carcinoma) but the literature regarding prognostic significance of this tumour subclass is as yet unclear [218–220], possibly due to inconsistency in diagnosis and biological heterogeneity within the lesions fulfilling the diagnostic criteria for mucinous carcinoma [221].

3.1.2.5 Differentiation

CRC are graded by the degree to which they form glandular lumena, within conventional-type adenocarcinoma this approach has proven prognostic significance [222,223]. According to the most recent international guidance [211] CRC should be divided into four grades, G1 (well differentiated with 95% lumen formation) to G4 (undifferentiated, with no appreciable gland formation), however the most recent UK RCPath guidance advocates use of a two-grade system dividing tumours into well/moderately differentiated and poorly differentiated according to the predominant grade; the simplified system has been advocated on a pragmatic level to improve interobserver correlation [196]. Both systems are only recommended for use in conventional-type adenocarcinoma, the behaviour of the adenocarcinoma subtypes does not correlate well with the standard grading according to UK or international guidelines.

3.1.2.6 Tumour Budding and necrosis

Tumour budding is defined as the presence of isolated single cells or small clusters composed of less than five cells at the invasive margin of the tumour. The presence of

this infiltrative pattern of growth has been shown to be independently predictive of survival and local recurrence [224–227] with and without neoadjuvent therapy [228]. However due to concerns regarding the reproducibility of the parameter, tumour budding does not form part of standard reporting protocols in the UK or Europe [196,229].

The proportion of tumour necrosis and inflammatory tumour response are two further histological features which have possible prognostic predictive power. These parameters have not formed part of the current reporting guidelines in the UK as they lack a clear consensus is to the most appropriate method of scoring and reflect complex processes involving tumour and patient factors [230–234].

3.1.2.7 Vascular invasion – intramural and extramural

Vascular invasion (VI) is associated with poor prognosis in many solid malignancies [235–238] and within CRC the understanding and classification of VI is an evolving area. VI in the context of primary CRC is divided into that involving veins out with the muscularis propria (MP) (extramural venous invasion (EMVI)) and that within the MP (intramural vascular invasion (IMVI)) which may involve blood or lymphatic channels.

The importance of EMVI has been appreciated since the 1930s [239] and is long established as an independent prognostic factor in CRC [240,241]. The identification of EMVI requires the identification of tumour with an endothelial lined space which either contains red blood cells or is surrounded by a rim of muscle. Despite the use of this concise definition there is significant variation in the rate of identification between histopathologists. The use of special stains to highlight the elastic lamina of venous channels may improve the rate of detection of EMVI, especially in cases displaying suboptimal tissue morphology due to autolysis or processing artefact [242–244].

The prognostic significance of IMVI has only recently been established possibly due to difficulty the accurate identification of tumour within the predominantly small vascular channels within the MP of the bowel [242]. However, due to the increasing rate of resection of early CRC and the use of immunohistochemistry, meta-analytical data has been produced which suggests that IMVI is predictive of lymph node metastases [245] and therefore this parameter has been included in the most recent UK reporting guidelines.

3.1.2.8 Perineural invasion (PNI)

In a similar fashion to IMVI recent meta-analytical data indicates [246] the importance of PNI in CRC and it is included in the most recent UK reporting guidelines. There is, in fact, a suggestion that the identification of PNI may show greater interobserver correlation than VI and that intra and extramural PNI may hold differing prognostic significance should be differentiated [247], however this suggestion does not form part of UK or international guidelines.

3.1.2.9 Response to neoadjuvant therapy

Neo-adjuvant chemoradiotherapy has become a standard part of rectal cancer treatment. It has been observed that those tumours which show a complete or significant response to therapy have a superior outcome to those which show no regression although many lesions show an intermediate response[248,249]. Several classification systems have been devised for the assessment of regression in rectal tumours, which resemble those used in other parts of the gastrointestinal tract and show a variable degree of interobserver variation[250]. The current system currently recommended for use in the UK and as part of the AJCC guidance [251] is a four-tier system shown in Table 2.

Table 2: Tumour regression scoring criteria as per AJCC [251] and RCPath [196] guidelines (modified from Ryan *et al* [252])

Microscopic description	Score
No viable tumour cells	0
Single cell or small groups of cancer cells	1
residual cancer outgrown by fibrosis	2
No fibrosis with extensive residual cancer	3

3.2 Aims and Objectives

1) Document the distribution of metastatic disease within a cohort of individuals (’donors’) with MCRC as part of the ‘Gift’ research autopsy project.

2) Retrieve clinical information pertaining to disease course and treatment for each donor.

3) Retrieve any tumour samples taken during life.

4) Confirm the diagnosis of colorectal adenocarcinoma within the primary tumour and each metastatic deposit.

5) Identify the nature of each metastatic deposit (i.e. intra-nodal, intravascular, peritoneal).

6) Confirm the mismatch repair status of each tumour by IHC.

3.3 Materials and Methods

3.3.1 Tissue sampling referral pathway and ethical approval

Patients were referred to the project by one of two methods, either opportunistically if an individual with MCRC contacted Leeds Teaching Hospitals NHS tissue services expressing a desire to donate their body to research after death or via oncology outpatient clinics where a poster was displayed advertising the 'Gift' project (see appendix 1: Gift referral poster). This latter method was implemented with the involvement and support of the oncology team at St James University Hospital (SJUH).

Patients who are referred to the 'GIFT' team were able to discuss the post mortem examination and donation process along with their next of kin. With this information a donor could give consent as to the type of post mortem (limited or full) and which tissues they wished to donate. Consent was also gained for the correlation of post mortem findings with clinical details and pathological findings during life. Once the donor had passed away, the patient's next of kin were then required to give their consent before the post mortem and tissue sampling could take place. A formal post mortem report was provided to the next of kin if requested.

Ethical approval for this project has been gained from Northeast and Tyneside NHS ethics committee (REC reference: 13/NE/0079).

3.3.2 Clinical Data

In accordance with the consent gained and ethical approval, a clinical history was gathered from an interview taken during the consent process and clinical notes. Data collected included:

- Age

- Past medical history

- Date of CRC diagnosis

- Surgical and drug history relevant to CRC

- Date of death

3.3.3 Autopsy Procedure

During the examination, within the limits of the expressed consent, standard autopsy procedure was observed. This comprises of patient identification, external and internal examination followed by sampling for histological examination or other special techniques.

During external examination the patient was identified and any distinguishing marks, features of illness or previous medical intervention were documented. The internal examination involved a systematic examination of the organ systems noting the weight and condition of the viscera. The size and position of any tumour nodules was documented, when possible photographs were taken (ensuring donor anonymity).

All tumour deposits were then sampled along with macroscopically normal (control) tissue material (usually spleen and liver). Each sample was taken in duplicate; one sample was formalin fixed paraffin embedded (FFPE) and the other frozen. Multiple sections were taken from large metastatic deposits and the primary tumour.

Matched fixed and frozen sections were taken in all cases (except GD2); FFPE material is more easily stored and provides superior histological morphology, however fresh tissue is preferable for sequencing platforms which require less fragmented, higher quality DNA. For GD2 only FFPE material was available as the examination and sampling was performed by an individual unfamiliar with the protocol for the project.

3.3.4 Histological staining

5μm sections of each FFPE tissue block were initially stained with haematoxylin and eosin (H and E) to confirm the presence of colorectal adenocarcinoma and identify the surrounding tissue thereby confirming the nature of the deposit e.g. hepatic metastasis, lymph node metastasis, vascular tumour embolus.

Histochemical staining for elastin was performed to confirm the presence of VI in each case and to aid the characterisation of tumour deposits which lay within adipose tissue without evidence of a surrounding lymph node or vascular structure on H and E staining. IHC staining for S100 was performed in areas suspicious for PNI by tumour.

Mismatch repair status was also established using IHC staining for MLH1, MSH2, MSH6 and PMS2 antigens. Two sections from each case were stained, one from the primary

tumour and the second from a well-preserved metastasis.

H and E staining was performed according to the following protocol:

- Sections de-waxed
- 2 minutes (min) Meyers haematoxylin
- Rinsed in running tap water
- 2 min Scott's tap water
- Rinsed in running tap water
- 1 min eosin
- Rinsed in running tap water
- Sections dehydrated in ethanol and xylene
- Mounted in dibutylphthalate polystyrene xylene (DPX)

Elastin staining was initially performed according to the following protocol:

- Sections de-waxed
- 30 min Verhoff's haematoxylin
- Rinsed in running tap water
- Differentiated in 2% ferric chloride (15 seconds)
- 1 min 5% sodium thiosulphate
- Rinsed in running tap water
- 5 min Sirius red
- Air-dry sections
- Sections dehydrated in ethanol and xylene
- Mounted in DPX

Due to inadequate staining an alternative elastin staining protocol was followed:

- Sections de-waxed

- 2 minutes acidified potassium permanganate

- Rinsed in running water

- 1 min 1% oxalic acid

- Washed in 70% ethanol

- 45 min Weigert's Resorcin Fuchsin

- Rinsed in running water

- Differentiated in acid alcohol

- 5 minutes Sirius red

- Rinsed in running water

- Sections dehydrated in ethanol and xylene

- Mounted in DPX

IHC staining for S100 antigen was performed, by hand, according to the following protocol:

- Sections de-waxed.

- Antigen retrieval performed in Menarini Access Revelation solution (Menarini Diagnostics, Berkshire, UK) (1 in 10 dilution in deionized water) @ 125°C for 2 minutes and cool to 90°C in antigen retrieval unit.

- Wash in water.

- Endogenous peroxidase (PO) activity blocked in Menarini PO block,100 µl per slide for 10 min @ room temperature.

- Slides washed for 5 minutes in Menarini wash buffer

- Non-specific antibody binding blocked with Casein Menarini Block solution, 100 µl per slide for 10 min @ room temperature.

- Rinsed in Menarini wash buffer.

- 100µl primary antibody (Dako anti-rabbit S100 cat Z0311) added, (diluted 1:500 Zymed) (Life Technologies, Carlsbad, USA), incubated for 5 min @ room temperature.

- Washed three times in Menarini wash buffer for 5 minutes.

- Incubated in 100 µl of Menarini X-Cell Polymer Horse Radish Peroxidase (HRP) reagent for 30 minutes @ room temperature.

- Washed three times in Menarini wash buffer for 5 minutes.

- Incubate in 100 µl Menapath diaminobenzidine (DAB) solution for 5 minutes @ room temperature.

- Washed in Menarini wash buffer for 5 minutes and then in tap water

- Counterstained in Mayers Haematoxylin for 2 minutes.

- Differentiated in Scotts tap water for minutes.

- Sections dehydrated in ethanol and xylene

- Mounted in DPX

IHC for MMR status was performed using the Dako Autostainer Link 48 automated stainer (Agilent technologies, Santa Clara, California, USA) according to the manufacturer's protocol. Antibody clone, antigen retrieval and immune detection solutions are described in Table 3.

Table 3: MMR IHC protocol

		MLH1	MSH2	MSH6	PMS2
Antibody	**Type**	Dako monoclonal hMLH1 mouse antibody Clone ESO5 (ready-to-use)	Dako monoclonal hMSH2 mouse antibody Clone FE11 (ready-to-use)	Dako monoclonal hMSH6 mouse antibody Clone EP49 (ready-to-use)	Dako monoclonal hPMS2 rabbit antibody Clone EP51 (ready-to-use)
	Incubation temperature	Room temperature	Room temperature	Room temperature	Room temperature
	Incubation time	1 hour	1 hour	1 hour	1 hour
Antigen retrieval		High pH Dako target retrieval buffer	High pH Dako target retrieval buffer	High pH Dako target retrieval buffer	High pH Dako target retrieval buffer
Immune detection		Dako Flex Envision High pH	Dako Flex Envision High pH	Dako Flex Envision High pH	Dako Flex Envision High pH
Counterstain		Haematoxylin	Haematoxylin	Haematoxylin	Haematoxylin

3.4 Results

3.4.1 Recruitment and clinical data

Eight individuals who had died with advanced colorectal cancer underwent autopsy examination as part of this project.

3.4.2 Clinical data and resection histology

The clinical data for these patients is summarised in Table 4. The age range within the group was large (40-92 years). 6 of the 8 patients ('Gift' donors (GD) 1-2 and 4-7) presented with disseminated, unresectable disease; only 3 patients (GD3, 4 and 8) underwent resection of the primary tumour and only one donor (GD8) survived more than 15 months from the time of diagnosis. Two patients (GD1 and 8) received systemic chemotherapy, which comprised multiple cycles of conventional chemotherapy with the addition of a monoclonal antibody in the event of disease progression. The monoclonal antibody-type drug used in each case was different, GD1 received an EGFR antagonist (cetuximab) and GD8 was treated with VEGF blockade (aflibercept). GD3 underwent a resection of a locally advanced rectal tumour and adjuvant therapy was not administered due to co-morbidity and patient wishes, although this individual did receive palliative radiotherapy (DXT) to recurrent local disease. Those patients who did not receive chemotherapy (GD2-7) had significant disease burden at the time of diagnosis and/or were not likely to tolerate or benefit from palliative chemotherapy.

Four of the eight primary lesions were in the rectum, two were caecal and single instances of sigmoid and transverse colon were also analysed. The primary tumours resected from GD3, 4 and 8 were all described by the reporting pathologists as moderate to poorly differentiated tumours which were locally advanced. The primary tumour in GD3 extended into the bladder and was therefore excised by a multiorgan resection (involving resection of part of the bladder) whereas the other two patients underwent conventional colectomy surgery. The primary tumour in GD3 was staged as pT4a (TNM5), the other two resected primary tumours perforated the visceral peritoneum and were classified as pT4b (TNM5). All three resections contained involved lymph nodes ('pN1') and displayed EMVI. A peritoneal deposit was also resected from GD3 during the resection of the primary tumour.

Table 4: Clinical details of MCRC post mortem cases showing the clinical course and extent of disease.

		GD1	GD2	GD3	GD4	GD5	GD6	GD7	GD8
Age		54	73	59	80	92	40	69	61
Date of diagnosis		June 2010	May 2012	December 2012	September 2013	March 2013 (clinical Dx)	May 2014	February 2015	April 2011
Date of death		September 2011	May 2012	November 2013	July 2014	September 2014	September 2014	May 2015	June 2015
Primary site		Rectal	Rectal	Rectal	Caecum	Caecum	Sigmoid	Transverse colon	Rectum
Surgical resection		Tumor in-situ	Tumor in-situ	Tumor resected April 2013 (pT4a N1 M1) (TNM5)	Tumor resected (pT4b N1 Mx) (TNM5)	Tumor in-situ	Tumor in-situ	Tumor in-situ	Tumor resected May 2011 (pT4b N1 Mx) (TNM5)
Rx	**1st line**	Oxaliplatin, capecitabine	None	DXT to recurrence September 2013	Capecitabine (6 cycles, stopped due to comorbidity)	None	None	None	Neo-adjuvant DXT/5-FU
	2nd line	FOLFOX							FOLFOX
	3rd line	FOLFIRI							FOLFIRI Aflibercept
	4th line	Cetuximab							Palliative DXT

3.4.3 PM examination findings and sampling

The distribution of disease was varied, ranging from intra-abdominal disease as seen in GD3 to disseminated carcinomatosis in GD8.

The primary tumours which were in-situ (in individuals GD1, 2, 5-7) were all locally advanced extending either through visceral peritoneum or into adjacent viscera. The entirety of these lesions was sampled, except for the primary tumour in GD1, this tumour contained a stent which significantly impaired sampling. Two of the three cases which had undergone resection (GD3 and 4), displayed bulky tumour deposits at the site of previous resection.

Local lymph nodes were present in all cases except for GD6, distant lymph node metastases were present in GD1, 5, 7 and 8.

Distant metastases were extensive in all cases except GD3, this case showed only locally recurrent disease in the pelvis plus soft tissue deposits in the abdominal wall and a peritoneal metastasis. Bulky intrahepatic deposits were present six of the eight cases; GD3 and GD5 did not display liver lesions however the latter case showed multiple lung metastases, as did GD2 and GD8.

Peritoneal metastases were present in GD1, 3, 4 and 7 and tumour deposits within the pancreas were identified GD1 and 7.

The sampled disease distribution is shown in Figure 5 and Figure 6, the number of FFPE samples taken per lesion is shown in Table 5.

In total 375 FFPE samples were taken from 266 tumour deposits, including blocks retrieved from previous surgery in cases GD3, 4 and 8; multiple blocks were taken from large metastases. As previously described, in cases where the primary tumour was in-situ the entirety of the primary tumour was sampled, except for GD1; in this case a stent within the rectum impaired systematic sampling.

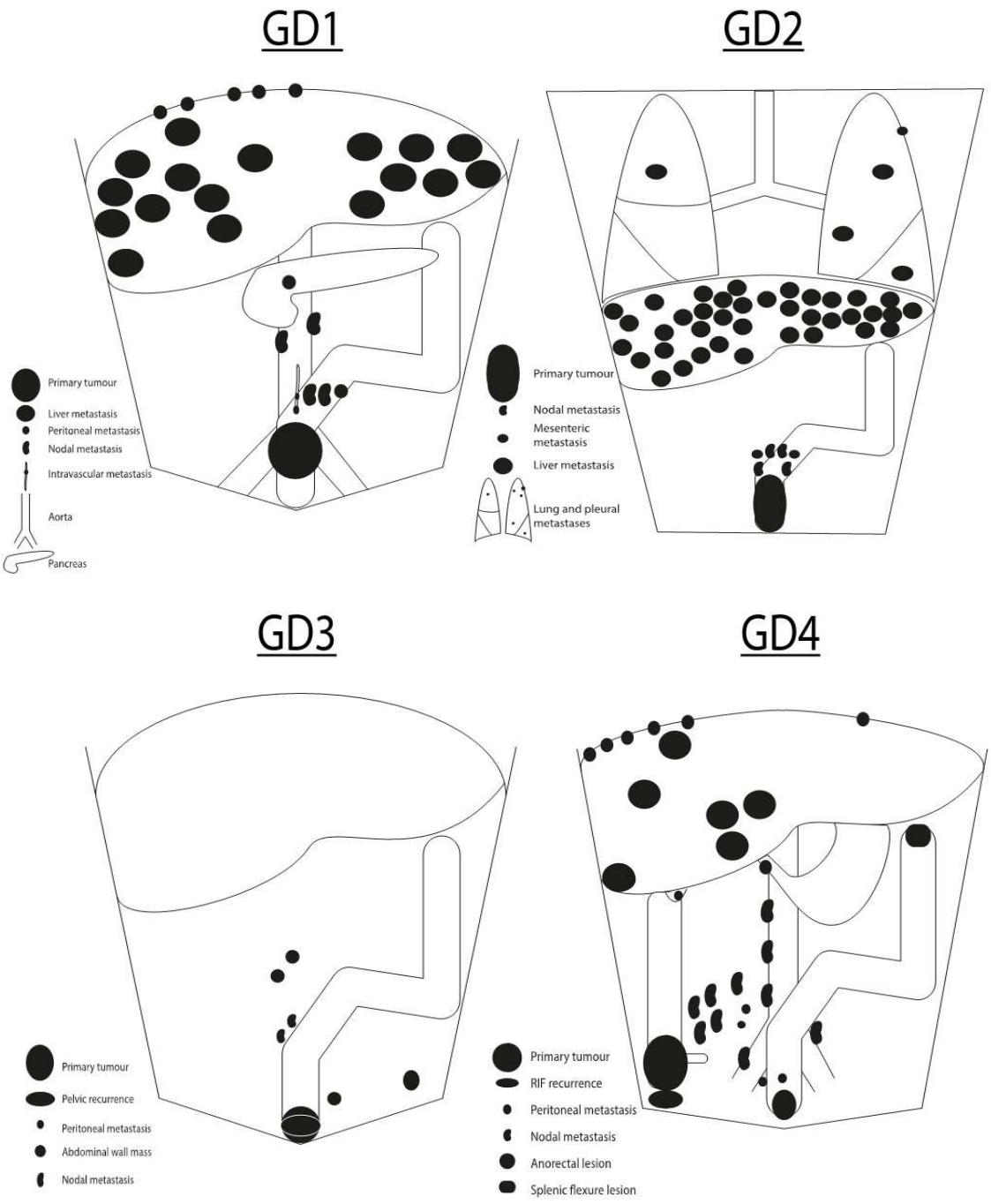


Figure 5: Diagram showing distribution of sampled tumour deposits in GD1-4. Accompanying each body diagram is a key showing the types of deposit identified in each case, the primary tumour overlies the colon in each diagram.

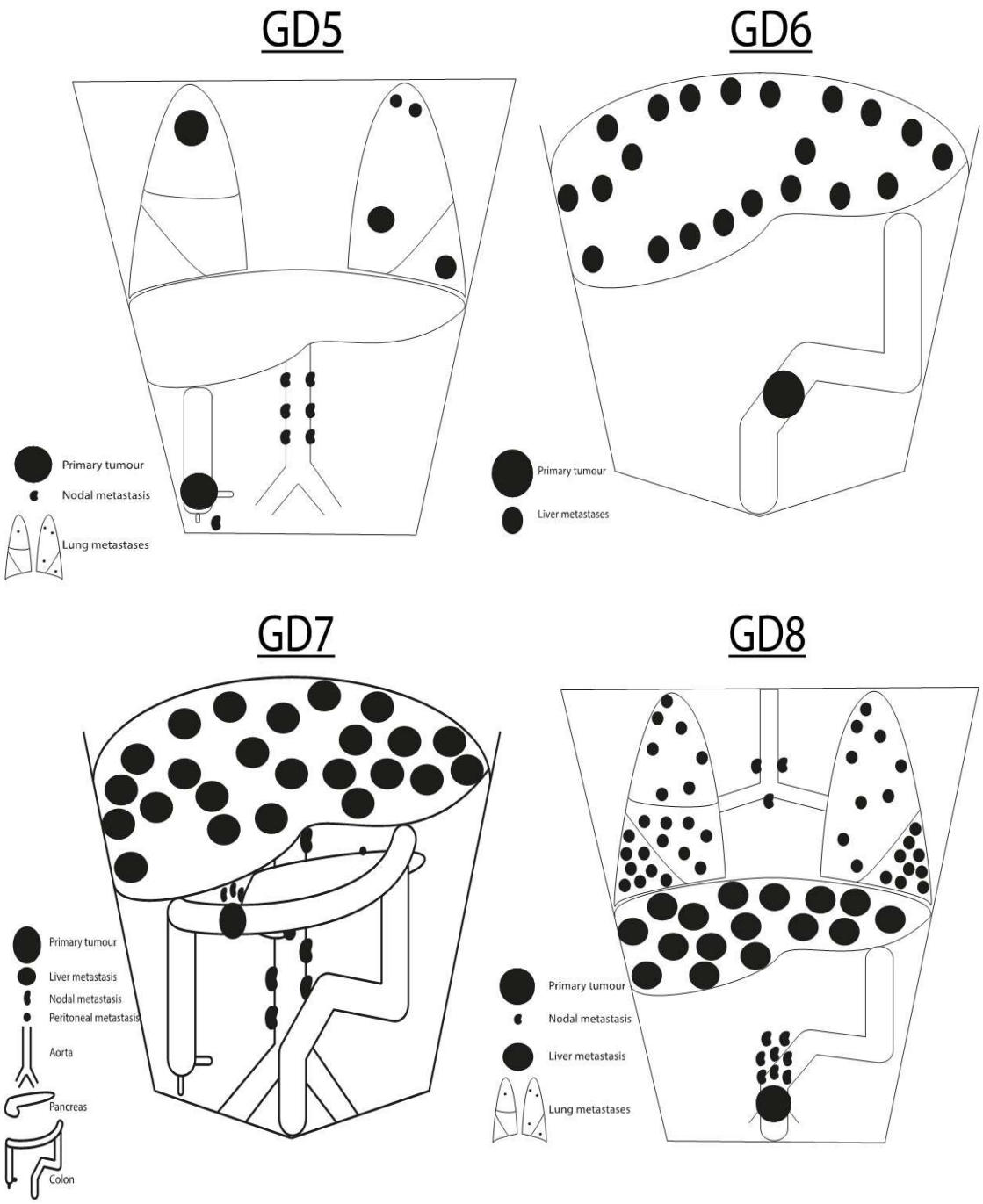


Figure 6: Diagram showing distribution of sampled tumour deposits in GD5-8. Accompanying each body diagram is a key showing the types of deposit identified in each case, the primary tumour overlies the colon in each diagram

Table 5: Table showing the number of tumour deposits and samples taken from each location

	GD1		GD2		GD3		GD4		GD5		GD6		GD7		GD8	
	No. of Lesions	No. of samples	No. of Lesions	No. of samples	No. of Lesions	No. of samples	No. of Lesions	No. of samples	No. of Lesions	No. of samples	No. of Lesions	No. of samples	No. of Lesions	No. of samples	No. of Lesions	No. of samples
Primary	1	3	1	13	1	7	1	8	1	9	1	12	1	8	1	3
Recurrence	0	0	0	0	1	10	1	5	0	0	0	0	0	0	0	0
Mesentery	7	7	6	6	3	5	7	8	0	0	0	0	2	3	8	8
Peritoneum	6	6	0	0	2	6	12	17	1	1	0	0	6	9	0	0
Retro-peritoneum	3	3	0	0	0	0	4	4	6	6	0	0	4	4	0	0
Liver	18	18	43	51	0	0	6	9	0	0	17	20	24	26	17	17
Lung	0	0	4	5	0	0	0	0	5	13	0	0	0	0	43	43
Total	35	37	55	75	7	28	31	51	13	29	19	33	37	51	69	71

3.4.4 Histological staining

The histological features demonstrated by H and E staining were similar for all cases examined and are summarised in Table 6. Despite a variable degree of post mortem autolysis, it was possible to confirm all cases were conventional-type colorectal adenocarcinoma. The majority of cases were a mix of moderate and poorly differentiated carcinoma; only GD4 displayed predominately poor differentiation within the primary and metastatic lesions.

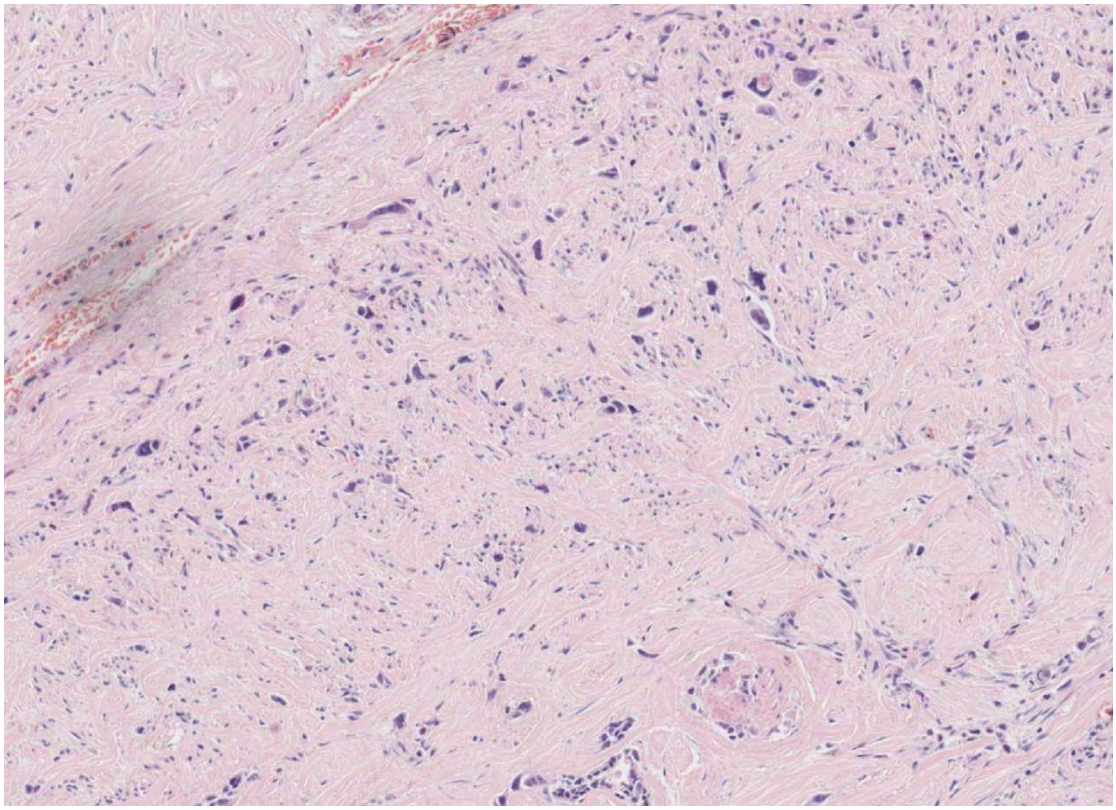


Figure 7: Photomicrograph showing area of poorly differentiated tumour from GD1

The primary tumour from GD2 contained mucin lakes (as displayed in Figure 8), which formed a minor component of the lesion; the tumour was therefore a conventional adenocarcinoma with mucinous areas. A significant mucinous component was not identified in any of the metastases in this case.

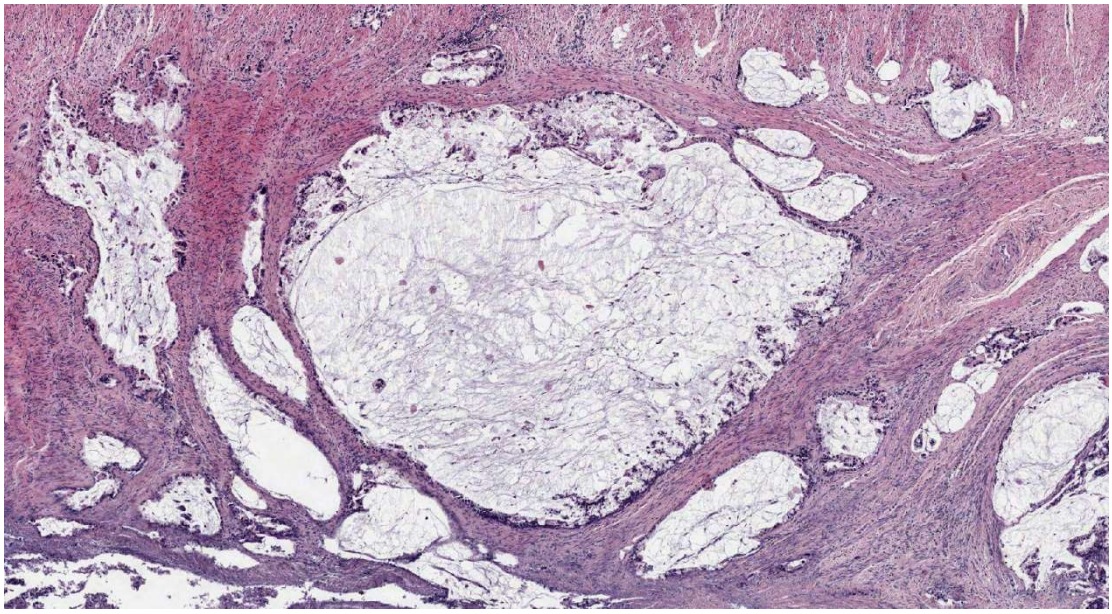


Figure 8: Photomicrograph showing mucinous area from the primary tumour in GD2

It was also possible to identify the presence of vascular and perineural invasion in all cases with the use of S100 and elastin staining (as displayed in Figure 10); GD6 was the only case not to display nodal metastases but did show extensive vascular invasion. Within some blocks it was possible to identify more than one mode of spread, for example sample 63 from GD1 Figure 9 shows a nodal deposit with adjacent vascular invasion.

Table 6: Table showing the features demonstrated by light microscopy in each case examined

		GD1	GD2	GD3	GD4	GD5	GD6	GD7	GD8
Histological type		Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma
Differentiation of primary tumor		Moderate / poor	Moderate / poor with mucinous areas	Moderate / poor	Poor	Moderate / poor	Moderate / poor	Moderate/poor	Moderate/poor
Differentiation of metastatic deposits		Moderate / poor	Moderate / poor	Moderate / poor	Poor	Moderate / poor	Moderate / poor	Moderate/poor	Moderate/poor
Modes of spread identified	**Vascular**	Present	Present	Present	Present	Present	Present	Present	Present
	Nodal	Present	Present	Present	Present	Present	Absent	Present	Present
	Peri-neural	Present	Present	Present	Present	Present	Present	Present	Present
	Trans-coelomic	Present	Absent	Present	Present	Present	Absent	Present	Absent

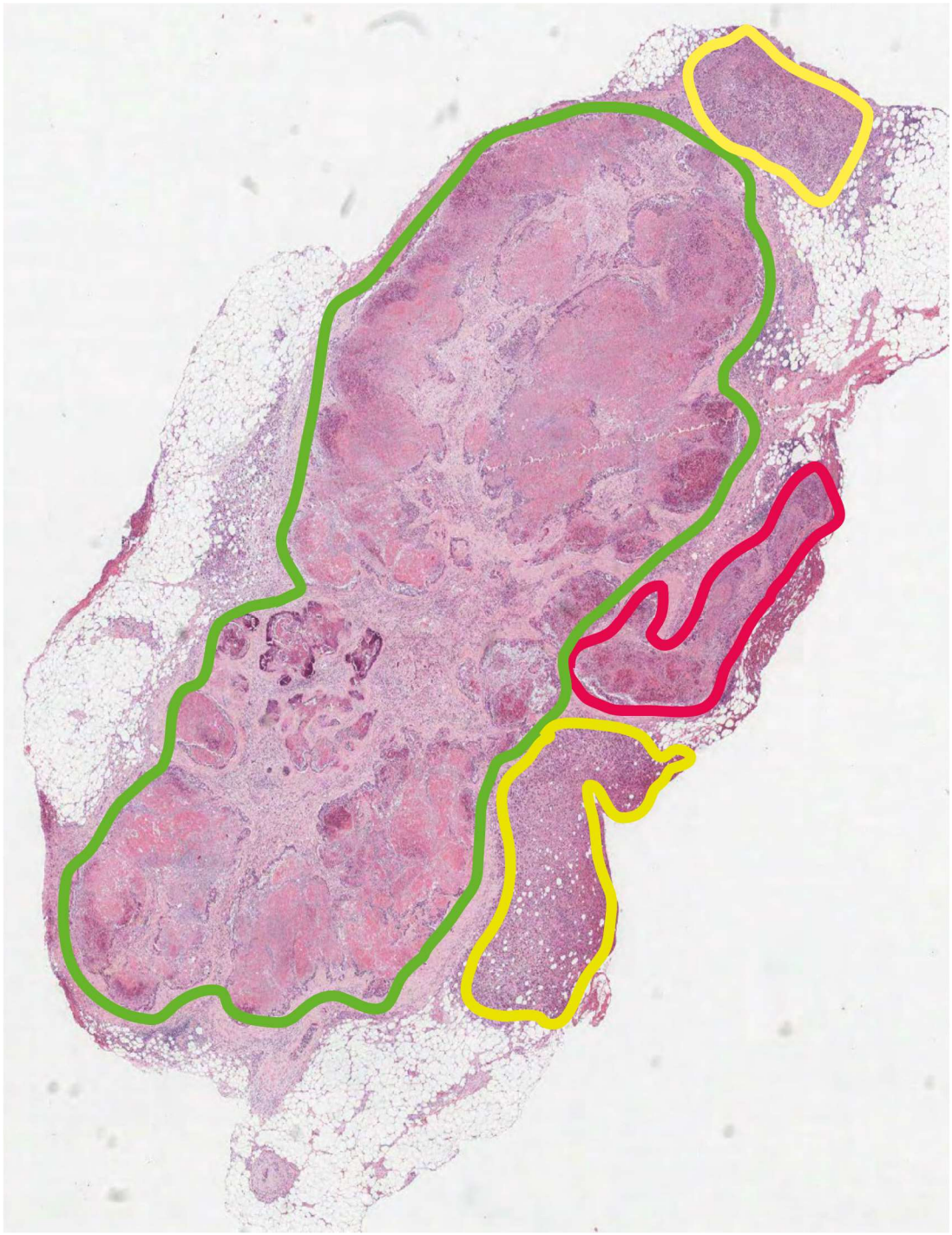


Figure 9: Photomicrograph displaying sample 63 from GD1. Within this image the intranodal tumour is outlined in green, the intravascular tumour is outlined in red and the extra-nodal tumour is highlighted in yellow.

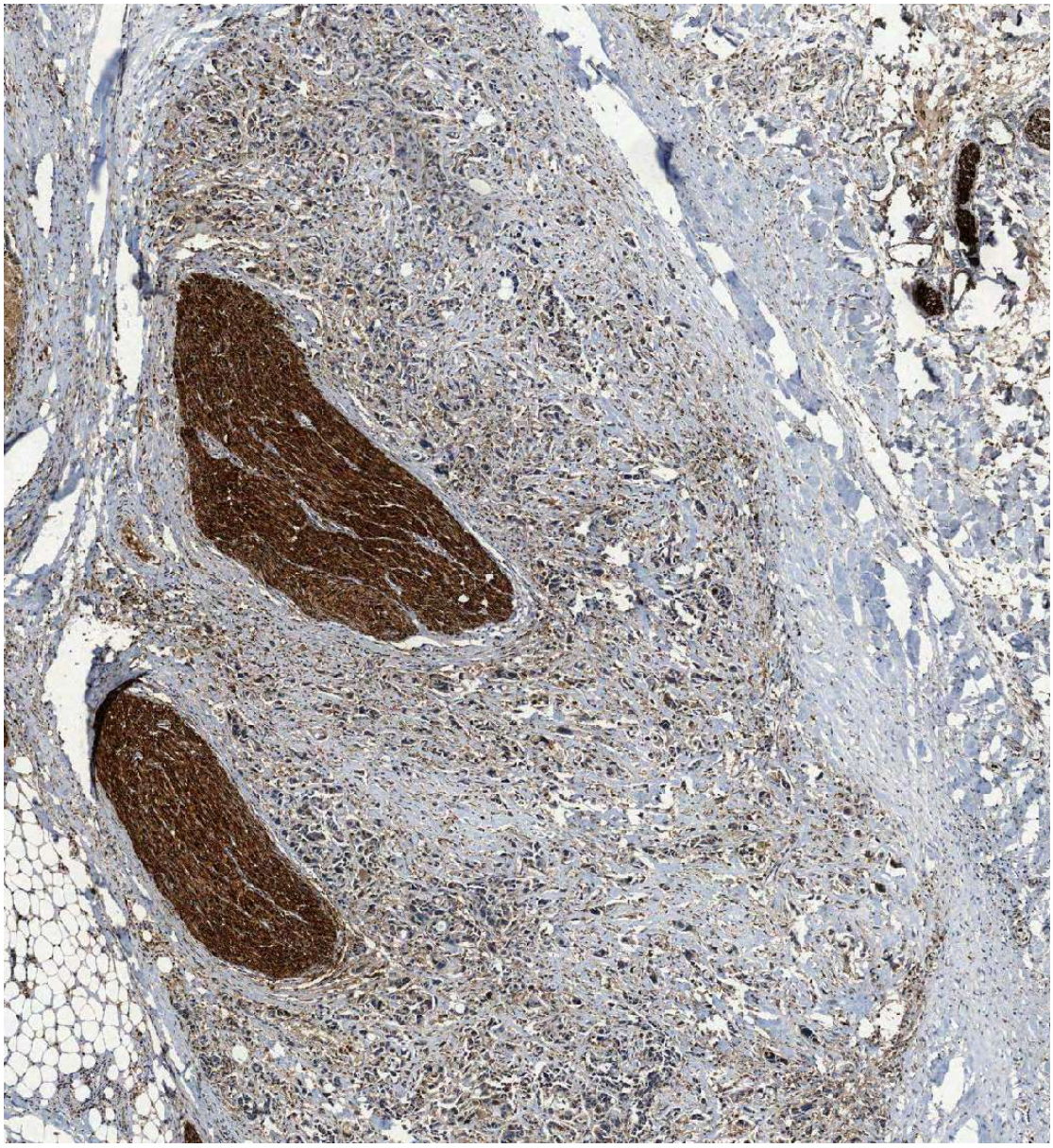


Figure 10: Photomicrograph showing tumour infiltrating the perineural space (S100 stain)

3.4.5 MMR status

Two sections, one from the primary two and the other from a well-preserved metastasis, from each case were successfully stained for MLH1, PMS2, MSH2 and 6; the results are summarised in Table 7. Six of the cases examined appeared to be MMR proficient, whereas GD3 and GD4 did not express MLH1/PMS2 and MSH2/6 respectively; these patterns of staining are those associated with MMR deficiency and are shown in figure 11 and 12.

Table 7: Table displaying the expression of MLH1, MSH2, MSH6 and PMS2 in the eight autopsy cases

	MLH1	MSH2	MSH6	PMS2	MMR deficient?
GD1	+	+	+	+	No
GD2	+	+	+	+	No
GD3	-	+	+	-	Yes
GD4	+	-	-	+	Yes
GD5	+	+	+	+	No
GD6	+	+	+	+	No
GD7	+	+	+	+	No
GD8	+	+	+	+	No

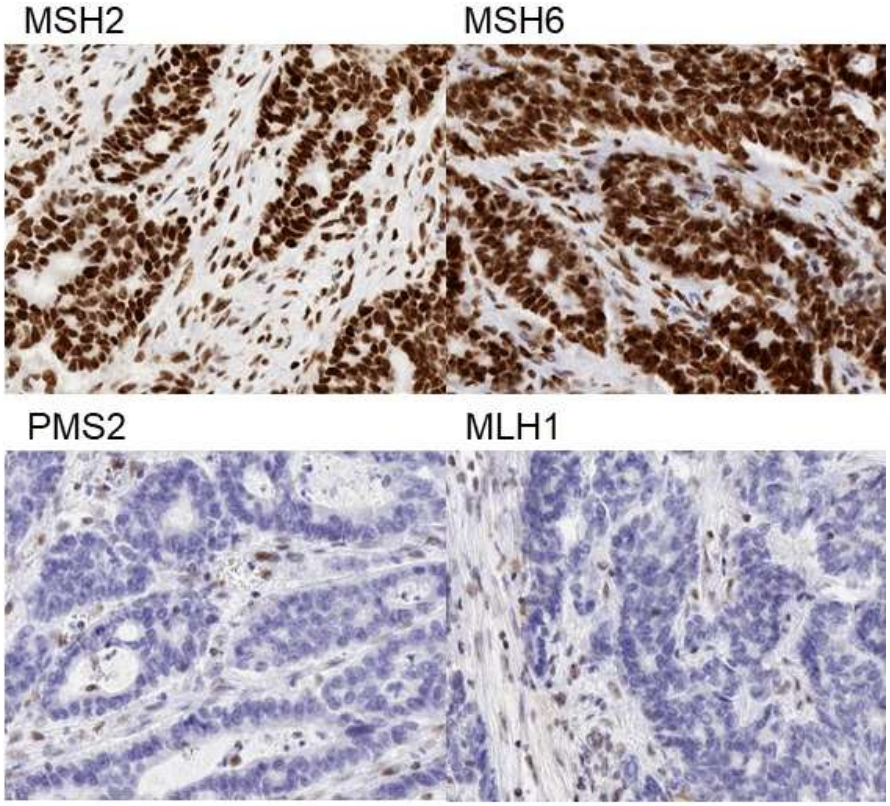


Figure 11:Photomicrographs showing immunohistochemistry from GD3 demonstrating loss of PMS2/MLH-1 and retention of MSH2/6 staining

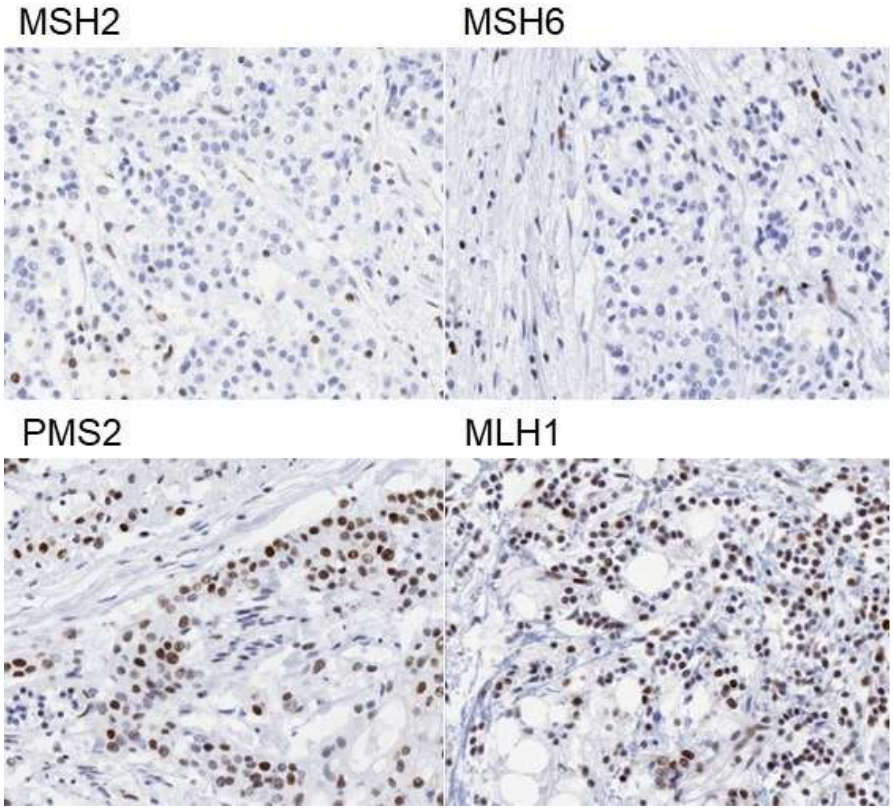


Figure 12: Photomicrographs showing immunohistochemistry from GD3 demonstrating loss of MSH2/6 and retention of MLH1/PMS2 staining

3.5 Discussion

We have performed autopsy examinations on eight individuals with advanced CRC, which displayed a wide variety of clinical, macro and microscopic features.

A dominant feature of the cohort was the advanced stage at presentation, only three of the cases were deemed operable at presentation and the patients who did undergo surgical resection were shown to have local nodal metastases, GD3 also had metastatic disease within the abdomen. This group is therefore representative of the, approximately, 40% of CRC patients who present with stage 3 disease or worse [253]. As would be expected in a group of patients with high stage disease (along with other adverse clinical and pathological features such as EMVI) only one of the eight donors survived more than 15 months from the date of diagnosis. The advanced nature of the disease at presentation is also reflected by the fact that only two patients received systemic chemotherapy; the chemotherapy received (FOLFOX/FOLFIRI) was in keeping with the current NICE guidelines [130,254]. One donor died with only locally recurrent and pelvic disease (GD3), this individual had significant comorbidities and had not received adjuvant treatment at the time of resection or palliative chemotherapy due to this complicating factor. It is interesting to note that, although the pathological stage of the primary tumour in GD3 was more advanced than either of the other two resected tumours, the patient only experienced relatively limited recurrent disease.

The distribution of disease observed is in keeping with that which would be expected in CRC i.e. a predominance of intra-abdominal disease with bulky liver and nodal metastases. Three cases also showed lung metastases, synchronous lung metastasis are present in approximately 11-12% of CRC, whereas approximately 5% suffer metachronous lung metastases [255–257]. Lymph node metastases were present in all but one case (GD6); the absence of lymph node disease and presence of bulky liver metastases in this case supports the suggestion that regional lymph node and distant metastases do not necessarily occur sequentially or from the same regions of a primary tumour [258].

Histologically the cases were fairly uniform, all the primary tumours displayed areas of poor differentiation, which was appreciable despite post mortem autolysis. The use of tinctorial and immunohistochemical staining demonstrated the presence of perineural and vascular invasion in all eight cases. The only slightly outlying cases were GD2 which

showed focal mucinous differentiation and GD4 which was predominantly poorly differentiated. All eight cases displayed virtually all the histological features associated with poor prognosis according to the RCPath guidelines [196,222,240,259–261] and special stains allowed classification of each lesion.

Within the three cases which underwent resection (GD3, 4 and 8), two (GD3 and 4) experienced local tumour recurrence; tumour recurrence at the site of previous resection carries a poor prognosis across all subtypes of CRC [262,263]. Both locally recurrent tumours in our cohort displayed an immunotype suggestive of dMMR which is associated with hypermutator-type tumours. The presence of two dMMR cases is slightly surprising in this cohort, as dMMR cancers have been associated with good prognosis[264,265]; it is postulated this is due to endogenous immune mediated tumour suppression triggered by the expression of neoantigens expressed on tumour cell surfaces formed by the large number of allelic changes characterizing this tumour subtype. This traditional viewpoint is, in part, contradicted by the consensus classification of CRC, which found that the hypermutator, CMS1 group of tumours, into which dMMR CRC fall, tend to show a poor prognosis following recurrence[50], GD3 and 4 therefore conform to this observation. Additionally GD3 was atypical of dMMR CRC as the primary tumour arose within the rectum, the majority of tumours of this type effect the proximal colon [61].

The immunohistochemical profiles present in the dMMR cases are also worthy of comment. GD3 showed loss of PMS2/MLH1 expression, which is a common pattern of loss in sporadic dMMR CRC, whereas GD4 demonstrated loss of MSH2/6 antigens, a profile much more closely associated with germline mutations of MSH2 mutation resulting in hereditary non-polyposis colorectal cancer (HNPCC) or Lynch syndrome [266]. As displayed in Table 4, the donor, GD4, was aged 79 at the time of diagnosis and it would seem improbable that a gentleman of this age (whose only relevant past medical history was prostatic carcinoma treated with radiotherapy in the donors seventies) would be suffering from a hereditary cancer syndrome which has a mean age of presentation of 45 years of age[267]. Although this may in fact be the case, the alternative explanations would include a rare somatic mutation or hypermethylation of either MSH2 or 6 or repeated failure of IHC (however the protocols were repeated for all 4 antigens in GD3 and 4 producing identical results); a further possibility is the presence of a germline MSH6 mutation, which is an uncommon observation and has been associated with a more attenuated presentation of HNPCC[267]. The presence of somatic and germline mutations within all eight cases is examined in subsequent sections.

To further examine the heterogeneity between and within the cases we have examined the genomic landscape via several methodologies. Firstly, pyrosequencing was performed to determine the mutation status at the loci used to predict therapeutic response to EGFR blockade and subsequently Next Generation Sequencing (NGS) was used for a broader examination of the genomics of each case.

4 Targeted mutational sequencing of therapeutically significant loci in disseminated colorectal cancer

4.1 Introduction

As described in the introduction the significance of EGFR related pathways is well-established in CRC, as such monoclonal antibody therapy blocking ligand binding to these receptors have been incorporated in the neo-adjuvant and palliative chemotherapy for CRC[130,268]. The use of these therapies has been refined by the identification of predictive biomarkers, principally mutations in *KRAS* codons 12, 13, 61 and 146, *NRAS* codons 12, 13 and 61, *BRAF* codon 600 and *PIK3CA* codons 542, 545, 546 and 1047 [8,145]. Within clinical practice several methods have been employed to identify mutations at these hotspots; these assays are described below and employ a range of techniques which confer differing degrees of sensitivity and specificity [269].

4.2 Mutational Analyses Guiding EGFR therapy

4.2.1 Sanger sequencing

The oldest, currently employed, method for DNA sequencing is the "chain termination" technique introduced by Sanger [269]. This method involves running DNA synthesis reactions for the same fragment of DNA in 4 wells each containing a different, labelled dideoxy-ribonucleotide form of one base; the addition of this base will terminate DNA synthesis. The terminating base will be present in each well at a low concentration amongst deoxy-ribonucleotide bases and will therefore only be incorporated in a small minority of additions of each base. This produces fragments of different lengths representing different additions of each base; each well can then be run on an electrophoretic gel and the position of each base maybe determined. The major drawbacks of this technique are speed, sensitivity [270] and cost; whole genome sequencing using the Sanger technique takes up to six months and 12 million dollars [271], the newer NGS platforms can produce 16 whole genome sequences in 3 days hours at a cost of slightly over one thousand dollars per genome [272]. Sanger sequencing is only reliably able to detect mutations present in at least 15% of the DNA analysed [270]. This threshold (or 'variant allele frequency' (VAF)) for detection is adequate to detect germline genomic changes or those within a well preserved, relatively pure tumour sample.

However, significant mutations in clinical samples may be heterozygous or even subclonal events and clinical samples are often mixtures of 'target' tumour DNA and DNA from background normal tissue. Newer sequencing techniques are quicker, cheaper, more sensitive and therefore more useful in the clinical setting [273]. At the time this project was started, nationally the most established and widely used technique used for mutation detection in CRC was pyrosequencing [273].

4.2.2 Pyrosequencing

Pyrosequencing is a sequencing "by synthesis" method, which exploits the release of inorganic pyrophosphate (PPi) by DNA polymerase upon base addition during DNA synthesis. PPi in the presence of adenosine triphosphate (ATP)-sulfurylase and adenosine monophosphate (AMP) produces ATP; this is used by luciferase to oxidase luciferin producing light [274]. Using these basic steps, the synthesis of a DNA strand complementary to a primed, immobilised, template DNA fragment (the loci of interest) is performed; each nucleotide is added in sequence and if that base is incorporated to the complementary strand, light is produced. After the application of each base either a wash step can be performed or a fourth enzyme (apyrase) can be added to digest unreacted reagents. The fluorescence signals can be detected by eye or, more commonly, by automated detection systems.

According to NICE, at the time this work was performed (2013), pyrosequencing was the most commonly employed of several techniques used to establish *KRAS* status in the UK [273]; these include mutational arrays, high-resolution melting analysis (HRMA) and electrophorectic methods (single-strand conformation polymorphism (SSCP) and denaturing gradient gel electrophoresis (DGGE)).

Pyrosequencing was the most commonly employed [273] as it provides a balance of sensitivity, specificity and cost. Pyrosequencing is able to identify mutations with a VAF as low as 5% [275], which is superior to traditional Sanger sequencing (15-20%) and comparable to HRMA [276]. Although electrophoretic methods may be more sensitive they are technically more complex procedures [277]. Allele-specific PCR is a highly sensitive (0.01%) mutation detection method in which PCR primers are used which will only amplify mutated alleles. The major drawback is a high rate of false positivity, especially when amplifying low frequency mutations, which is the major application of this assay [278].

It is of note that any technique which employs a single primer to amplify regions of DNA via PCR may be vulnerable to error. Although errors introduced during the PCR cycle are infrequent [279], the power of PCR to amplify DNA may be problematic if the sample contains only a few copies of a target gene (due to low sample quality or volume). PCR will amplify a single copy of a gene to produce a sequence-able sample; this enriched sample may not be representative of the actual tumour or disease state. As the use of PCR is almost ubiquitous in the preparation of samples for sequencing, the use multiple primers or primers that barcode individual strands of DNA have been produced to sequence low volume samples or even single cells and provide more reliable estimates tumour copy number within the original sample and VAF [280–282]. As these newer techniques require significant financial and informatic input they have, until recently, been unsuitable for clinical use.

Pyrosequencing provides the sequence of any identified mutation; this is an advantage compared to other commonly used techniques (except for Sanger sequencing) used for extended *KRAS* testing [273]. Array techniques may not include rare mutations therefore giving false negative results; electrophoretic and melting analyses do not document the exact base change when a mutation is identified, missing potentially important information.

Since the publication of the NICE overview of extended *KRAS* testing the role of NGS in clinical genetics has expanded significantly [283] ; this technology will be discussed in the following sections.

There are two technical problems identified in the use of pyrosequencing, which limit the application of the technique. Firstly, pyrosequencing is only able to sequence relatively short fragments of DNA (it produces a 'read length' of up to 100 bases); this problem arises due to accumulation of degradation products from each base addition and dilution from iterative addition of bases (despite the volume of fluid being ~200 nl). This limitation allows only a small portion of the genome to be analysed in each assay and separate assays will usually be required to examine different genomic loci, even if they lie within the same gene. Therefore pyrosequencing is inappropriate [284] for analysis of large genes which lack well-defined mutational hotspots, such as the breast cancer 1, early onset gene (*BRCA1)*.

Secondly, pyrosequencing can also be prone to errors when sequencing regions of the genome containing homopolymeric tracts (sequences of DNA containing repeats of the

same base). These errors arise as signal production during pyrosequencing and number of bases incorporated at each base addition do not share a linear relationship, thus it can be difficult to assess how many bases have been incorporated consecutively. The presence of these regions therefore limits the applicability of pyrosequencing to some genes and associated disease states such as the cystic fibrosis transmembrane conductance regulator gene (*CTFR*) in cystic fibrosis [285,286].

In summary, when correctly applied, pyrosequencing is a relatively, sensitive and specific sequencing technique, which may be used to identify a limited number of mutations in a focused manner. We have, therefore, chosen to examine the tumour samples taken from the 'Gift' autopsies described in section 3 using pyrosequencing for the most common *N/KRAS*, *PIK3CA* and *BRAF* mutations. This will demonstrate whether a commonly used, current technology is able to demonstrate clinically important intra or inter-tumoural mutational heterogeneity. If this were the case, it would be reasonable to conclude that testing of recurrent disease or multiple metastatic sites maybe beneficial in the treatment of MCRC.

4.3 Aims of this section

1) To assess the presence of intra and intertumoural heterogeneity in mCRC using pyrosequencing to assess the following genomic loci:

   KRAS codons 12, 13, 61 and 146

   NRAS codons 12, 13 and 61

   PIK3CA codons 542, 545, 546 and 1047

   BRAF codon 600

4.4 Materials and methods

4.4.1 DNA extraction

DNA for pyrosequencing analysis was extracted from 5 x 10 µm FFPE sections, extraction was repeated with 10 sections if initial extraction failed to obtain 200 nanograms (ng) of DNA (according to Fluroskan Ascent Microplate Fluorometer measurement (Thermo Fischer Scientific, Loughborough, UK).

De-waxed sections were macrodissected according to regions highlighted on H and E staining as shown in Figure 13.

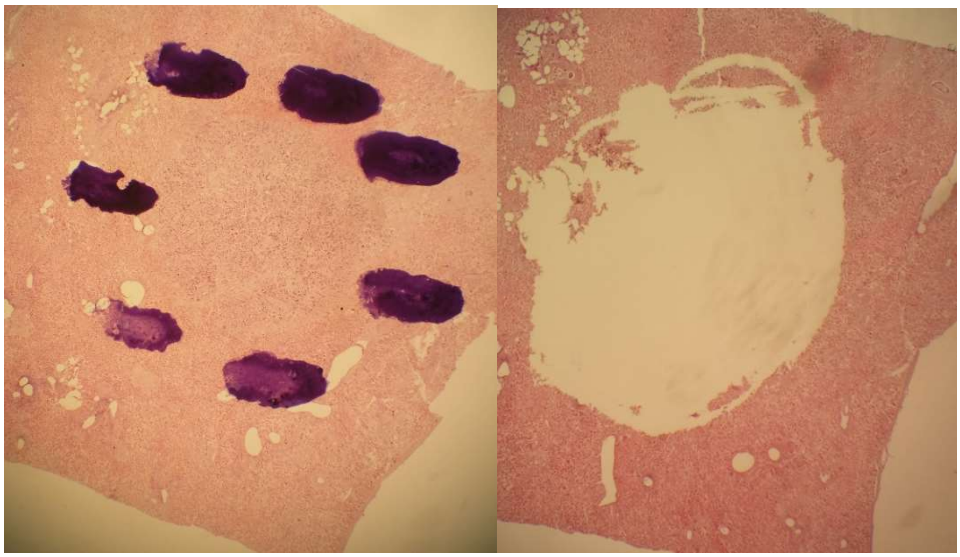


Figure 13: Photomicrograph showing microdissection of liver metastasis

DNA was extracted from the dissected tissue using the Qiagen QiAMP DNA micro kit according to the manufacturer's protocol (Qiagen, Crawley, UK).

4.4.2 DNA Quantification

The concentration of nucleic acid within each sample was initially quantified using a Nanodrop-1000 spectrophotometer (Thermo Fischer Scientific, Loughborough, UK). This measurement was used to dilute the extracted samples for dsDNA quantification using the Quant-iT dsDNA Assay Kit (Thermo Fischer Scientific, Loughborough, UK) and Fluroskan Ascent Microplate Fluorometer (Thermo Fischer Scientific, Loughborough, UK).

4.4.3 Pyrosequencing

The primers used for pyrosequencing are listed in appendix 2: pyrosequencing primers.

The Pyromark Q96 ID (Biotage AB, Uppsala, Sweden) platform was used to perform the pyrosequencing according to manufacturer's protocol.

The reagents used for the PCR master mix are listed in Table 8. They were combined with a 2 μl of sample with a nucleic acid concentration of 10ng/μl according to Nanodrop 1000 measurement.

Table 8: Pyrosequencing PCR reagent mix

Reagent	Volume (μl)
Sample DNA (10ng/μl)	2
HotStar Taq mastermix (2x)	12.5
Forward primer (100uM)	0.05
Reverse primer (100uM)	0.05
Additional $MgCl_2$ (25mM)	0.5
Molecular Biology Grade Water	9.9
Total	25

The PCR was performed in a thermal cycler and the conditions are listed in Table 9.

Table 9: Pyrosequencing PCR conditions

Stage	Temperature (°C)	Time (min:sec)	Cycles
Initial denaturation	95	12:00	1
denaturation	94	0:10	40
Anealing	55	0:20	
Extension	72	0:20	
Hold	15	∞	

The data produced by the detection of the fluorescence signal and the accompanying software is presented as a 'pyrogram' and is displayed in Figure 14.

This plot shows the sequence of bases added and the strength of fluorescence produced

by the addition of the base, allowing for identification of any mutations within the short amplicons sequenced during pyrosequencing. This platform also calculates the percentage of reads showing any mutation within a sample. However, due to the impact of PCR, DNA quality (as described above) and other factors such as tumour ploidy, this figure only provides a rough estimate of the genuine VAF and tumour cell percentage within the original sample.

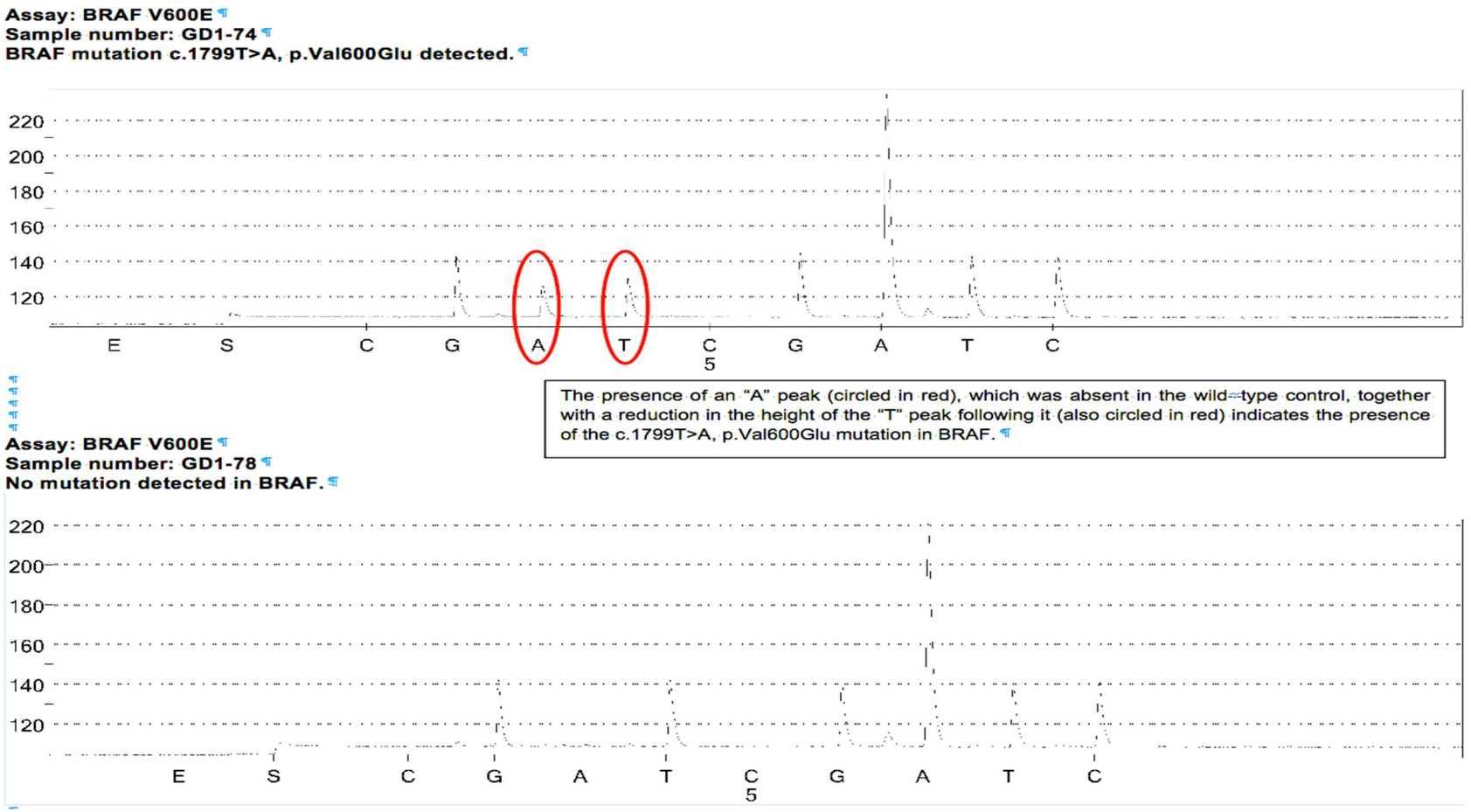


Figure 14: Example pyrograms showing *BRAF* mutation in a tumour sample from GD1 (top plot) as compared to a wild-type plot from the control sample from the same case

4.5 Results

Pyrosequencing for mutations in *KRAS* (codons 12, 13, 61 and 146), *NRAS* (codons 12, 13 and 61), *PIK3CA* (codons 542, 545, 546 and 1047) and *BRAF* (codon 600) were performed and a summary of results is shown in Table 10, complete results are presented in Appendix 3: Pyrosequencing results by sample.

As is shown in Table 10, a clinically significant mutation was identified in each case; *BRAF* codon 600 mutation was present in GD1 and 3, *NRAS* codon 61 mutation was identified in GD6 and the remaining cases bore mutations in *KRAS*. No double mutant cases were identified using this mutational panel. The mutations present in each case were ubiquitous across every sample from that individual, except for three WT samples identified in cases GD2, 4 and 8.

The WT samples in GD4 and 8 were samples taken from lymph node metastases sampled, whereas the WT sample in GD2 originated from the primary tumour within the rectum.

The percentage mutant or variant allele frequencies (VAFs) are presented in Appendix 3 and are summarised in Table 11 and Figure 15. The VAFs for each case were relatively comparable, the mean VAF for all 8 cases was between 22 and 44% although a wide range of VAF was present within most cases. The plots presented in Figure 15 show that the distribution of VAF in each case was either normal or slightly right-shifted. Only GD1 and GD7 showed any significantly outlying results both of which were samples with high VAF.

Table 10: Pyrosequencing results showing number of tumour samples submitted, successfully amplified and the mutation identified in each case.

	Tumour samples submitted	No. successfully amplified	Mutation present	No of samples mutant
GD1	37	37/37	*BRAF* c.1799 T>A	37/37
GD2	70	70/70	*KRAS* c.35 G>A	69/70
GD3	28	27/28	*BRAF* c.1799 T>A	27/27
GD4	51	51/51	*KRAS* c.38G>A	50/51
GD5	29	29/29	*KRAS* c.437C>T	29/29
GD6	32	32/32	*NRAS* c.182A>T	32/32
GD7	50	50/50	*KRAS* c.38G>A	50/50
GD8	71	71/71	*KRAS* c.35G>A	70/71

Table 11: Table showing the range and mean percentage variant allele frequency for mutant samples as demonstrated by pyrosequencing

	GD1	GD2	GD3	GD4	GD5	GD6	GD7	GD8
Range	23-69	18-61	10-34	4-36	13-43	22-51	16-58	5-75
Mean	43	39	25	22	26	38	31	39

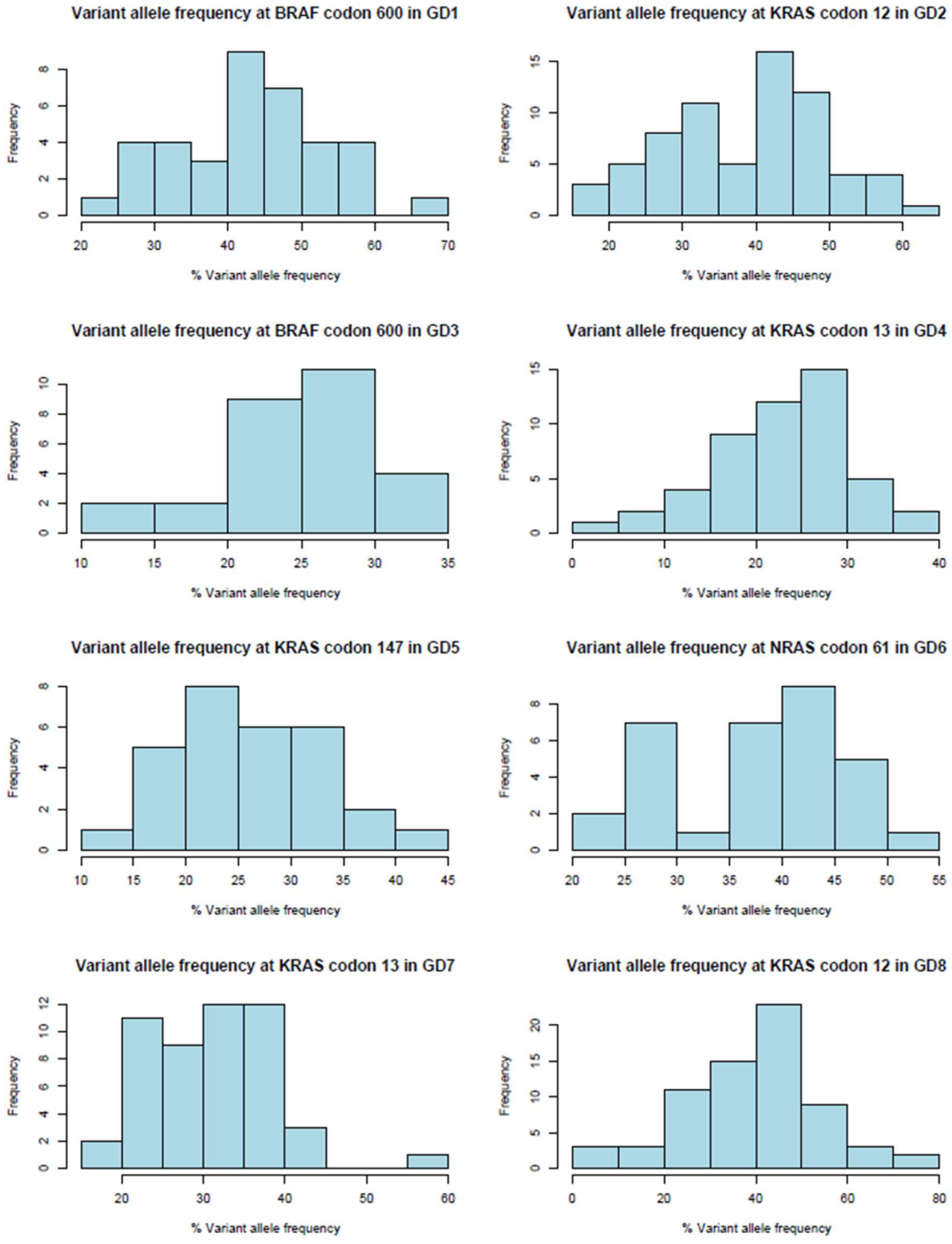


Figure 15: Plots displaying mutant (variant) allele frequencies identified by pyrosequencing in GD1-8

4.6 Discussion

We have performed pyrosequencing on 367 tumour deposits from 8 cases of MCRC demonstrating the presence of a therapeutically significant mutation in each case. Within this cohort *KRAS* mutations within exon 2 (codons 12 and 13) were the most prevalent (being present in 4 of the 8 cases), *BRAF* codon 600 mutations were present in two cases and *KRAS* codon 147 and *NRAS* codon 61 mutation was identified in one case each. The relative frequencies of these mutations are in keeping those identified in genome wide analyses, published as part of the Cancer Genome Atlas Network [65], and those identified as part of large clinical trials of EGFR blockade [8,145]. *KRAS* exon 2 mutations are identified as the most common in both contexts, present in approximately 40% of CRC, with *BRAF* mutations present in approximately 15% of tumours (although they arise in almost half of hypermutator type tumours). Mutations in other regions of *KRAS*, *NRAS* and *PIK3CA* arise in 8%, 4% and 5% of CRC respectively.

The homogeneity of the mutations identified across all lesions in each case is a striking feature of this cohort. Only 3 of the 367 tumour samples sequenced were WT, these were samples from three different individuals and were discordant with the other samples taken from that individual. The high degree of mutational concurrence between primary tumour and metastases has been observed in the largest previous examinations of matched, resected colorectal tumours and metastases; Brannon *et al* [81] showed 100% concurrence of *KRAS*, *NRAS* and *BRAF* in 69 matched CRC primary and metastases, Goswami *et al* [82] were able to demonstrated concurrence of mutational status in 99% of 265 CRC tumour pairs at the loci examined in this cohort. Interestingly both of these reports documented a higher rate of new mutations in *PIK3CA*, although no mutations were highlighted by the analyses included in this report. It is therefore reasonable to conclude, from our evidence and that cited above, any *K/NRAS* or *BRAF* mutations present within the primary tumour will be represented in the large majority of metastatic lesions. We have also not identified any new *K/NRAS* or *BRAF* mutant clones within this cohort of individuals; although the low number of patients in this sample weakens the strength of this evidence, this study does include 312 samples taken from 258 tumour deposits and is in keeping with the literature suggesting new therapeutically significant or driver mutations are rare events, at least within treatment-naive CRC.

The non-concurrent samples were present in GD2, GD4 and GD8 and were samples from the primary tumour and two nodal deposits sampled at surgical resection of the

primary colorectal tumour respectively. The WT sample from GD2 (sample 82), originating within the primary tumour, likely represents genuine heterogeneity within the invasive component of the primary tumour, as the remainder of the blocks taken from the primary tumour showed mutant allele frequencies greater than 20%, suggesting reasonable tumour cell content within the tumour and the histological sections taken from this sample display reasonably well preserved adenocarcinoma within the submucosa (as displayed in Figure 16). The presence of heterogeneity in *KRAS* mutation status as identified by pyrosequencing has previously been documented by Kosmidou *et al* [287]; this group compared the centre and periphery of 75 CRC and in 19 lesions identified non-concurrence *KRAS* status between different regions of the tumour. The presence of intratumoural heterogeneity between separate regions of 6 rectal tumours was also investigated with a more in-depth approach using NGS and SNP array techniques by Hardiman *et al* [288]; this work also identified intratumoural heterogeneity but documented a wide range of heterogeneity using a more broad based approach than employed within this section.

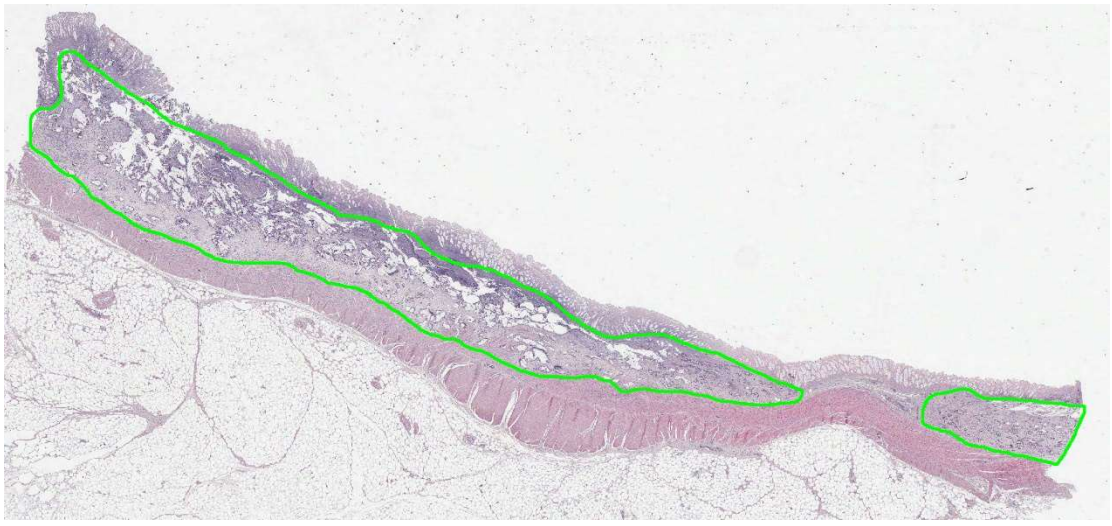


Figure 16: Photomicrograph of sample 82 from GD2 with the region of submucosal tumour marked in green

The two WT samples from GD4 (sample A6) and GD8 (sample A20) were both tumour deposits within local lymph nodes, which were sampled during resection of the primary tumour prior to the development of disseminated disease in either case. Local lymph nodes represent the most common form of metastasis (being present in approximately one third of of newly diagnosed CRC nationally [289]) and as such it is possible that these WT tumour deposits may represent very early metastases arising prior to the

development of *KRAS* mutation or a metastasis from an unsampled *KRAS* WT portion of the tumour. A further explanation for these non-concurrent samples is the loss of genetic material at this locus, so-called “loss of heterozygosity” however pyrosequencing lacks the capacity to exclude this possibility. Whilst these discordant findings appear anomalous in the context of the larger studies cited previously, meta-analytical data has suggested that there may be a higher rate of mutational discordance between primary tumours and regional lymph node deposits as compared to primary tumours and distant metastases [86]; this data should be viewed with caution as the studies included involved a range of different sequencing modalities and technical protocols and it would appear unlikely that two lesions from two separate cases would lose genetic information in identical loci. Therefore, although the possibility of LOH, early metastasis or metastasis from an unsampled portion of the primary tumour cannot be refuted, it is most likely that as both lesions are small (2mm in GD4 (Figure 17) and 0.77mm in GD8 (Figure 18)) macrodissection of the tumour sample (as described in the materials section) has resulted in low tumour cell content. Pyrosequencing has a sensitivity of approximately 5%, representing the ability to identify a heterozygous mutation present in 10% of cells sequenced and it is possible that in both instances, the tumour cell content of the macro-dissected sample was below 10%. GD8 was submitted for more sensitive NGS for *KRAS* mutation as part of a target capture sequencing panel, these results are described in section 6.

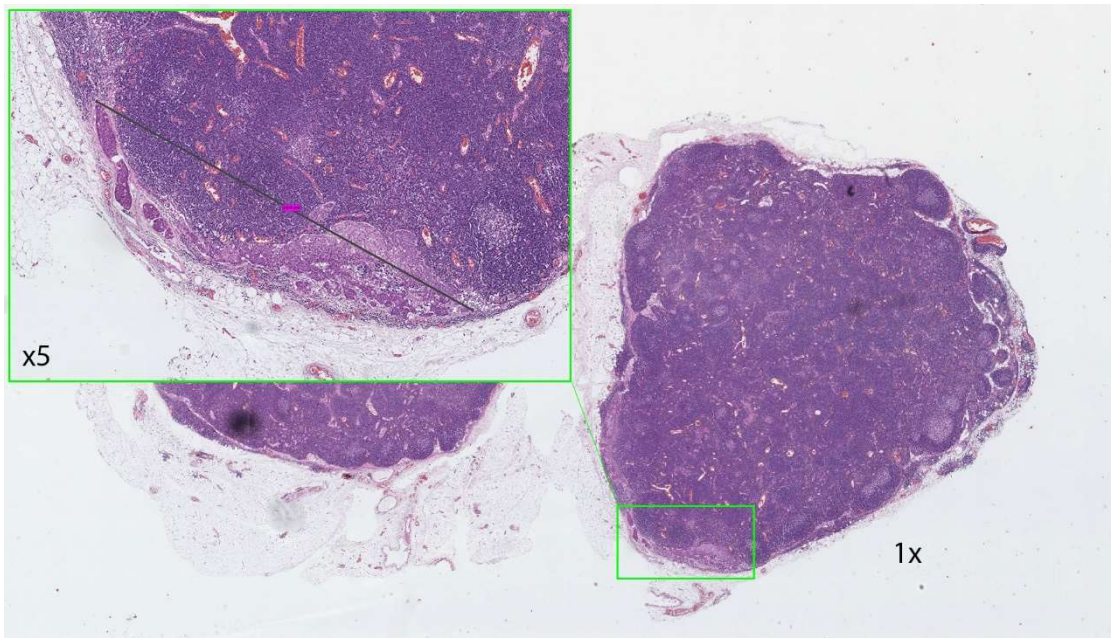


Figure 17: Photomicrograph showing a 2 mm sub-capsular tumour deposit from GD4 which was *KRAS* WT according to pyrosequencing

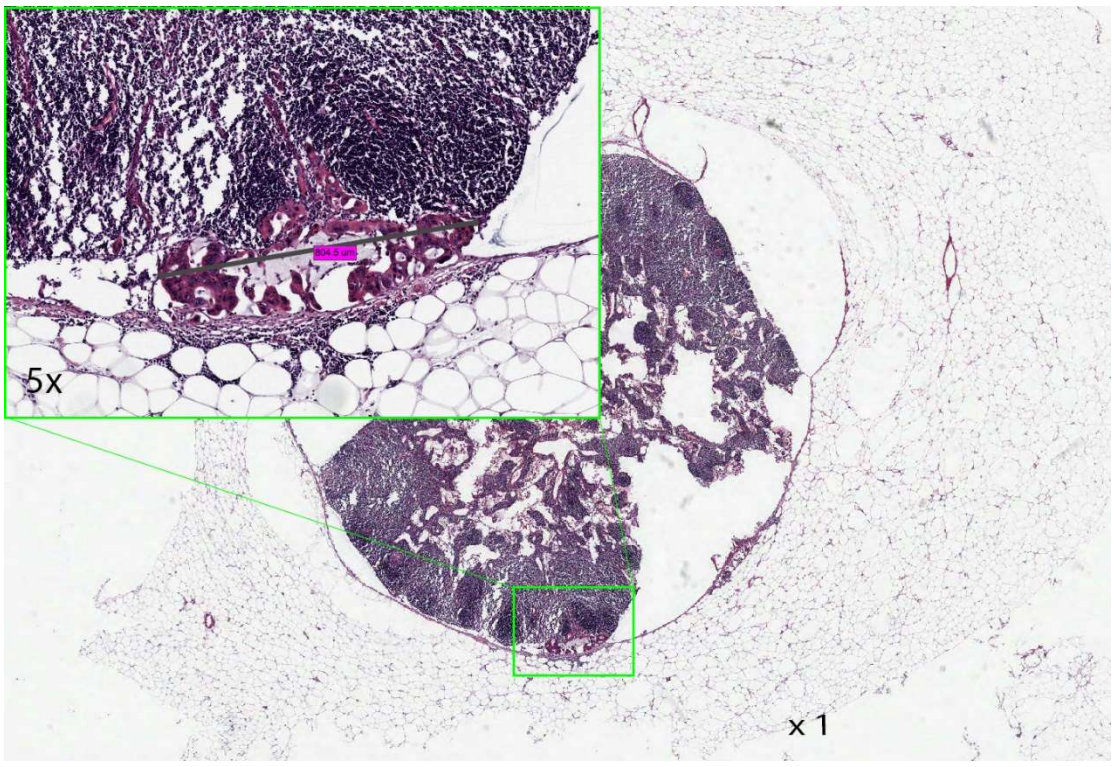


Figure 18:Photomicrograph showing a 0.8 mm sub-capsular tumour deposit from GD8, which was *KRAS* WT according to pyrosequencing

The mean VAF observed in each case presented in Table 11 show a range of mean VAF from 22-43%, which assuming heterozygous mutation in diploid cells, equates to a mean range of tumour cell content of 44-86%; these values would be in keeping with the range of tumour cell content in primary CRC identified as 22-84% by West *et al* [290]. Although the cited work was in primary CRC and therefore direct comparison is not completely valid, it is slightly surprising that these cases fall into the upper range of tumours examined by West *et al*, as this work identified high tumour cell content to be a favourable independent prognostic predictor of survival. However VAF likely only provides an approximation of tumour cell content as at least a large minority of the tumours sampled are not diploid [51,52], especially as it has repeatedly been shown that advanced tumours show a greater degree of aneuploidy than early lesions [51,52,291], but this estimate is in keeping with the published literature. Further analyses such as copy number analysis by NGS or comparative genomic hybridisation (CGH) would be required to establish the degree of aneuploidy in each tumour.

In conclusion we have examined 367 tumour samples from 8 cases of disseminated CRC using a panel of mutations know to be predictive of current targeted therapy. We have identified a clinically significant mutation in each case, these mutations were ubiquitous across all lesions from that individual, with two exceptions, both of which were samples possibly of low tumour cell content. It was, therefore not possible to identify intertumoural heterogeneity with distant metastases using the current panel of mutations by pyrosequencing (the most widely used sequencing modality in *KRAS* mutation detection nationwide [273]), although a minor degree of heterogeneity was present in locoregional disease.

5 Copy number alteration analysis in disseminated CRC by next generation sequencing

5.1 Introduction

As described in section 4 we did not identify convincing evidence of inter-tumour heterogeneity in eight cases of MCRC, using the panel of mutations currently employed to predict the efficacy of EGFR blockade. However, as described in section one the majority of tumours are not characterised by an abundance of point mutations, but by the gain or loss of large portions of or whole chromosomes, so-called 'chromosomal instability' [51,52], also described as copy number alteration (CNA).

The CNAs identified in the development of CRC, were initially documented by meta-phase cytogenetic analysis [292] and have been confirmed and refined by further work, enabled by comparative genomic hybridization [293–296] and subsequently high-throughput sequencing technologies, such as that performed the Cancer Genome Atlas Network (TCGA) [65], which remains amongst the most comprehensive single analysis of CRC genomics to date.

This work identified a spectrum of CNA in CRC with regions of the genome recurrently affected by large and small changes in genetic content. Areas bearing large regions of loss identified were 1p, 4q, 5q, 8p, 14q, 15q, 20p and 22q, in addition to losses of 18q (the genomic region containing the tumour suppressor *SMAD4* and *DCC*) and chromosome 17 (which contains *TP53*); these latter two changes were observed in 66% and 54% of tumours respectively. Smaller focal loses were also recurrently identified affecting tumour suppressor genes in the Wnt and RAS-RAF pathways. A smaller number of regions showing recurrent gain were also identified including 1q, 7p, 7q, 8p, 8q, 12q, 13q, 19q and 20p and q plus focal gains in coding regions of a range of oncogenes including *ERBB2* (which encodes the HER2 receptor), cyclin dependent kinase 8, avian myelocytomatosis viral oncogene homolog (*MYC*) and insulin-like growth factor receptor 2 (*IGFR2*).

As described in the introduction, large chromosomal changes are traditionally thought to arise from abnormalities in the mechanics of cell division, however the process by which focal CNAs (such as described in *MYC* by TCGA) arise is less well characterised. Recent evidence [297,298] suggests that enzymatically driven site-specific alterations in histone or chromatin structure may be responsible for these smaller CNAs. Additionally

it is suspected that these some of these oncogenic events may be driven by changes in the tumour microenvironment such as hypoxia [299] and may in fact be transient. This is in keeping with the observation of adaptive CNA as a physiological phenomenon observed in multiple organ systems in humans and other mammals[300–302].

It would therefore be reasonable to surmise that CNA is likely central to the neoplastic process; as such the occurrence of new CNAs in subclones of a tumour cell population is likely to confer advantageous (as well as deleterious) properties to a tumour cell clone. As a proportion of these lesions are large events (in the context of the genome), they provide a good opportunity to identify the emergence of polyclonality within and between tumour deposits. As discussed in the introduction, several groups [92,94,303] have attempted to identify evolution and polyclonality in metastatic CRC by CNA analysis, however these studies only included limited numbers of resected primary and metastatic CRC, most of which were not from the same individual or 'non-matched'. These previous studies therefore may only reflect the biology of resectable MCRC (rather than more advanced, fatal disease which poses the most important clinical problem) or, in the case of studies including non-matched primary tumours and metastases, differences between the more aggressive metastasising tumours and those which have not spread. The cohort of patients examined during the 'Gift' autopsy project is however an opportunity to examine the spectrum and evolution of new CNAs in the setting of MCRC, as it is the most comprehensive collection of lesions within individuals with disseminated CRC.

5.1.1 Identification of CNA

Chromosomal abnormalities detected as CNA have traditionally been assessed using comparative genomic hybridization (CGH) [293,295,296] and single nucleotide polymorphism (SNP) arrays. The former technique involves hybridising tagged tumour and normal DNA to a normal chromosome in metaphase. During hybridization the tags are released producing a fluorescent signal. The copy number may be determined as the intensity of the signal is proportional to the quantity of DNA in the sample at any point on a chromosome, the normal sample is run in parallel to act as a control for comparison.

SNP arrays function by a similar method to CGH, whereby small, complimentary fragments of DNA ('probes') are formed into an array to which the labeled sample DNA is hybridized, generating a signal (usually fluorescent), the intensity of which equates to the sample copy number at the genomic loci represented by the probes. This technique differs from CGH by virtue of the fact that the probes in SNP arrays may be highly

specific, only hybridizing to a specific mutation, therefore SNP arrays provide mutational as well as copy number data [304].

Massively parallel sequencing has been shown to have comparable or slightly superior ability to detect small deletions as CGH and SNP arrays [305], and with the ability to multiplex up to 96 samples at one time this technology is also cheaper and more versatile than the commercially available CGH/SNP platforms. This sequencing modality has produced a large advance in the volume and quality of sequencing data available for analysis.

5.1.2 Next generation sequencing

Next generation (NGS) or massively parallel sequencing represents a significant improvement upon pyrosequencing and provides the capacity to sequence an entire genome or simultaneously sequence multiple targets from hundreds of individuals [306].

NGS is another 'sequencing by synthesis' method in which DNA is fragmented and undergoes 'library preparation'. Library preparation is the process during which adaptors and index primers are attached or 'annealed' at either end of the fragments. These fragments are amplified by PCR and then inserted into a 'flow cell'. This a glass slide with a number of 'lanes', which are covered in small strands of nucleotides (or 'oligos'), complementary to and therefore bind, a sequence within the adaptor annealed to the sample DNA. Once bound, the sample DNA undergoes amplification or 'cluster generation'; during this stage DNA polymerase is used to produce approximately 1000 copies of each strand.

These strands are then sequenced, all four nucleotides, tagged with different fluorescent probes are added and the base complementary to the template strand will be attached. The attached probe is then excited and the emitted light is recorded. Superficially there is some similarity to pyrosequencing, in that the addition of each base produces light, however NGS directly measures each base addition whereas pyrosequencing infers the number of base additions from the strength of the signal. In this manner up to 10 million clusters are sequenced at one time and the sequence of nucleotides produced by each fragment is called a 'read'. The index primer attached to each fragment during library preparation is also sequenced, identifying the read as belonging to a specific sample.

Once all fragments from a sample are sequenced the adaptor and primer sequences are removed and the reads aligned to a reference genome. When performing genome wide

sequencing there is considerable overlap between fragments, this improves the accuracy of sequencing as if differences occur a consensus can be reached between several reads.

So NGS (and the associated bioinformatics programs) allows fragmented DNA to be sequenced and reassembled, bypassing the limited read length of pyrosequencing and increasing sequencing speed compared to Sanger sequencing. This technology can therefore be used for sequencing the entire genome or the data containing exomes. Alternatively, NGS can be exploited to sequence multiple mutational sequences at one time; only areas of interest are amplified (rather than the whole genome), adaptor and primer tagged and then sequenced. Early work, using a relatively low output Roche 454 Junior sequencer (Roche Diagnostics, Mannenheim, Germany), has confirmed NGS can be as sensitive as pyrosequencing for the detection of *KRAS* and *BRAF* mutations in CRC [307]; higher output platforms allow mutant allele frequencies as low as 0.01% to be identified [308,309]. A high degree of sensitivity is possible as a relatively small amount of DNA is submitted for sequencing and a large number of reads are produced for each region of interest. This high sensitivity may allow for the identification of different mutational clones present within a tumour; this mutational heterogeneity is a potentially important to treatment failure.

In addition to the ability to detect low frequency mutations, by multiplexing multiple samples into a single flow cell, it is possible to perform low-depth whole genome sequencing [90,305,310]. Although this technique is unlikely to provide a sufficient number of 'reads' at any one loci to reliably identify point mutations, the number of reads produced within a larger portion of the genome, for instance 300 kilobases (kb) (the resolution of the Affymetrix Oncoscan copy-number SNP array [311]) is sufficient to identify copy number change with a similar resolution to that described with SNP arrays [90,310].

We have therefore performed CNA analysis by NGS upon each sample taken from the GIFT autopsies described in section 3. Firstly, to detect any change in CNA between and within primary and metastatic CRC and attempt to determine the phylogeny of the lesions present within MCRC by examining the pattern of CNA in each sample. As described in the introduction, the examination of tumour phylogeny by CNA analysis within an individual with multiple tumour deposits is not a standardised process and several, automated bioinformatic approaches have been devised for determining tumour phylogeny from CNA data; either by the overall similarity of genomic changes between

tumour deposits (such as unsupervised hierarchal clustering [91,312]) or by examining the occurrence of shared breakpoints [125].

Hierarchal clustering analysis may be used to group samples together by overall similarity or dissimilarity of the data of CNA, by expressing the 'distance' between two samples, this value is represented in a dendrogram or family tree. Within the context of identifying the phylogeny of lesions by shared CNAs, this approach is inappropriate as it favours large events (i.e. those which produce a large 'distance') over small breakpoints, whereas in phylogenetic terms those events, which are mutually exclusive, large or small, hold the greatest value.

As such we have adopted a semi-automated Dirichletian approach (as described in the introduction) by which, shared copy number breakpoints are used to group the lesions from each case and thereby determine a phylogenetic tree for each individual. A semi automated approach was favoured over a fully automated technique such as TuMult [125], to minimise the potential confounding effect of DNA degradation in the 'Gift' autopsy samples; this issue is particularly pertinent within this collection of samples, as they comprise FFPE material and have most likely undergone a degree of post mortem autolysis (as demonstrated by the histological findings) [313–316]. Additionally, a semi-automated approach provides a greater flexibility to incorporate a larger number of samples and evolutionary steps within each case, TuMult and other automated informatic approaches [117,125,126] are limited either by the number of samples or evolutionary steps which can be incorporated into a phylogenetic tree.

5.2 Aims

1) Assess the presence, frequency and severity of CNA in each of the eight GIFT autopsy cases

2) Determine the phylogeny of each tumour deposit by the presence of shared and private CNAs.

3) Correlate the CNA data with the clinical, histological and mutational data described thus far.

5.3 Materials and methods

5.3.1 NGS for CNA

The library preparation for CNA was performed using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, Massachusetts, USA); this protocol requires 200 ng of dsDNA in 50 µl of molecular grade water.

200 ng of extracted DNA is sheared into 200 base pair (bp) fragments using the Covaris S220 Focused-ultrasonicator (Covaris, Brighton, UK). Sonicators produce high frequency sound that fragments nucleic acid (NA); the intensity and duration of sound can be adjusted to control the size of fragments produced. The settings for the Covaris sonicator are described in Table 12.

Table 12: Settings for DNA shearing with Covaris S220 sonicator

Duty Cycle	20%
Intensity	5
Cycle Burst	200
Time	140 seconds
Temperature	4°C

Incubation with End Repair Reaction Enzyme (New England Biolabs) for 30 min @ 20°C and then 30 min @ 65°C repairs any overhanging, unequal DNA fragments producing blunt ending fragments for the addition of adaptors. The reagent mix for this step is described in Table 13.

Table 13: Reagent mix for end repair reaction

Reagent	ng per well
End Prep Enzyme Mix	3
End Repair Reaction Buffer (10x)	6.5
Fragmented DNA	*55.5*

Adaptor ligation is the next step; this involves the addition of a dsDNA molecule, which allows binding of DNA fragments to the sequencing flow cell and act as a bridge to

primers. This process occurs in the presence of a ligation enhancer, blunt/TA ligase Master Mix and Adaptor for Illumina (New England Biolabs,) whilst incubated for 30 min @ 20°C. The reagent mix is described in Table 14.

Table 14: Reagent mix for adaptor ligation reaction

Reagent	ng/µl For 1 well
Blunt/TA Ligase Master Mix	15
NEBNext Adaptor for Illumina	2.5
Ligation Enhancer	1

The addition of (3ul) USER enzyme (New England Biolabs) and incubation for 15 min @ 37°C exposes a binding site for primer addition.

Agencourt AMPure XP beads (Beckman Coulter, Pasadena, California, USA) are then used to "size select" out the DNA fragments 200 bp in length. These paramagnetic beads reversibly bind to DNA in the presence of an appropriate buffer or 'crowding agent'; the size of fragment selected is reliant upon the ratio of volume of bead solution to sample volume.

An initial addition of beads within the polyethylene glycol (PEG) buffer at a low ratio compared to solution (0.55:1.0) binds any large fragments of DNA (>300 bp). A magnet is then used to remove the beads from the PEG/sample solution, which is placed in a clean tube and another aliquot of bead solution is added increasing the PEG buffer: sample ratio to 0.8; at this concentration all the remaining fragments larger than 200 bp are bound. At this point the beads are placed on a magnet and the solution (containing all fragments <200bp) is discarded. The beads then undergo several washes and are re-suspended in a solution within which DNA-bead binding is reversed; the beads are then removed by binding to a magnet.

A primer for PCR and an index primer are then annealed to the 200-300bp fragments present in the eluted solution; this fragment of DNA is then amplified using the NEB PCR master mix (New England Biolabs, Ipswich, Massachusetts, USA). The PCR primer allows DNA polymerase to bind to the sample DNA and the index primer allows identification of sample DNA during parallel sequencing of multiple samples. This process takes place within the thermal cycler which initially heats the solution to 98°C which "denatures" the dsDNA into single strand DNA (ssDNA) and the temperature is

then lowered to 65°C at which the primers bind to the ssDNA. At 72°C the polymerase present within the PCR Mix produces a ssDNA complementary to the sample DNA with annealed primers ("extension" stage). This process is repeated exponentially increasing the number of 200bp fragments with attached PCR and index primers. The reagent mix is described in Table 15.

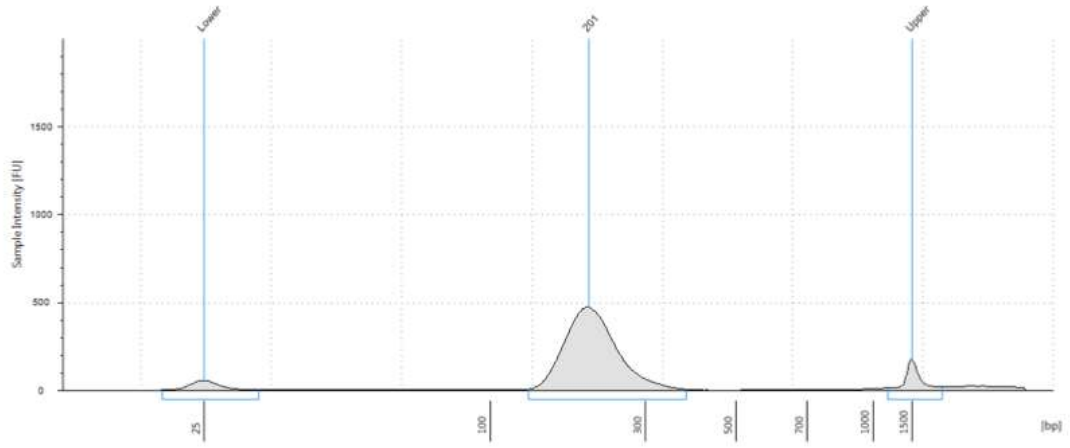
Table 15: Reagent mix for PCR amplification

Reagent	µl per well
NEBNext High Fidelity 2X PCR Master Mix	25
Universal PCR Primer	1
Index Primer	1

A final addition of AMPure XP beads at a ratio of 1.25:1 is then performed to bind all DNA fragments within the solution larger than 200bp. The beads are drawn out of solution with a magnet, the PEG buffer is discarded and the DNA is eluted into 33 µl of buffer EB (Qiagen, Crawley, UK). This final solution is the final "library" upon which sequencing takes place.

5.4 Library preparation quality control

The size of DNA fragments within each the library was confirmed with the Agilent 2200 TapeStation System (Agilent Technologies, Santa Clara, California, United States) and high sensitivity D1K ScreenTape (Agilent Technologies) plus reagents according to the manufacturers protocol. The target fragment was 200bp without significant (<5%) 'adaptor peak'. An adaptor peak is seen as a second peak on the electropherogram at approximately 20bp and represents unannealed adaptor within the library; unbound adaptor will occupy oligos in the sequencing flow cell thereby reducing the sequencing depth of the sample. This electrophoretic system and accompanying software provides an electropherogram and the concentration of any detected peaks (Figure 19).



G12/12_80

Peak Table

Size [bp]	Calibrated Conc. [ng/µl]	Assigned Conc. [ng/µl]	Molarity [nmol/l]	% Integrated Area	Peak Comment	Observations
25	4.24	-	261	-		Lower Marker
201	63.8	-	488	100		
1500	6.50	6.50	6.67	-		Upper Marker

Figure 19: Agilent TapeStation 2200 electropherogram and peak table showing fragment length and library concentration of sample.

Following quantification by TapeStation and Quant-iT dsDNA Assay Kit (Agilent Technologies), 20 ng of up to 96 samples were pooled for sequencing.

5.4.1 Illumina HiSeq NGS and data generation

20 ng of each library was submitted for sequencing on the Illumina HiSeq2500 (Illumina inc., San Diego, California, USA) producing paired 101 bp reads. "Paired" means each DNA fragment is sequenced twice, once in the forward direction and then in the reverse direction; this process improves the quality of read alignment to the reference genome.

Once sequenced the bases corresponding to the adaptors were trimmed using cutadapt [317] and aligned to the human genome (hg38) using 'bwa' [318]. This aligned DNA sequence is then broken down into fragments of approximately 400 kilobases (kb); CNAnorm was then used to calculate any alterations in copy-number whilst correcting for all changes in ploidy and contamination of tumour DNA with normal tissue [310]. The control samples used were microscopically normal tissue sampled from the donor; the pooled control reads were trimmed to match the adaptor trimming for each sample. Regions of amplification and deletion (breakpoints) were identified using DNAcopy [319] to produce a

plot of CNA as shown in Figure 20.

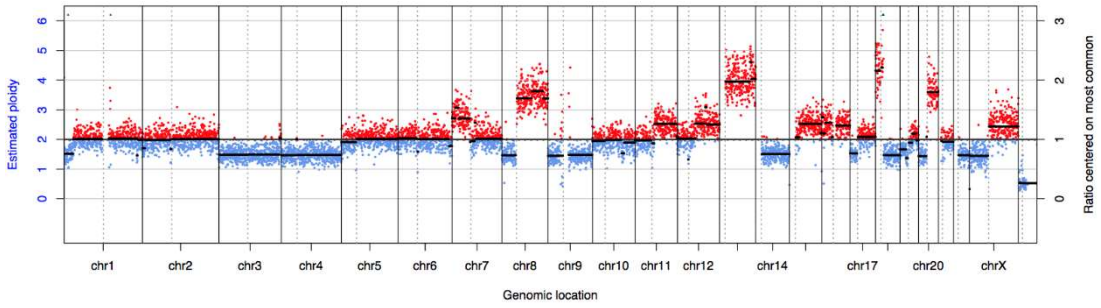


Figure 20: CNA plot showing abnormalities throughout the tumour genome

Approximately 5 to 10 million reads were generated per sample, the equivalent of 0.5-1 mega base (Mb) or 0.16-0.33x coverage.

5.4.2 Mapping tumour heterogeneity

The evolution or phylogeny of metastases was reconstructed by comparing co-occurrence of breakpoints using a method based upon Dirichletian principals previously used to describe the evolution of breast and oropharyngeal neoplasia [320,321]:

Common breakpoints across samples were identified.

Then through correlation with the tumour cell content of each section, the approximate proportion of cells carrying each breakpoint was inferred.

Groups or 'clusters' of breakpoints that occurred in the same samples at the same frequencies were plotted. Events which may have occurred independently such as whole chromosome changes or breakpoints at centromeres and breakpoints, which might have been masked by genomic losses in overlapping regions were taken into consideration.

Each group of breakpoints was interpreted as a tumour subclone and therefore by comparing the presence of subclones in each sample the relationship between each sample was determined. Heterogeneous samples were those, which contained more than two characteristic clusters of changes. A worked example of this method along with the phylogenetic tree for GD1 is displayed in Figure 21.

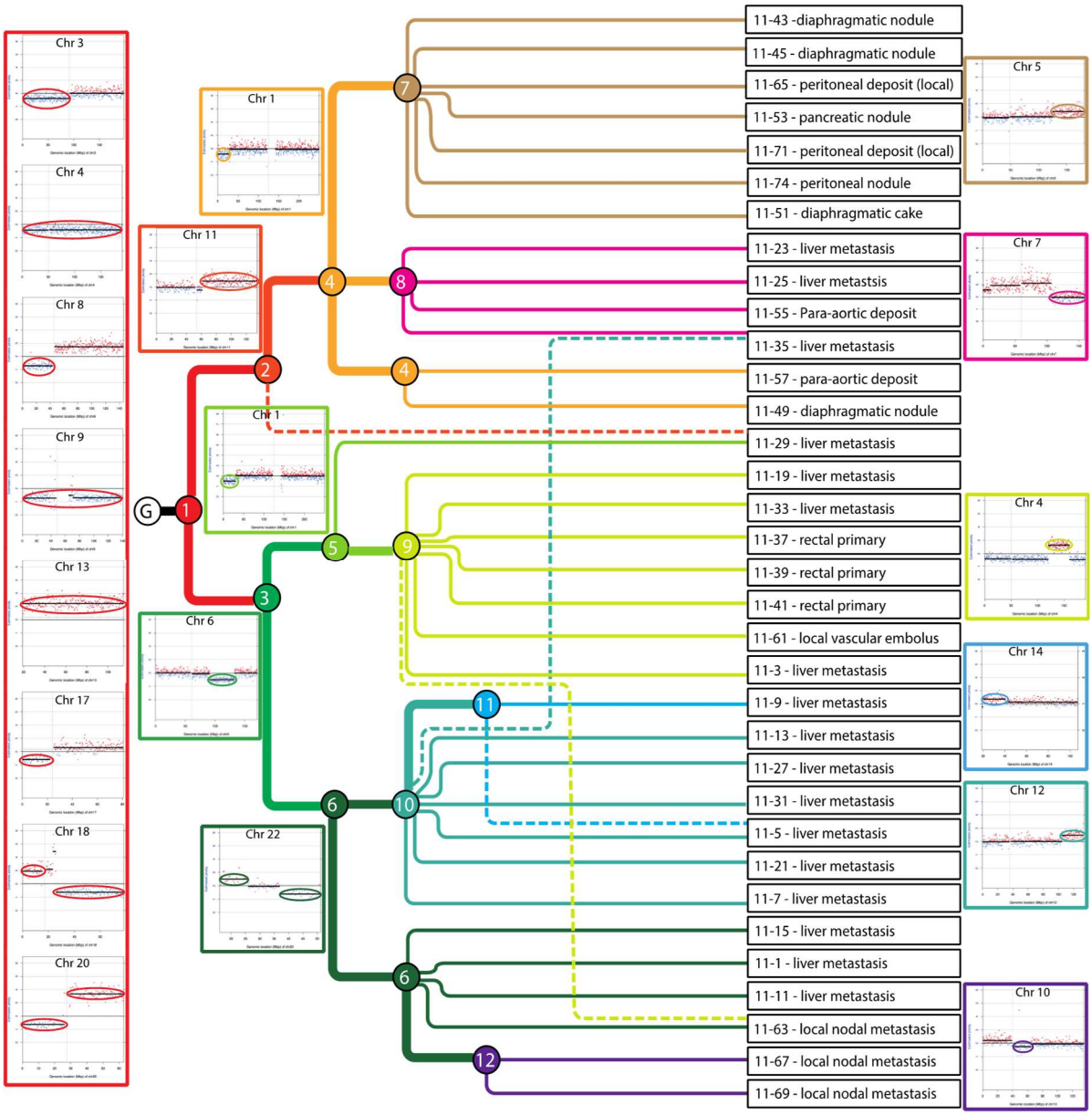


Figure 21: Diagram illustrating the construction of the phylogenetic tree for GD1. This diagram shows the CNAs identified by DNAcopy, which are either shared by all lesions ('truncal' events in the red box on the left) and those, which are only present in a subgroup or 'branch' of the family tree (all other events). The G represents germline events present in the normal tissue. The majority lesions only contained changes typical of branch of a single branch of the phylogenetic tree, however several lesions contained changes typical of multiple branches and were considered polyclonal, for example sample 11-35 contained changes of both cluster 10 in addition to those of 1, 2, 4 and 8. The broken lines within the diagram represent the minor subclone within the polyclonal samples

5.5 Results

5.5.1 Sample numbers and quality control

In total 368 of 376 tumour samples were submitted for CNA analysis by next generation sequencing, matched normal samples were also submitted for each case. 8 samples failed quality control (QC), comprising of 3, 4 and 1 samples from GD1, 2 and 6 respectively. The 5 samples from GD2 and GD6 did not produce a sufficient quantity of DNA for library preparation (despite repeated attempts to extract DNA from up to 10 x 10 µm sections of tissue), the three samples from GD1 produced 200 ng of DNA however they repeatedly failed to produce a library of sufficient fragment size and concentration for sequencing. The samples which did not pass QC were all blocks taken from the primary tumour in GD1 and 2 and a section from a peritoneal deposit in GD6.

All but one of the samples which were submitted for sequencing were successfully sequenced (producing at least 1 million reads per sample), aligned, processed by CNAnorm and DNAcopy and incorporated into a phylogenetic tree using the semi-automated process described in materials and methods. The excluded sample, from GD8, was not included within the phylogenetic tree for the respective case due to low tumour cell content. The number of samples successfully processed per case is presented in Table 16.

Table 16: This table shows the numbers of samples submitted for each 'Gift' case. The left column shows the number submitted over the number taken for each case. The right column shows the number samples included in the phylogenetic analyses

	No. passed library preparation QC (excluding control samples)	No. successfully sequenced and included in phylogenic tree
GD1	36/39	36/36
GD2	71/75	71/71
GD3	28/28	28/28
GD4	51/51	51/51
GD5	29/29	29/29
GD6	32/33	32/32
GD7	50/50	50/50
GD8	71/71	70/71

5.5.2 Phylogenetic analysis

The 8 cases examined produced a wide range of CNA, the detail of each case will be described below, but there were some general patterns also observed. Two of the cases (GD3 and 4) only showed a small number of CNAs, whilst the rest showed a greater degree of CNA. In these more aneuploid cases, the majority of CNAs (distinct in each case) were shared between all lesions, with a small number which were exclusive to a subgroup of samples. They were considered as representing the emergence of a new clone of tumour cells and were a 'branch' in the phylogenetic tree for that case into which a 'cluster' of lesions would fall. By identifying the pattern of these branch events a phylogenetic tree was produced. In several cases displayed below, multiple 'branch' events occurred, producing a more complex pattern of clonality.

The majority of the branch events occurred in a mutually exclusive fashion i.e. those samples which showed the presence of a new 'branch' CNA, were a subgroup of those in a previous branch of the phylogenetic tree for that case. There were, however, several samples (in GD1, 7 and 8) which appeared to contain CNAs from multiple branches of the overall phylogenetic tree, not obeying the overall branching pattern for the case. Additionally, the confounding CNAs appeared to be present in varying proportions of the tumour cell content for the sample as they produced smaller deviations in the relevant copy-number plots. These samples, which did not obey the overall structure of each case and appeared to contain multiple clones of tumour cells were deemed 'polyclonal' lesions and were placed within the tree according to the most predominant pattern

present, the origin of the minor subclone within the sample is also represented in the figures used by a dashed line. These analyses only included those events, which were shared between more than one sample, the individual samples had additional, private 'leaf' events.

For each case the CNA data is presented as a grid, which shows the pattern and position of the CNAs (to the nearest mb); the resultant inferred phylogenetic trees are also presented alongside this grid. To allow greater ease of interpretation we have also colour coded each branch or cluster and produced a colour coded body diagram showing the anatomical distribution of each clone.

5.5.2.1 GD1

This case showed a complex pattern of evolution with multiple CNAs, the majority of which were shared and are shown in Figure 22 as cluster 1, coloured red and include loss of 8p, 18q and chromosome 4 along with gain of 8q, chromosome 13 and 20q. The branches within the phylogenetic tree represent mainly single, mutually exclusive CNAs into which individual samples fall; overall 12 clonal events were identified (including the first, ubiquitous cluster of events). There were, however, several samples, which did not obey this pattern of mutual exclusivity, these were samples 29, 63 and 35. The majority of cells from these samples showed a copy number characteristic of a certain branch of the phylogenetic tree but also displayed a CNA (in a subclone of tumour cells) which was observed in an otherwise unrelated cluster of samples.

The distribution of the clones by anatomical location is shown in Figure 23. Within this case the three samples taken from the primary tumour were homogenous in-terms of their pattern of CNA, this pattern was shared with several liver metastases, although the samples taken from liver metastases were relatively diverse; all but one CNA cluster was identifiable within the group of liver lesion samples. The samples taken from nodal metastases were similarly diverse, although the two local lymph nodes fell into the cluster whereas the para-aortic, non-local nodes were distinct from this and each other.

The most strikingly monomorphic group within this case were samples taken from the peritoneal metastases. All 9 lesions from within the peritoneal cavity showed CNAs characteristic of cluster 2 and 4, a subgroup of 5 these samples showed an additional gain of genetic material on 5q.

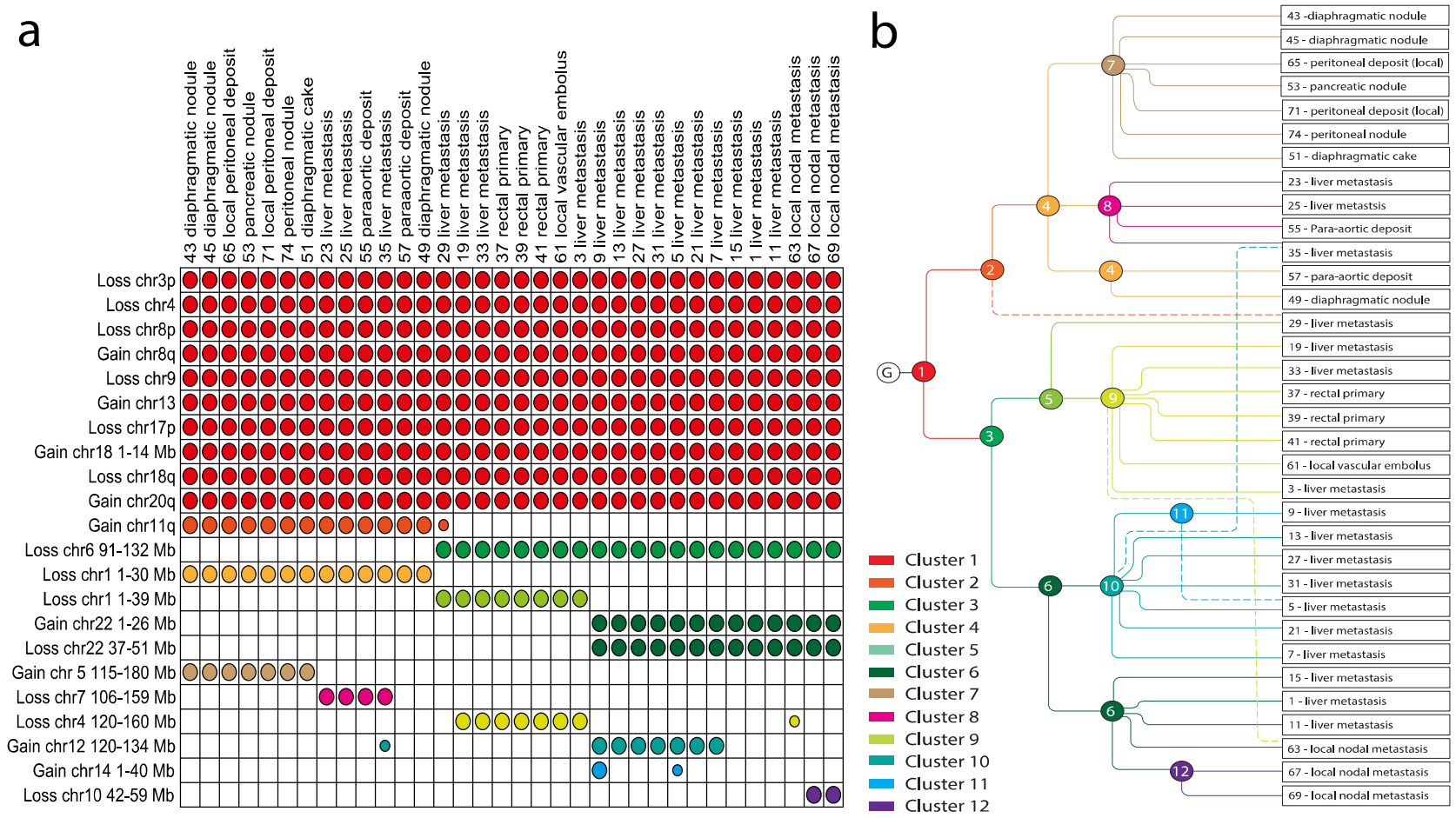


Figure 22: This figure displays the pattern of CNAs in GD1, (a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from GD1 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. (b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in (a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

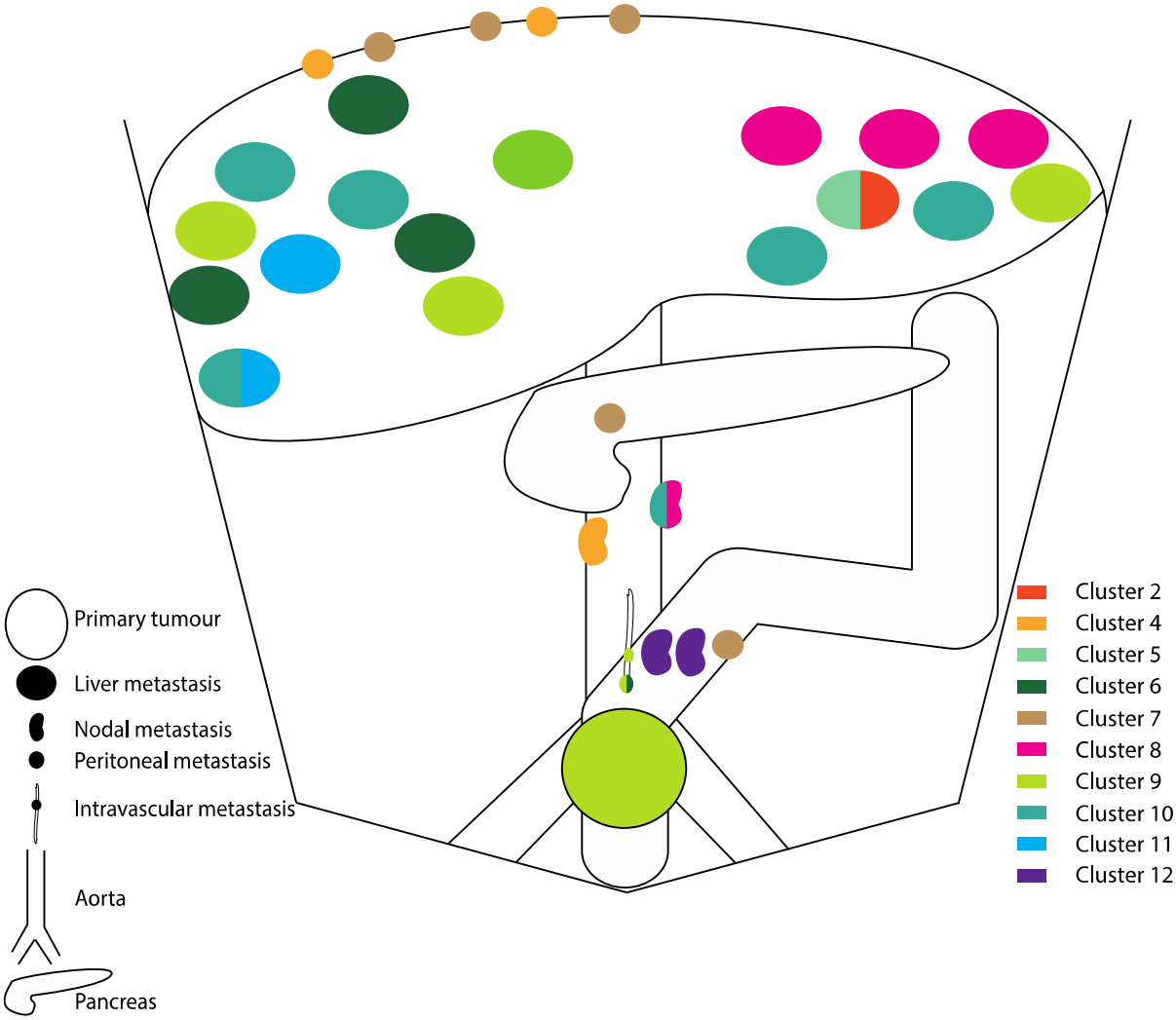


Figure 23: This displays the distribution of tumour cell clones within the abdomen of GD1 as inferred from the distribution of CNA. The key on the left describes the position of each lesion, the key on the right describes the colour code which corresponds to that in the phylogenetic tree and grid for this case, note cluster 1 is not present in this diagram as it only displays the terminal position of the samples within the grid and all lesions have gained additional CNAs to those present in cluster 1. Additionally, some of the samples shown in the figure contain more than one colour, these are the polyclonal samples described in the main text which contained CNAs characteristic of more than one cluster from the phylogenetic tree for this case.

5.5.2.2 GD2

This case showed a lower degree of aneuploidy and clonal diversity than was displayed in GD1, but once again all lesions shared numerous CNAs including gains in 7q and chromosome 8; a loss within 18q was present in all put one local nodal metastasis (the full list is shown in Figure 24a). These CNAs, forming cluster 1 and 2, were the only shared events identifiable in the lung metastases, whereas further events were evident in the primary tumour, nodal and liver metastases. Firstly, the primary tumour showed the emergence of multiple CNAs exclusive to this lesion (cluster 5, 6, 8 and 9) and the pattern of these CNAs was in keeping with the location of the samples within the primary tumour, the blocks were taken proximal to distal i.e. sample 73 was the most proximal sample taken from the rectal primary tumour and sample 83 was the most distal. Little clonal evolution was evident amongst the liver metastases, with a small subgroup of the 47 lesions lacking cluster 3 (displayed by the majority liver lesions and the primary tumour) and showing the acquisition of two CNAs, which are also present within local metastatic disease within the mesentery, adjacent to the rectal primary tumour. The phylogenetic tree and clonal distribution for this case are displayed in Figure 24b and Figure 25 respectively.

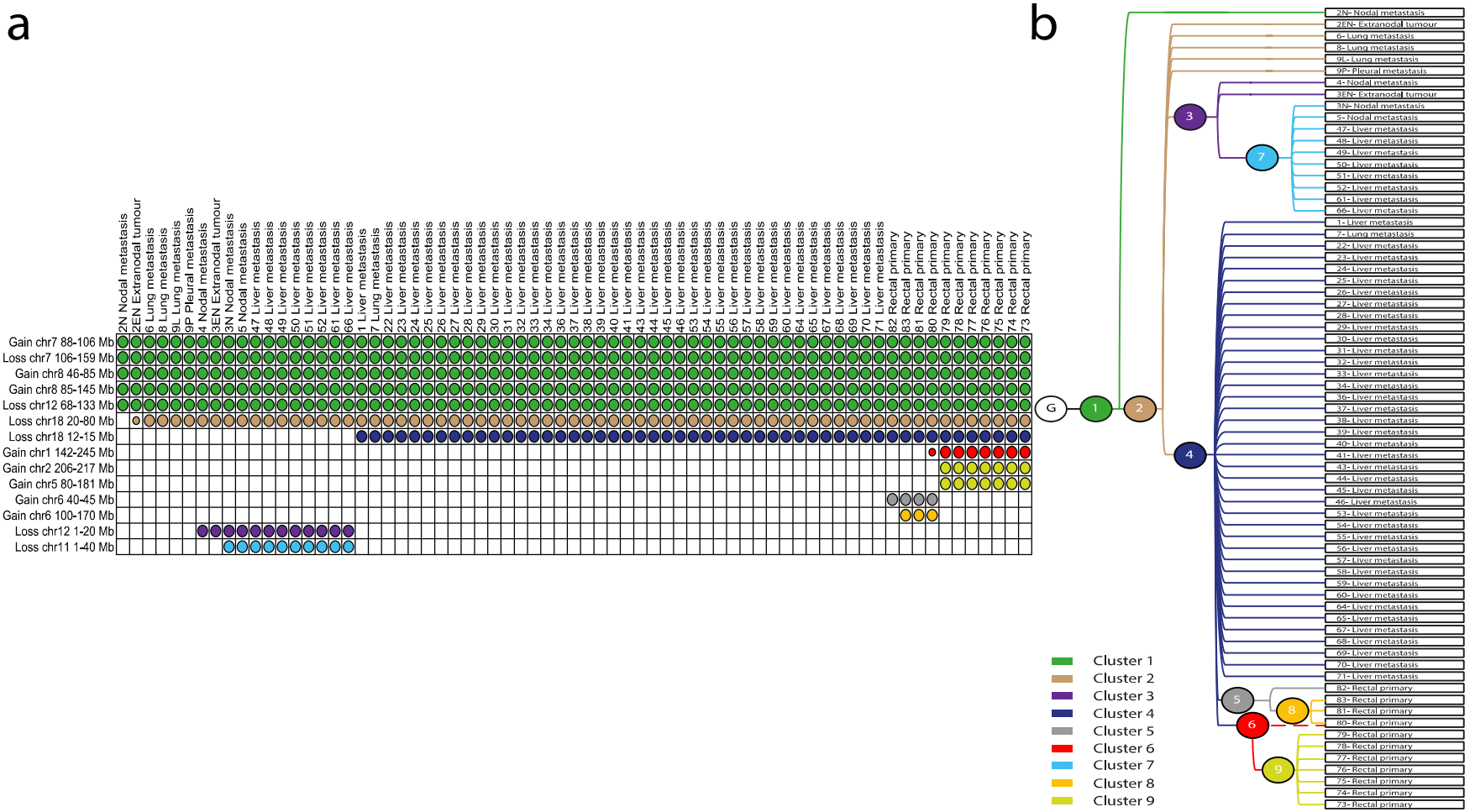


Figure 24: This figure displays the pattern of CNAs in GD2, (a) lists all shared, 'truncal' and 'branch' CNAs identified in all the lesions from GD2 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. (b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in (a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as only being 'subclonal'. The G in this figure represents the germline state for the individual.

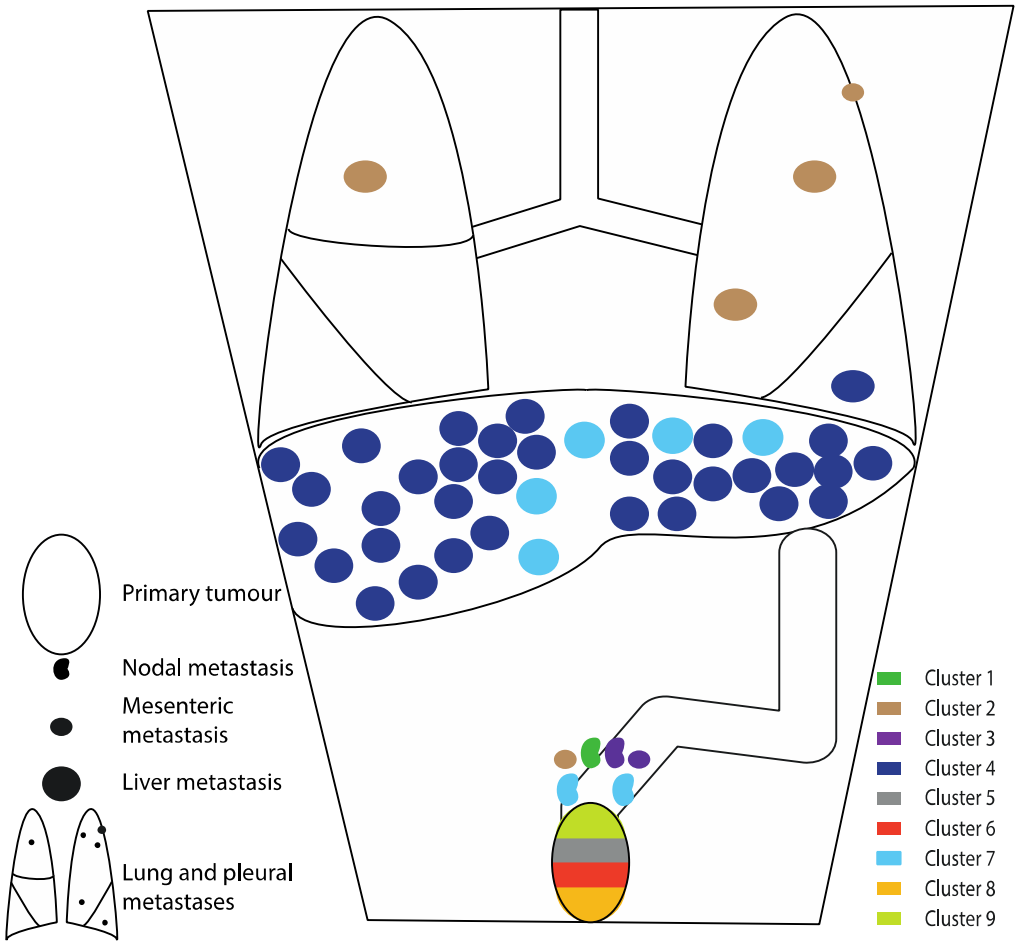


Figure 25: This diagram shows the distribution of tumour cell clones within the thorax and abdomen of GD2 as inferred from the distribution of CNA. The key on the left describes the position of each lesion, the key on the right describes the colour code which corresponds to that in the phylogenetic tree and grid for this case.

5.5.2.3 GD3

This was the first of two minimally aneuploid cases, showing only three CNAs between all samples taken from this case (including a gain of chromosome 8); note that 9 of the samples were those routinely taken for histopathology after resection of the primary tumour in 2011. The samples in this case were therefore temporally and anatomically separate but still show no evidence of clonal evolution by CNA. The list of cluster defining copy number changes, phylogenetic tree and map of tumour cell clones are displayed in Figure 26a, Figure 26b and Figure 27

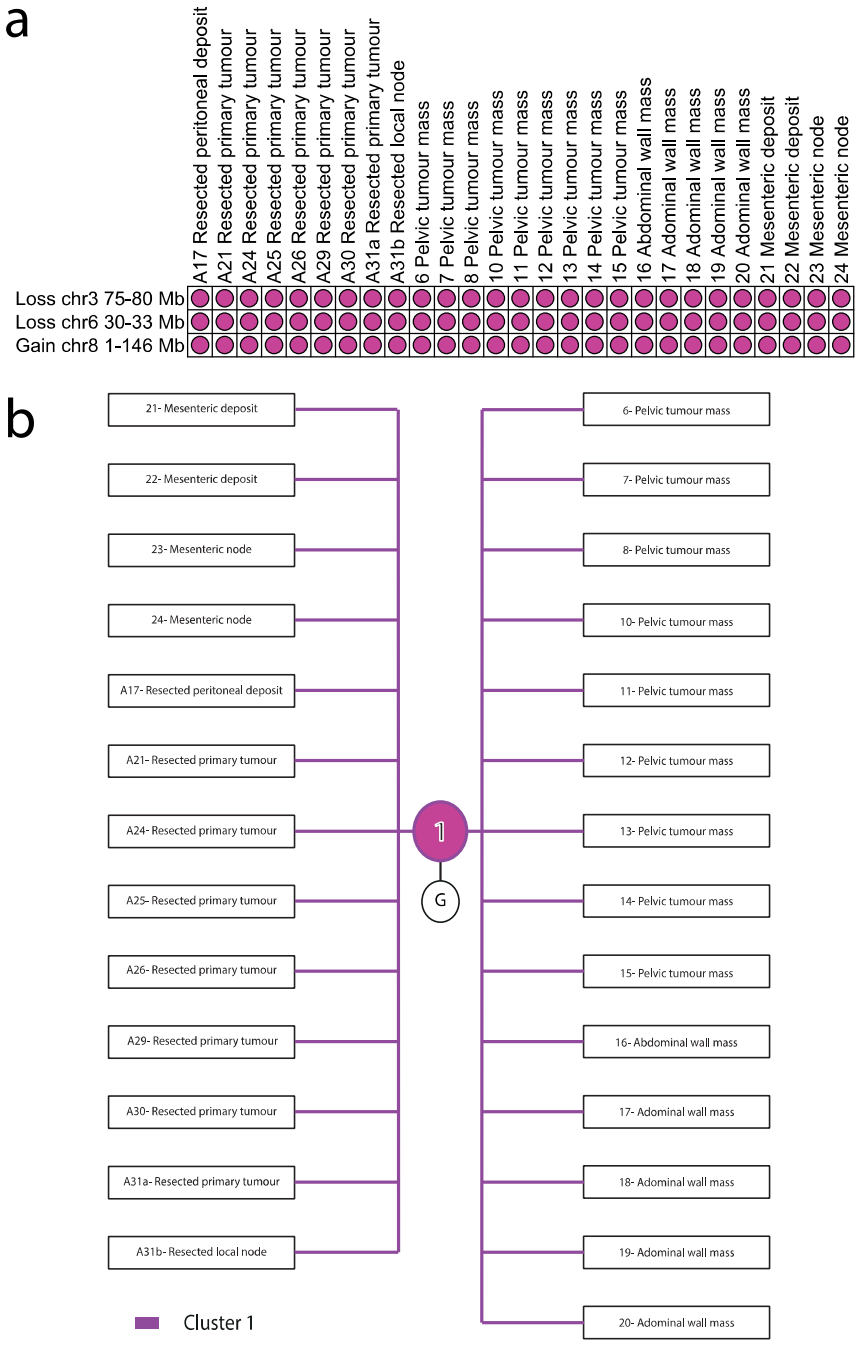


Figure 26: This figure displays the pattern of CNAs in GD3, (a) lists all shared CNAs identified in all the lesions from GD3 on the y-axis, all lesions shared only the 'truncal' CNAs. (b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in (a). The G in this figure represents the germline state for the individual.

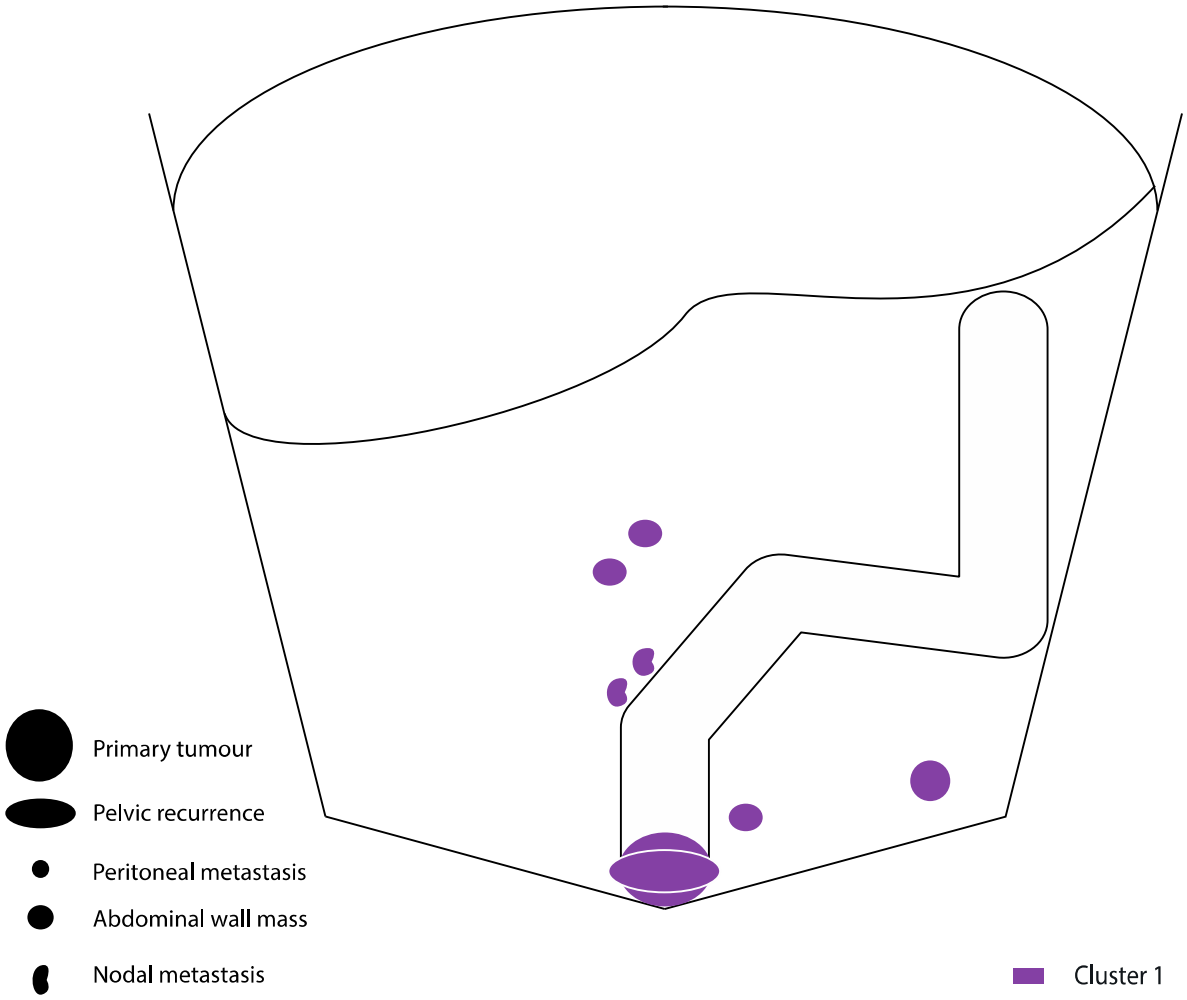


Figure 27: This diagram shows the distribution of tumour cell clones within the thorax and abdomen of GD3 as inferred from the distribution of CNA. The key on the left describes the position of each lesion, the key on the right describes the colour code which corresponds to that in the phylogenetic tree and grid for this case.

5.5.2.4 GD4

This is the second of the cases which displayed minimal CNA; the copy-number plots for this case were almost completely flat except for a small loss in 6p, identified in a subset of blocks taken from the surgical resection during life, and a loss in 8q present in a para-aortic metastasis sampled at autopsy. Therefore, within this case there are CNAs characteristic of the primary tumour which are not present within the recurrent disease. This second CNA was also present in a subpopulation of cells in a second para-aortic node and local recurrence also sampled at post mortem. The CNA data, phylogenetic trees and distribution of CNA throughout the body is displayed in Figure 28a, Figure 28b and Figure 29.

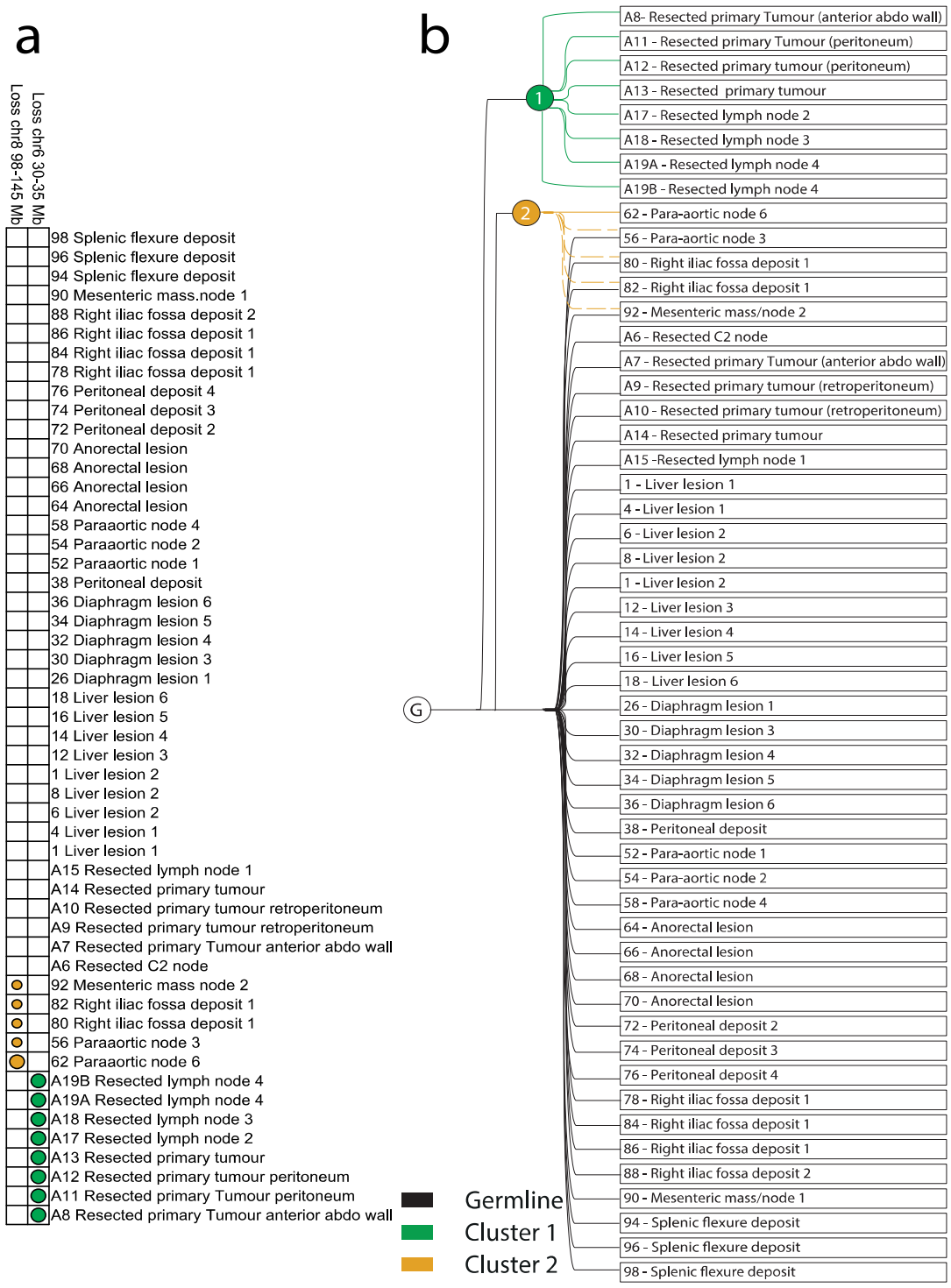


Figure 28: This figure displays the pattern of CNAs in GD4, (a) lists all shared, 'truncal' and 'branch' CNAs identified in all the lesions from GD4 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. (b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in (a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as only being 'subclonal'. The G in this figure represents the germline state for the individual.

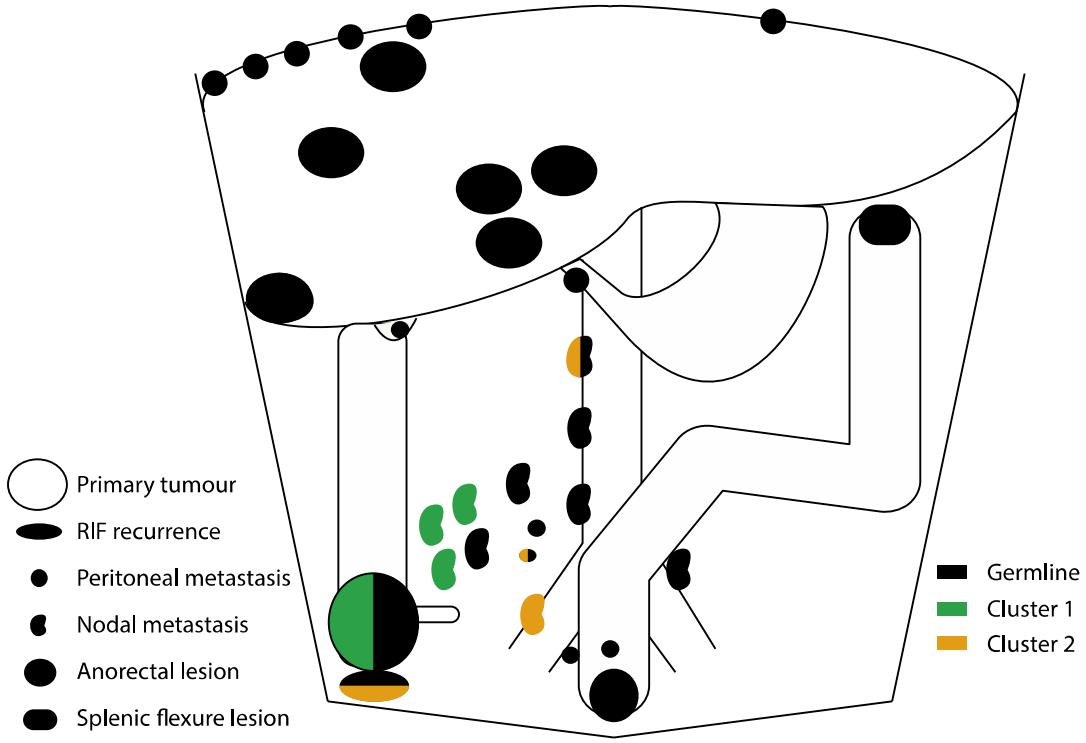


Figure 29: This diagram shows the distribution of tumour cell clones within the abdomen of GD4 as inferred from the distribution of CNA. The key on the left describes the position of each lesion, the key on the right describes the colour code which corresponds to that in the phylogenetic tree and grid for this case (note the black lesions are those which do not differ from the matched normal from this individual.

5.5.2.5 GD5

This case displayed a primary tumour (which was in-situ at post mortem examination) that displayed multiple patterns of CNA, including loss in 1p, 5q, 17p and gain in 13q. These changes are represented within the metastases, but the distribution of CNAs closely correlated with the anatomical location of the lesions in this case. Those lesions sharing CNAs changes in cluster 2 with the primary tumour comprise solely of the lung lesions (as displayed in Figure 31). The other early branch present in the phylogenetic tree for this case is cluster 9, which represents a small loss in 1p; only samples from the primary tumour and nodal metastases show this CNA. In this case, therefore, there is evidence of aneuploidy and clonal evolution represented by CNA, which is associated with the anatomical location of the lesions.

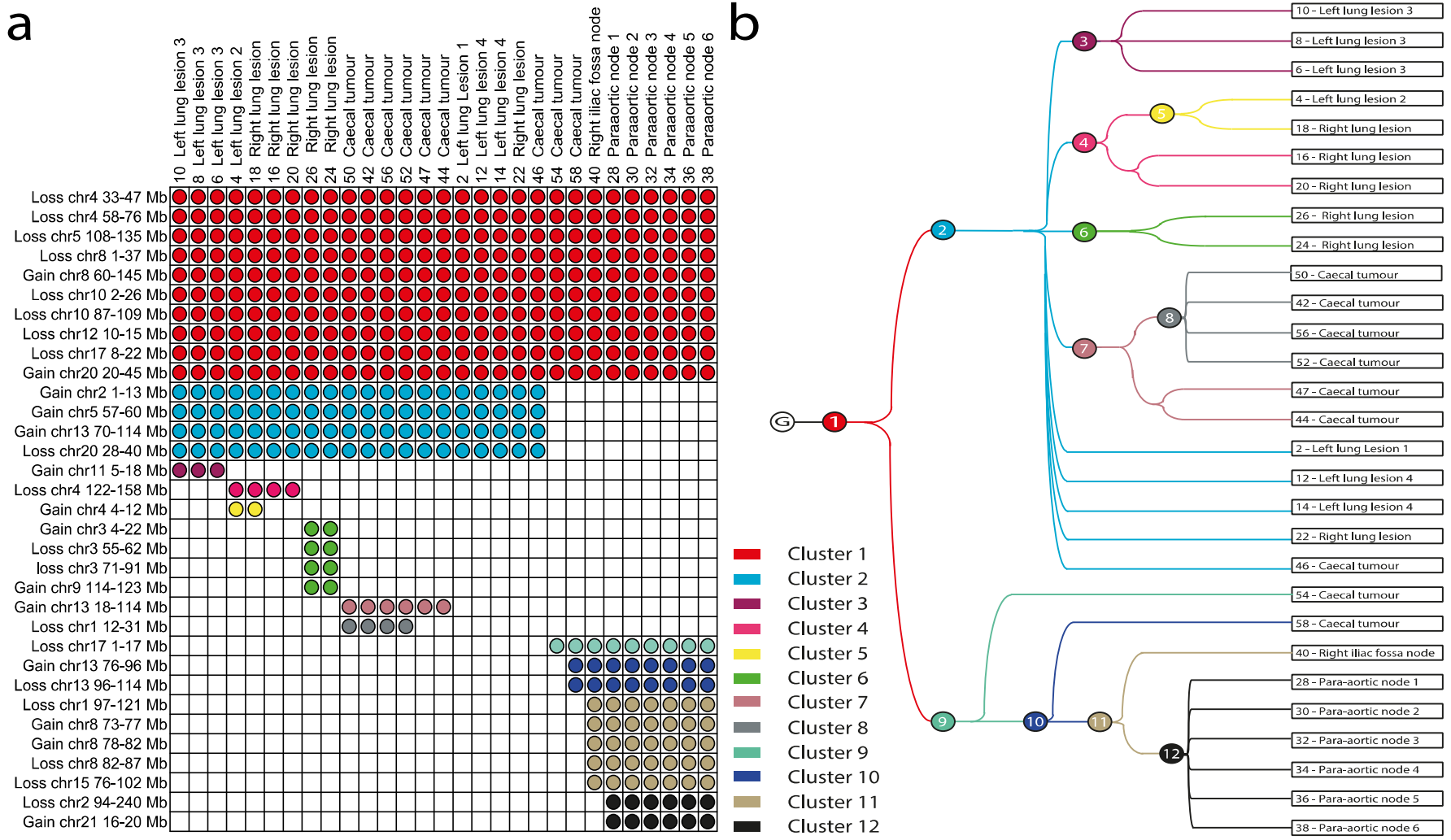


Figure 30: This figure displays the pattern of CNAs in GD5, (a) lists all shared, 'truncal' and 'branch' CNAs in all the lesions from GD5 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. (b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in (a). The G in this figure represents the germline state for the individual.

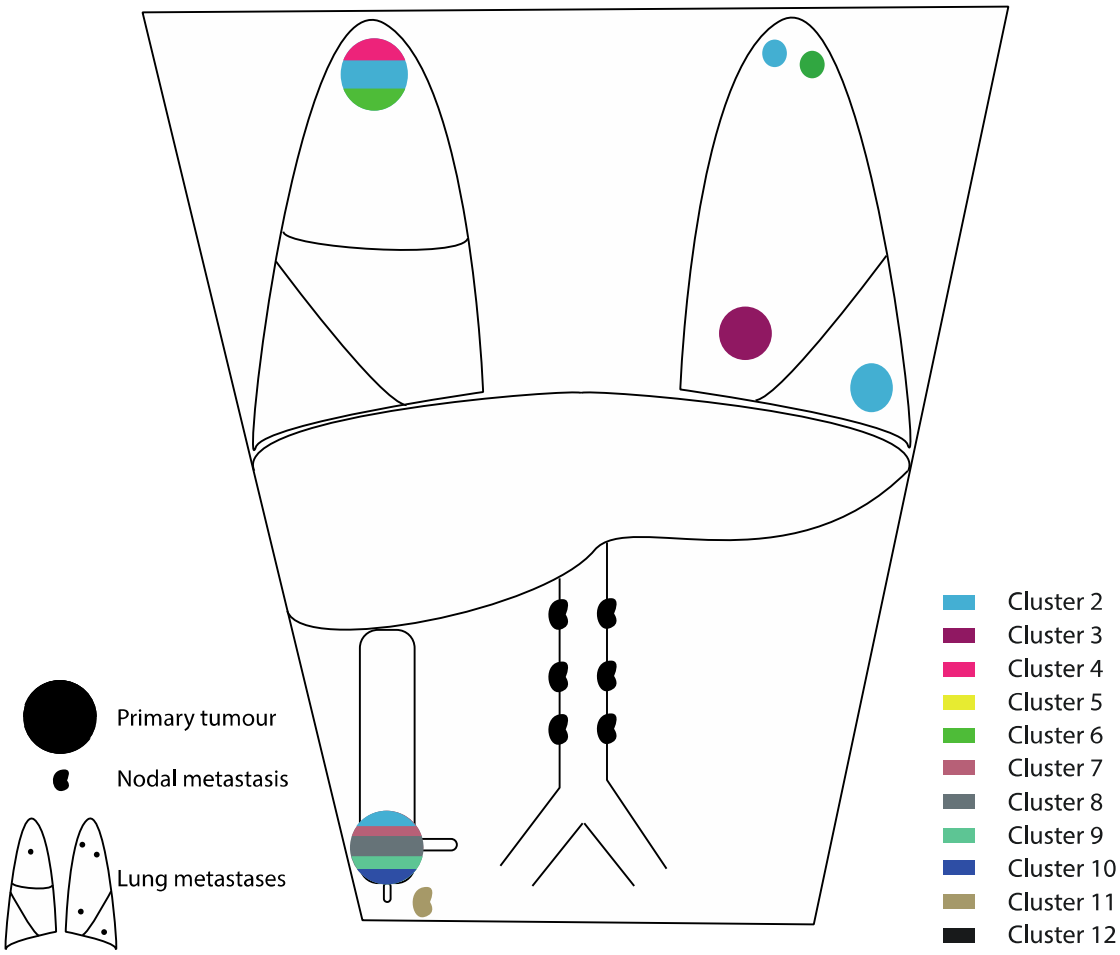


Figure 31: This diagram shows the distribution of tumour cell clones within the abdomen and thorax of GD5 as inferred from the distribution of CNA. The key on the left describes the position of each lesion, the key on the right describes the colour code which corresponds to that in the phylogenetic tree and grid for this case, note cluster 1 is not present in this diagram as it only displays the terminal position of the samples within the grid and all lesions have gained additional CNAs to those present in cluster. Additionally, some of the lesions shown in the figure contain more than one colour, these are the polyclonal lesions, described in the main text, which contained CNAs characteristic of more than one cluster from the phylogenetic tree for this case.

5.5.2.6 GD6

As presented in Figure 32a number of chromosomal aberrations were identified in this case, including loss of 1p, chromosome 4, 5q, 17p and chromosome 18 and gains in chromosome 20, which were ubiquitous and only a small gain in 15q which was seen in three samples taken from the primary tumour. This case, therefore, displayed evidence of chromosomal instability by the presence of aneuploidy, the nature of which was almost identical in every deposit sampled, so minimal evidence of clonal evolution was identifiable by CNA analysis in this case.

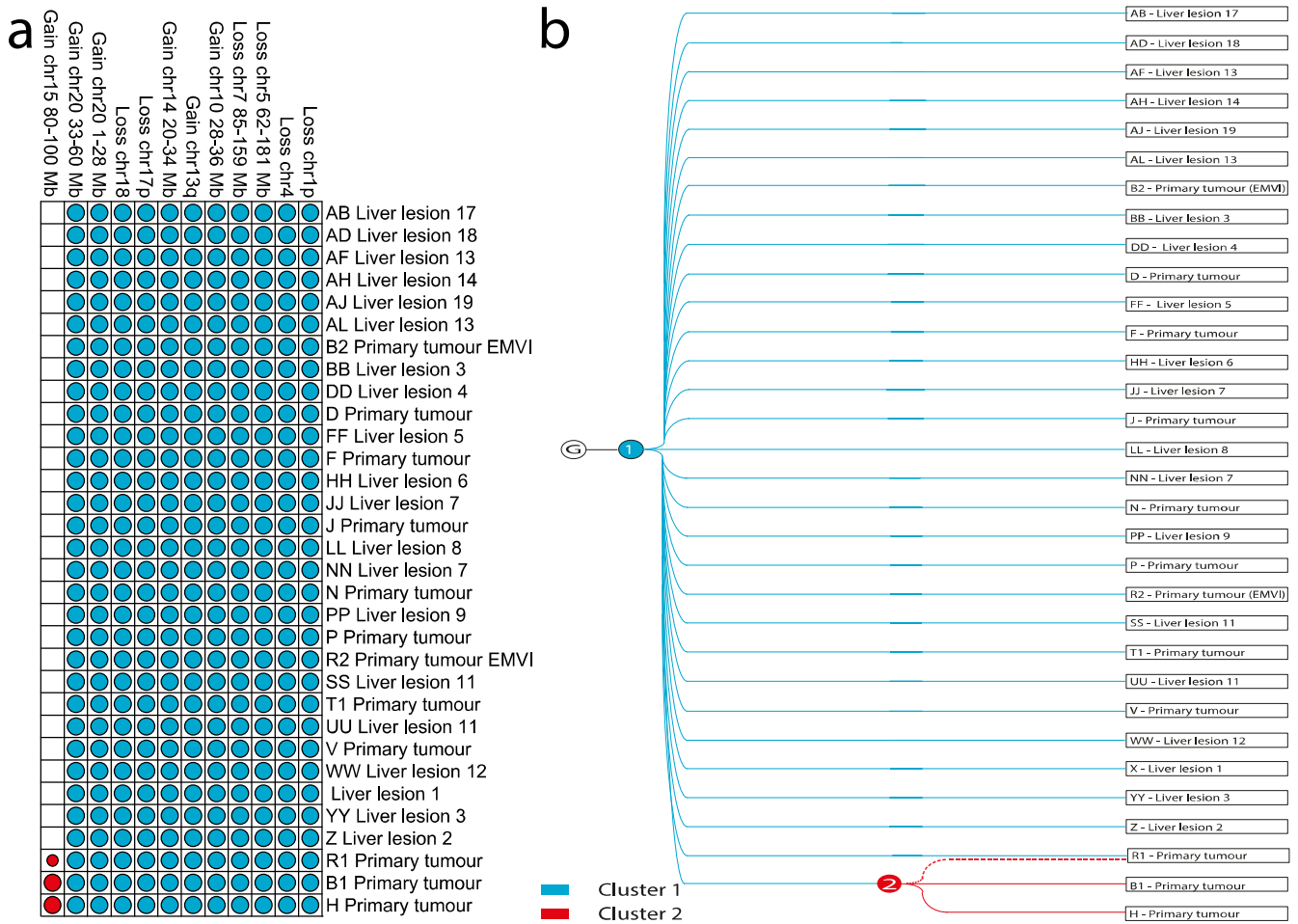


Figure 32: This figure displays the pattern of CNAs in GD6, (a) lists all shared, 'truncal' and 'branch' CNAs in all the lesions from GD5 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. (b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in (a). The G in this figure represents the germline state for the individual.

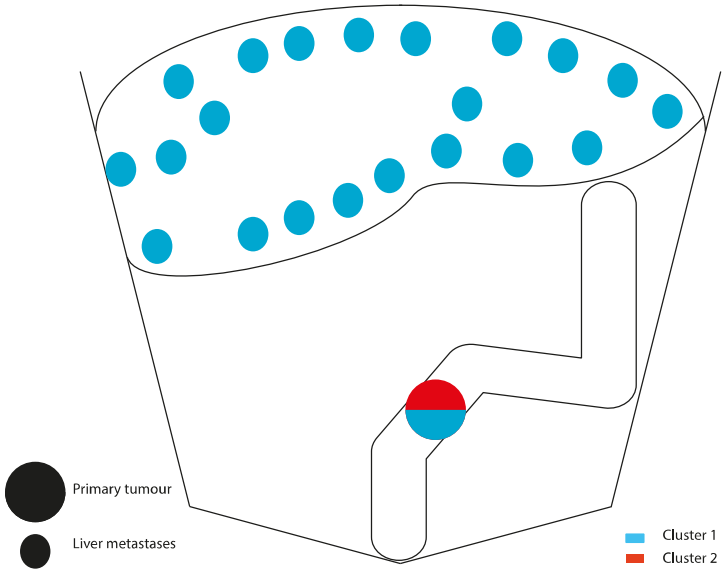


Figure 33: This diagram shows the distribution of tumour cell clones within the abdomen of GD6 as inferred from the distribution of CNA. The key on the left describes the position of each lesion, the key on the right describes the colour code which corresponds to that in the phylogenetic tree and grid for this case. In this case the samples originating from the primary tumour contained CNAs characteristic of more than one cluster within the phylogenetic tree for this case, this deposit therefore contains more than one colour.

5.5.2.7 GD7

This case showed numerous shared CNAs (including losses in 4q, 5q, 8p, 18q, 20p) along with a smaller number of 'branch' CNAs, all of which, except one (cluster 5) were identifiable in the primary tumour. This case therefore shows evidence of evolution represented by CNA, which appears to have originated within the primary tumour. Once again, the peritoneal deposits in this case have not diverged markedly from one-another. There is also a polyclonal metastasis in this case, sample 88i, which contains CNAs characterising cluster 4 and 6, which occur an in otherwise mutually exclusive fashion.

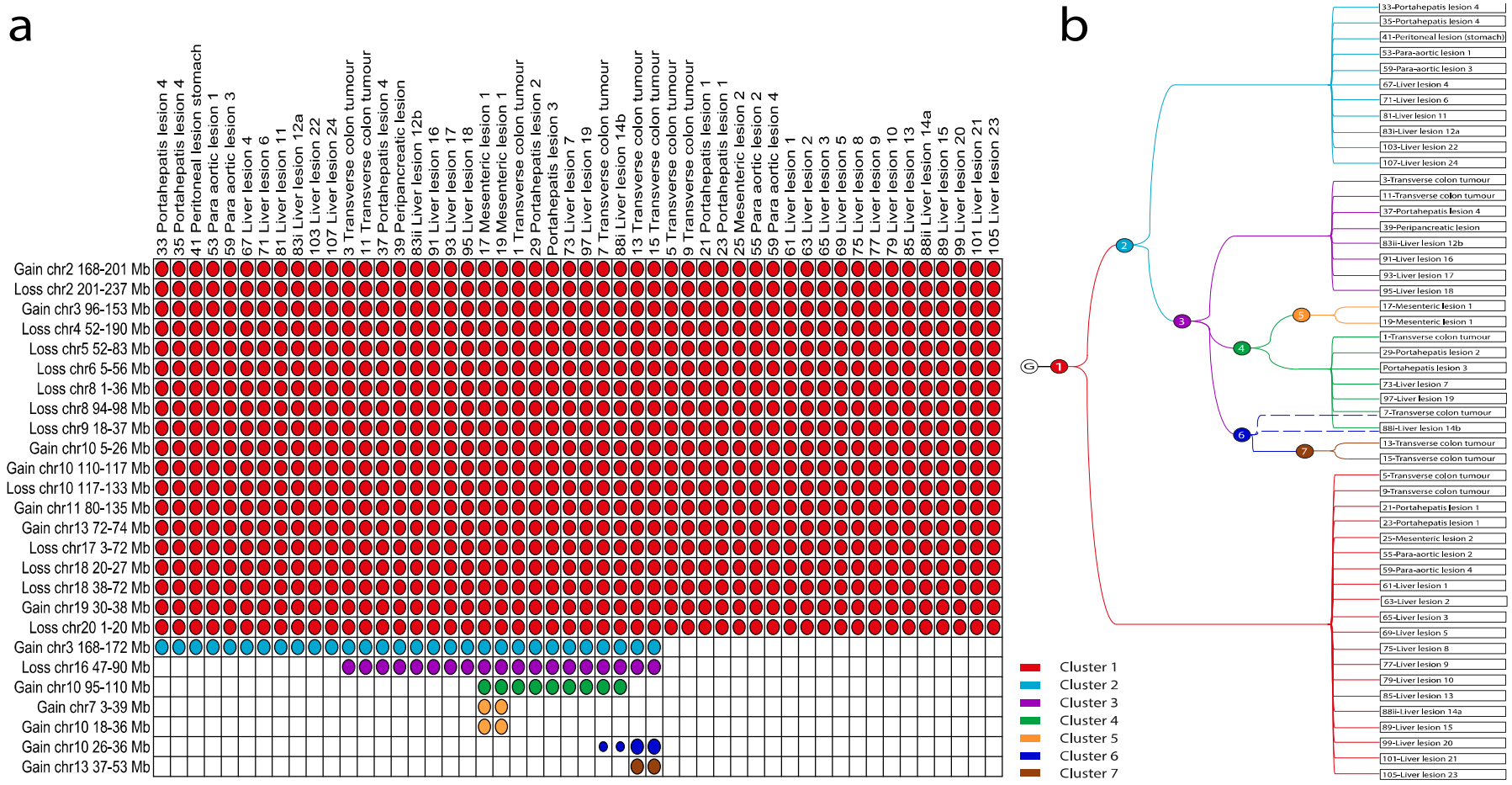


Figure 34: This figure displays the pattern of CNAs in GD7, (a) lists all shared, 'truncal' and 'branch' CNAs in all the lesions from GD7 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. (b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in (a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual

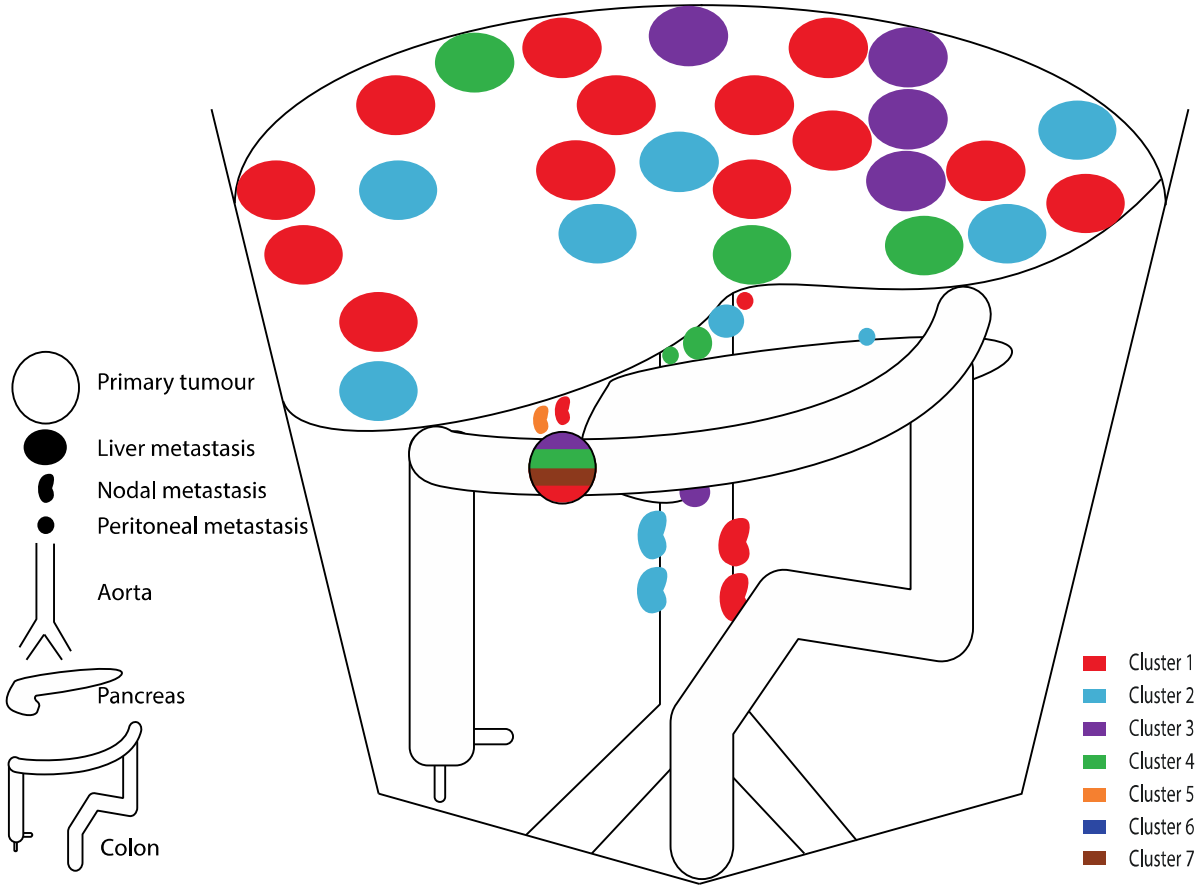


Figure 35: This diagram shows the distribution of tumour cell clones within the abdomen of GD7 as inferred from the distribution of CNA. The key on the left describes the position of each lesion, the key on the right describes the colour code which corresponds to that in the phylogenetic tree and grid for this case, note cluster 1 is not present in this diagram as it only displays the terminal position of the samples within the grid and all lesions have gained additional CNAs to those present in cluster. Additionally, some of the samples shown in the figure contain more than one colour, these are the polyclonal samples, described in the main text, which contained CNAs characteristic of more than one cluster from the phylogenetic tree for this case.

5.5.2.8 GD8

This case also showed marked aneuploidy, much of which is present as mutually exclusive branch events. There is divergence between the samples taken as part of the resection of the primary tumour and local metastases (in 2011) and the recurrent, metastatic disease (sampled in 2015). As in all cases the primary and metastatic disease share the largest cluster of CNAs (including loss of 8p, part of 15q, and 18q) however, except for a single lung metastasis (lesion 73), multiple steps of clonal evolution were observed which were exclusive within the resected and the metastatic disease. Also, 6 of the 7 resected local lymph nodes were observed as being divergent from the primary tumour. In addition to this inter-tumoural diversity the presence of polyclonal metastases was again noted in the liver (lesion 111, 119) and lungs (lesion 75, 39 and 19). In summary this case is highly aneuploid and shows a high degree of diversity arising at multiple stages before and after the occurrence of metastases and the resection of the tumour, it also shows polyclonal metastasis.

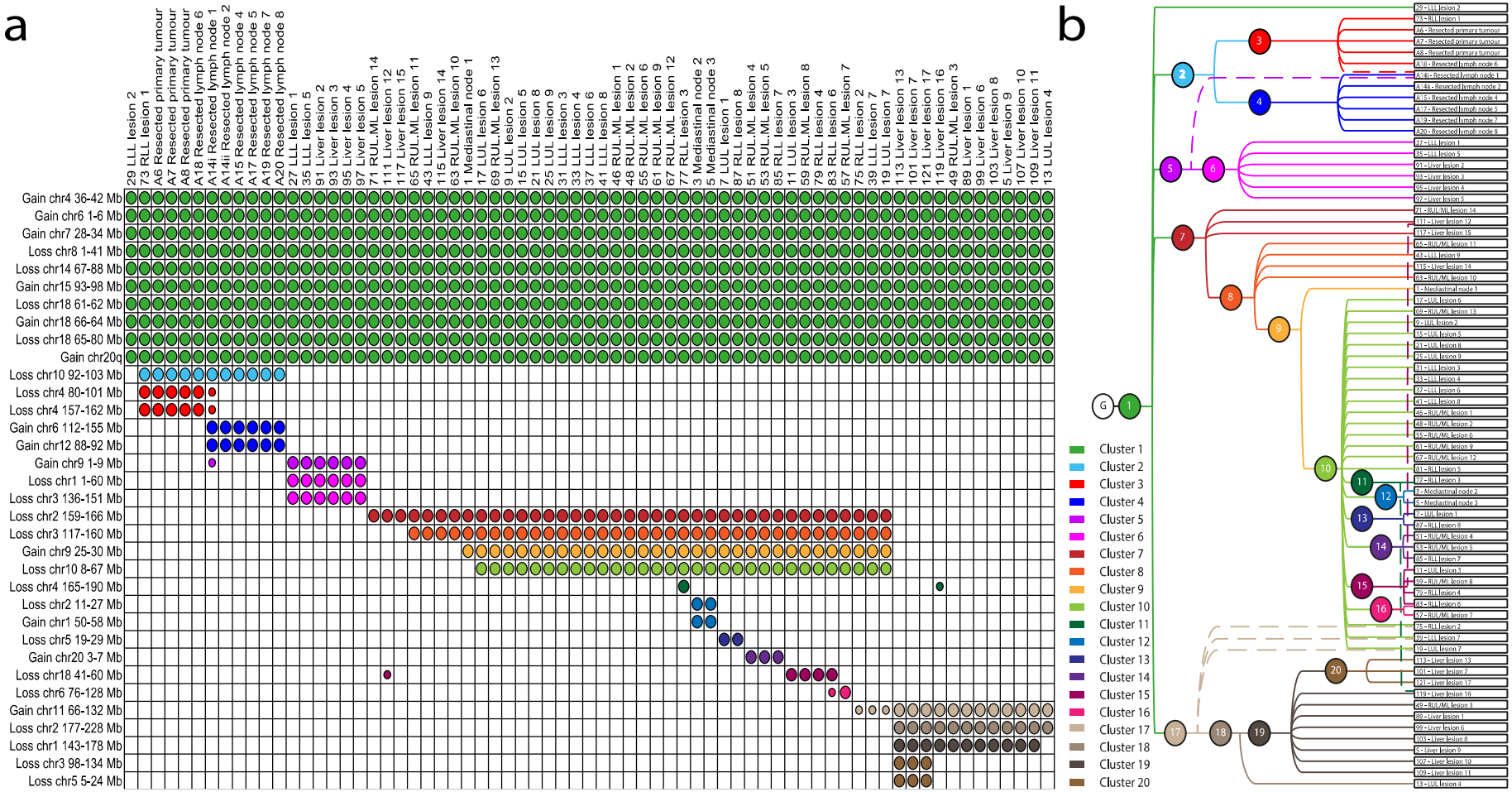


Figure 36: This figure displays the pattern of CNAs in GD8, (a) lists all shared, 'truncal' and 'branch' CNAs in all the lesions from GD8 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. (b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in (a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual

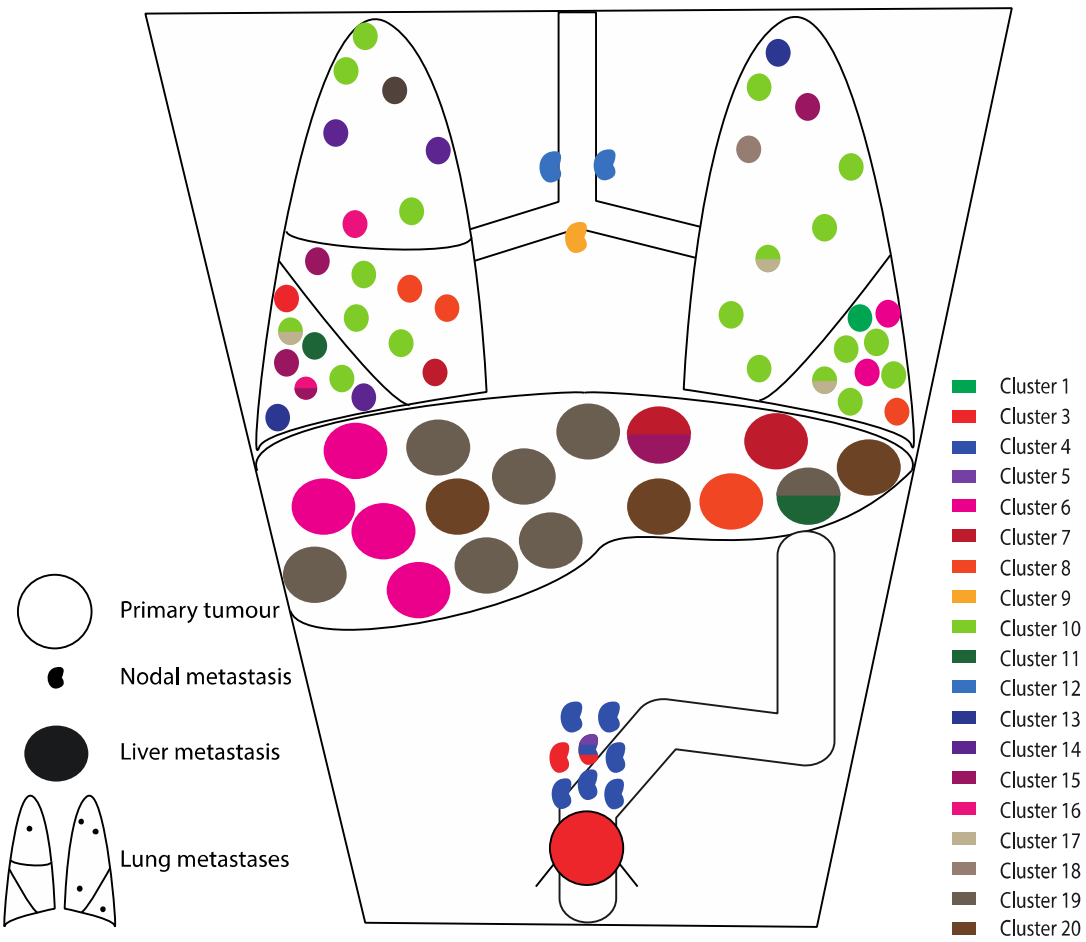


Figure 37: This displays the distribution of tumour cell clones within the abdomen of GD8 as inferred from the distribution of CNA. The key on the left describes the position of each lesion, the key on the right describes the colour code which corresponds to that in the phylogenetic tree and grid for this case, note cluster 2 is not present in this diagram as it only displays the terminal position of the samples within the grid and all lesions have gained additional CNAs to those present in cluster 2. Additionally, some of the samples shown in the figure contain more than one colour, these are the polyclonal samples described in the main text, which contained CNAs characteristic of more than one cluster from the phylogenetic tree for this case.

5.6 Discussion

We have performed copy number analysis by NGS on 367 of 375 tumour samples taken from 8 cases of MCRC, the majority of samples comprising this number were taken at post mortem examination and all extracted DNA originated from FFPE tissue blocks. The samples that did not produce a sufficient quality or quantity of DNA for library preparation and sequencing all originated from the primary tumour, except for one inadequate block from GD6. This was a small peritoneal deposit measuring approximately 1 mm and comprised an insufficient quantity of material to produce 200 ng of DNA required for sequencing.

In the case of the excluded samples from the primary tumours in GD1 and 2, it is unlikely that there was insufficient material available for DNA extraction as the lesions in each case were large, locally advanced lesions and ample tumour was microscopically visible in the sections taken from each of the samples. The poor quality of DNA yield was therefore most likely due to formalin fixation, poor fixation/autolysis of the samples taken or inadequate tissue processing. Variation in tissue processing is unlikely to account for the failure of the 7 samples from GD1 and 2, as all samples for each case were processed simulanteously, according to the same protocols and were therefore subject to the same conditions.

The deleterious effects of formalin fixation include those at the allelic level, such as increased numbers of nucleotide mismatches, particularly C·G > T·A [316] and more gross changes such as increased DNA fragmentation [322]. Highly fragmented DNA may not successfully provide 'libraries' for sequencing as if the majority of DNA is only present as small fragments they will be lost in the 'size selection' step of the NEB protocol. The effect of formalin fixation is the most likely explanation for the inadequate samples taken from GD1, as these tissue samples were three taken from the rectum following an extended period of fixation; all successfully sequenced samples from the primary tumour in GD1 were taken at the time of autopsy, the remainder of the samples (those which failed library preparation) were taken once the rectum was fixed and dissected at a later date as an in-situ stent within the rectum made sectioning of the fresh tissue impractical.

As all samples in GD2 were taken at the same time and underwent identical tissue processing and DNA extraction protocols, it is unclear why they did not provide adequate DNA yields for NGS. It is most likely that the inadequate samples were taken from a

portion of the primary tumour, which was more autolytic at the time of PM examination, although this was not appreciable on H and E examination.

5.6.1 CNA and phylogenetic analyses

In contrast with the macroscopic, histological and gene panel analyses (described in section 3 and 4), a large range of different levels of CNA was observed. Two patients (GD3 and 4) showed minimal CNAs and are therefore described as 'chromosomally stable', whilst the rest displayed aneuploidy characteristic of 'chromosomal instability'.

GD3 and 4 (the 'chromosomally stable' cases) were those, which were identified as being PMS1/MLH2 and MSH2/6 negative respectively by the IHC (in section 3) and as such fall into the mismatch repair/hypermutator subtype of CRC [50,61]. Previous studies examining the degree of chromosomal instability in dMMR CRC have consistently observed that dMMR CRC show a lesser degree of CNA [54,70] than chromosomally unstable types; the lack of CNAs in the dMMR CRC in this cohort is therefore in keeping with the existing medical literature. Additionally, amongst the 37 samples tested, there was no emergence of a CIN genotype, suggesting that dMMR CRC are unlikely to switch molecular subtype. To identify events representing clonal evolution in this subtype of tumour, it may therefore be necessary to perform a more comprehensive, possibly genome-wide, mutational analysis or examine tumours at the epigenetic, transcriptional or translational level.

As displayed by figures 22-37 the 6 chromosomally unstable cases (GD1, 2, 5, 6, 7 and 8) all showed a significantly greater degree of aneuploidy than the dMMR cases, although the uniformity of CNAs between samples within each case varied considerably.

All of the CIN tumours showed a shared, 'truncal' cluster of CNAs and these ubiquitous changes included many of the common chromosomal lesions previously identified in CRC [65]; loses of 1p, chromosome 4, 5q, 8p, 15q and 18q were recurrently identified, loss of 18q was the most common. Gains of the long arm of chromosomes 8, 13 and 20 were also identified more than once within this cohort. The presence of multiple shared events within each case (especially those identified in the regions of significant tumour suppressor genes such as *SMAD4* on 18q) suggests that all lesions in each individual arose from a common progenitor lesion; this is in keeping clinical and microscopic impression that all lesions were colorectal in origin. The fact that the majority of CNAs identified were shared is in concordance with recent studies suggesting that most

tumours acquire the majority of characteristics required for growth, invasion and metastasis early in their development [323–325] . This idea has been developed by a combination of mathematical modelling and experimental observation, particularly in work performed by Graham *et al* [323], which suggest that whilst it is possible to identify the occurrence of new genomic events within different regions of primary CRC, the majority of these are likely to be 'neutral', not conferring competitive advantage over other clones within the tumour and all major driver events have occurred at an early stage of tumour development and are ubiquitous. This work also suggest however that multiple clones may exist within a tumour which bear 'neutral' genomic events, as such it is possible that the variation observed within the CIN cases sampled is a simply the accumulation of biologically unimportant 'neutral' events. Work performed by the same group [326] also suggest that new 'driver' events which do occur within established lesions may not come to dominate the tumour in the fashion described by a Darwinian-type 'selective sweep', due to the architecture of the intestine. However, it is possible that the opportunity presented by the establishment of a new tumour deposit (such as a metastasis) may permit new driver events to become established within a population; work previously cited in treated CRC, RCC and prostate cancer [87,88,100] were able to demonstrate the emergence of new driver mutations in advanced cancer suggesting this may be the case. It is uncertain therefore whether the new CNAs observed in the CIN CRC in this cohort are biologically significant, however they do at least represent evolutionary events (some of which may be neutral in terms of their biology) and the emergence of new tumour cell clones. Therefore, by examining the occurrence of CNAs within each case we have identified a range of evolutionary complexity within each patient. This spectrum included GD6 at one end, which showed a large number of shared CNAs with almost no 'branch' events identified, and GD1, which displayed multiple new CNAs within subgroups of samples representing the emergence of new tumour cell clones, at the other.

Patient GD6 displayed a relatively large number of ubiquitous CNAs, with only three samples, all taken from the primary tumour, showing any clonal diversity, a loss in 15q not identified in the other samples from the individual. This tumour was therefore remarkably homogenous, both phenotypically and genomically, within the context of the assays performed up to this point. These findings are particularly interesting in the overall context of this case as the patient was a young woman who presented with a tumour, which had invaded mesenteric vessels and formed bulky hepatic deposits without developing observable heterogeneity as was seen in the other CIN CRC. The

short clinical course of this patient's illness either suggests the early development of an aggressive clone *de novo* or complete replacement of the primary tumour by a dominant clone with aggressive spread over a short period; bearing in mind the current literature [323,326] the former scenario is more probable. It is possible, with more in-depth characterisation of this tumour we may identify heterogeneity at the mutational level, as such whole genome sequencing of this case has been performed and is presented in section 6. Within the broader context of CRC, characterisation of such aggressive lesions may allow for early identification of such tumours which may be stratified into more intensive treatment stratagems.

Within the cases which showed intertumoural heterogeneity and clonal evolution between deposits, the emergence of new CNAs in regions of recurrent CNA (as identified in the TCGA publication [65]) were identified (such as losses in 18p, and 4q). However for the reasons stated above i.e. the majority of lesions are likely biologically neutral/passenger changes, and due to the inability of much larger studies of matched, resected CRC and metastases to identify recurrent CNAs which are associated with metastatic disease [92,94], the biological significance of these lesions in and of themselves is uncertain. The significance of CNA within this context is therefore as a marker of clonal evolution.

The distribution of CNAs appears to correlate with metastases within anatomically similar locations, particularly in the clonally simpler cases (GD2 and 5). Within GD2 there are patterns of CNA exclusive to the primary tumour, hepatic and lung metastases, GD5 shows clear separation between nodal and lung metastases. This separation was less marked in cases with higher degrees of aneuploidy, although in all cases hepatic metastases were observed which had no shared events with local nodal metastases (beyond the ubiquitous truncal changes). This suggests that the mechanisms underlying nodal metastases and those to other locations may vary, even within an individual. This observation was also made in recently published work by Naxerova *et al* [85], who observed a mutually exclusive pattern of passenger mutations in surgically resected nodal metastases and matched hepatic deposits. Gundem *et al* [87] also identified genomic similarities between matched metastases at similar anatomical sites in the post mortem setting.

Concurrence of genomic changes and anatomical location of metastasis has also been observed in a post mortem study of advanced pancreatic carcinoma [106]. The sequencing

analyses performed in this instance focused upon chromosomal rearrangements and demonstrated the occurrence of new potential driver events within metastases, most significantly amplification of *KRAS*. This observation is slightly at odds with the existing evidence in CRC, in which new driver events are thought to be rare [76,78,81]. This may reflect an essential difference in the evolutionary biology of pancreatic carcinoma and CRC; Sottoriva *et al* [323] described pancreatic carcinoma as tending toward non-neutral evolution as compared to CRC.

Within the complex pattern of evolution present in GD1, the peritoneal disease sampled displayed the greatest uniformity of copy number of the metastatic sites sampled. The striking similarity of the peritoneal disease suggests that all the peritoneal disease in this case arose due to single event (i.e. localised penetration of tumour onto the peritoneal surface and subsequent intra-peritoneal spread) rather than multiple independent seeding events, which would be more likely to produce a wider spectrum of CNA changes. As peritoneal disease is not easily or often sampled during life, little other data exists comparing this site of metastasis to others within the same individual. Diep *et al* [303] identified recurrent 5p and 12p in peritoneal lesions in a cohort of non-matched primary and metastatic CRC, this association has not been confirmed by data from matched primary CRC and peritoneal disease [92], however both of these studies only contained single deposits from individuals so the data from this study is not directly comparable. The observed homogeneity of peritoneal disease in this study is an important focus for future investigation as, if confirmed, therapeutic approaches specifically targeting a single clone causing intra-peritoneal disease may be valuable in this unresectable disease state.

Within the large number of samples sequenced from these individuals, we identified multiple samples showing CNAs from disparate branches of the phylogenetic tree for that case. In each instance of this phenomena the sample showed a dominant pattern of CNA with a smaller subpopulation of cells displaying a second incongruous CNA; as such these samples were interpreted as being polyclonal, containing two tumour cell clones, one more predominant than the other. The phenomena of polyclonal metastases has been previous documented in prostatic cancer[87] and in xenograft animal studies [327], although the mechanism by which these lesions occur are unclear and are likely multiple within the data documented in this work. As observed in section 3, the individuals with liver metastases in this cohort showed extensive replacement of the liver parenchyma by tumour, as such it is possible that the polyclonality observed in liver metastases from

GD1 and 8 reflects inadvertent sampling of two lesions, which were originally distinct lesions, which had grown into one another. In contrast to this, the polyclonal deposits sampled from the local lymph node in GD1, the right lung of GD5 and lungs and local lymph node in GD8 were distinct lesions surrounded by normal tissue. This group of distinct tumour deposits may have gained their apparent polyclonality by the seeding of one tumour to another, so called 'metastasis to metastasis seeding' or a polyclonal metastasis from the primary tumour. Metastasis to metastasis seeding is a phenomena which has been documented in animal studies [22,23]. It is postulated that this occurs either dues to a specific site being particularly attractive due to local factors which might be favourable to tumour growth, such as hypoxia [299] and hence attracting multiple metastatic events, or due to established metastases forming particularly receptive 'soil' for further circulating tumour cells possibly due to inter-clonal cooperation [102].

The possibility of individual metastatic deposits resulting from different pathways of spread (i.e. both vascular and neural invasion) has not been considered within the current literature. This appears to be present in the polyclonal 'nodal' lesion, sample 63, in GD1, as microscopically it is possible to identify tumor within both nodal and vascular structures in an individual deposit (see figure 9). This lesion underlines the potential for micro-dissection of primary CRC and local and distant metastases to identify genomic and phenotypic changes characteristic of specific modes of spread e.g. perineural, vascular and lymphatic.

The CNA patterns observed in GD8 also provide information regarding the timing and mechanisms of metastatic disease. As displayed in Figure 36 all samples deriving from the resected tumour and local metastases share a loss in 10q, and then either losses in chromosome 4 or gains in chromosome 6 and 12; these CNAs are exclusive to the resected blocks and one post mortem sample taken from a lung metastasis, the remainder of metastases sampled at post mortem only share the earliest, cluster 1, CNAs. This data would therefore suggest that the majority of the distant metastatic disease in this individual arose before the local nodal metastases and possibly that initial dissemination occurred by a vascular rather than lymphatic route. If the opposite were true then it might be expected that evidence of the genomic changes present within the systemic disease would be observed within some of the local lymph nodes. The importance of vascular dissemination in CRC [240,261,328] has been clearly established and therefore the observation that CRC does not solely rely upon lymphatic spread does not present a paradigm shift in the understanding of CRC. The occurrence of early

metastases is also well documented in CRC where approximately 10% of early stage (pT1) CRC will display local lymph node metastases if resected [245], however the presence of such disseminated, micro-metastatic disease is undocumented in CRC. This finding once again reinforces the idea that CRC can acquire the biological characteristics required for growth, invasion and metastasis at an early stage and any subsequent evolution may be neutral [323,325]. Identification of the likelihood of this having already occurred in early stage lesions would be an indicator to trial earlier adjuvant therapy in such individuals.

It is worthy of note that GD1 and GD8 demonstrated the highest degree clonal/evolutionary complexity by CNA and were also the two cases which received multiple cycles of chemotherapy. The suggestion that chemotherapy may drive diversity in CRC is supported by the emergence of treatment resistant cells during treatment with EGFR blockade [329], although this diversity arose secondary to a highly specific treatment. Clonal diversity driven by conventional chemotherapy has been investigated in small cohorts in breast and oesophageal cancer [113,330]. This work showed a variable effect of chemotherapy in which treatment resistant disease may emerge from single or multiple clones within the pre-treatment tumour thereby producing an unpredictable effect on the presence of genomic diversity within an individual. Within the two cases from this cohort the marked evolutionary complexity may reflect the effect of chemotherapy however many other variables including disease longevity, tumour micro-environment and features endogenous to the primary tumour may equally be responsible. GD1 and GD8 do differ slightly in the clinical history of each case, GD1 had the primary tumour in-situ at post mortem, whereas GD8 underwent resection of primary tumour 4 years prior to death, however both cases displayed divergence between the primary and metastatic disease similar to the other CIN cases within the cohort. In GD1 metastases emerged at multiple stages within the timeline of disease and within GD8 the primary tumour showed CNAs not present within the disease sampled 4 years later, suggesting that the clone (or clones) producing irresectable disease had already disseminated at the time of resection. Therefore, although the pattern of CNA change within these cases does provide insights as to the timing of metastatic seeding, further work incorporating prospective sampling of large cohorts of patients would be required to conclusively determine the effect of chemotherapy on the evolutionary pathways in CRC.

A further point to address from this data is the outlying *KRAS* wt sample (sample 82) from GD2. As demonstrated by sample 82 demonstrates the majority of CNAs exhibited

by the remainder of the primary tumour but lacks a small gain in the terminal portion of chromosome 6, it is therefore a slight outlier within the samples originating from the primary tumour. The deviations in the CNA plots produced by DNAcopy and CNAnorm provide an approximation of the relative tumour cell contents between samples and as shown in Figure 38 sample 82 displays deviations of comparable size to the adjacent samples; hence it is unlikely that the mutational and CNA differences between sample 82 and the remaining samples from the case are due to low tumour cell content. Additionally sample 82 does not show copy number loss in the region of the *KRAS* gene, therefore it is unlikely that the lack of detectable *KRAS* mutation is due to loss of this region although the possibility of loss of heterozygosity as an explanation for the *KRAS* wt is not excluded. The CNA data therefore provides further evidence that genuine heterogeneity is present at the mutational and chromosomal level in primary tumour from GD2. As shown in Figure 28, wt sample from GD4 did not display any CNA therefore this data cannot be used to infer the possible tumour cellularity of the sample. The final wt sample from the 'Gift' cohort, GD8-A20, showed only small deviations in copy number, suggesting the outlying pyrosequencing result from this case may have arisen due to low tumour cellularity.

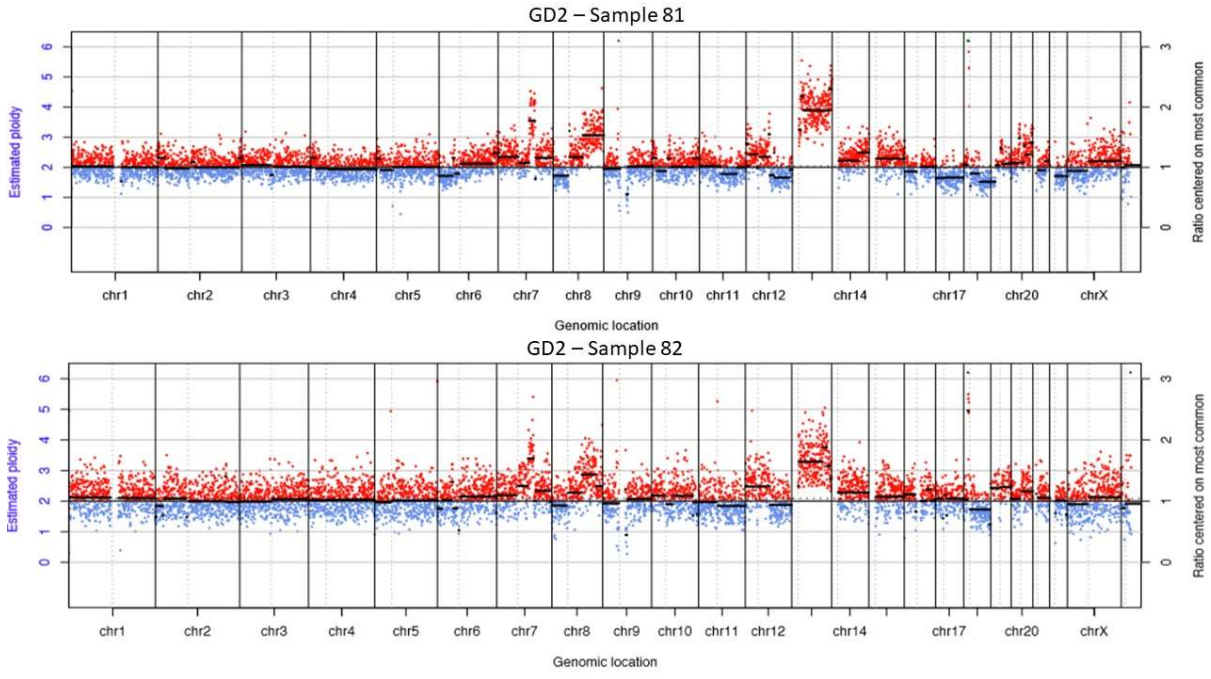


Figure 38: CNA plots from two samples taken from GD2. Sample 81 was *KRAS* mutant, whereas sample 82 was KRAS wt.

In summary in this section by examining the presence of CNA across multiple samples in cases of disseminated MCRC we have demonstrated a range of heterogeneity within

a group of non-dMMR/CIN CRC; the dMMR tumours did not show evidence of evolution by CNA in the disseminated disease state. Phylogenetic analysis of the CNA data showed a spectrum of clonal diversity ranging from simple patterns of clonal divergence associated with anatomical tumour to more complex cases with evidence of multiple seeding events from disparate portions of primary tumours giving rise to metastases in similar anatomical locations. However, even within the more clonally complex cases relatively homogenous groups of metastases emerged at specific metastatic sites suggesting either a biological predilection for a metastatic site within a specific tumour cell clone or passenger events arising in regions of the tumour with anatomical access to a site or mode of spread. The analysis of CNA data also suggested that disseminated disease may arise prior to the emergence of locoregional metastases. Additionally, the CNA data corroborates the minor degree of mutational intra-tumoural heterogeneity within GD2.

6 Cadaveric material mutational sequencing

6.1 Introduction

Up to this point the data in this thesis has demonstrated the existence of genomic heterogeneity within advanced colorectal cancer at the chromosomal level by performing low resolution WGS upon multiple samples taken from individuals with advanced disease. This variation is more pronounced in CIN CRC and appears to be an ongoing process during the progression and spread of CRC. Although the assay was not performed with a view to the identification of biologically significant genomic changes, it was possible to identify CNA at the genomic loci corresponding to canonical genomic loci in the development of CRC.

The mutational sequencing, detailed in section 4, performed via pyrosequencing did not identify mutational heterogeneity in disseminated disease. The sequencing performed in this section examined only a limited number of loci in *KRAS, NRAS, PIK3CA* and *BRAF*; these regions were chosen as they are the most established predictors of EGFR blockade response [8,145] and are of prognostic significance. Additionally *PIK3CA* status may have a relationship to the observed impact of aspirin in CRC [163]. This mutational homogeneity contradicts the theory that unsampled heterogeneity in *KRAS* mutation status underlies the modest improvements in outcome gained by EGFR blockade in patients shown to be *KRAS* WT on pre-treatment testing [100]. This lack of heterogeneity suggests that all important mutational 'driver' events have occurred early in the evolution of CRC and subsequent, late, sub-clonal events are likely to be neutral i.e. not conferring selective advantage to the tumour cell clone, as suggested by statistical inferences made from solid tumour sequencing data [323]. However, pyrosequencing has a sensitivity of approximately 5% (i.e. a mutation must be present in at least 5% of molecules sequenced to be detected) therefore therapeutic resistance may still be due to growth of very minor subclones present below the threshold of the sequencing modality. Pyrosequencing is also only practical for examining a relatively small number of genes and a broader spectrum approach, using either array techniques or NGS, may identify heterogeneity in other significant loci.

The presence of mutational genomic heterogeneity has been described in the Introduction, however in summary, work examining the presence of intratumoural heterogeneity in primary CRC suggested that the degree of heterogeneity present may

be significant [287]; Kosmidou *et al* used pyrosequencing to examine *BRAF, KRAS* and *PIK3CA* genes at therapeutically significant loci showing frequent new mutations within *KRAS*. This finding is of significance as any metastases arising from these tumours may reflect this heterogeneity. Subsequent work using NGS [83,288] contradicted this study finding a very high concordance of *RAS* mutation throughout multiple sections taken from primary CRC. It is probable that the work performed using NGS is more reliable as a greater number of sections of primary tumour were examined and the sequencing modality used provides a greater sensitivity as compared to pyrosequencing; due to the relative insensitivity of pyrosequencing it is possible that, in samples of low tumour cell content, mutations may be missed which would be detected with deep sequencing. With the acquisition of larger cohorts of matched primary and metastatic samples, it was been suggested, using pyrosequencing, that within chemotherapy naive CRC, the concordance of mutations within *RAS, BRAF* and *PIK3CA* was greater than 90%, not reflecting the trend observed in Kosmidou *et al* [78]. The application of more extensive mutational panels, via NGS or array panels, examining key pathways in CRC carcinogenesis including Wnt, RAS-MAPK, PIK/AKT/mTOR showed little variation in the known therapeutically predictive genes but described a variable rate of heterogeneity in one or more 'driver' genes [80–82,312] ranging from 15-50 %. All of the cited literature only included single resected metastases, as such this pattern may only reflect the genomic landscape of more indolent disease which produces large, resectable metastases. Additionally, the majority of these studies only sampled a small proportion of the primary tumour therefore potentially underestimating the variation within the primary tumour and overestimating the variation between the primary and metastatic disease; Brannon *et al* (2014) performed the most extensive sampling of the primary tumour and displayed the lowest rate of discordance between primary and metastatic disease [81].

A single study [83] has incorporated multiple matched deposits from primary tumours, resected metastases and biopsies from unresectable lesions including brain and lung metastases. This study used a targeted panel of 100 'driver' genes with very high depth NGS (to a depth of 1500x) and demonstrated only 4 instances of intertumoural heterogeneity within 27 cases of aCRC. This study demonstrated strong methodology regarding sequencing depth, however it once again included only a relatively small number of metastases from each individual, therefore possibly underestimating the degree of heterogeneity present in the fatal, disseminated disease state. The collection of cadaveric samples analysed in section 3, 4 and 5 therefore provide a more complete examination of heterogeneity within an individual and although the cost of ultra-deep

sequencing is prohibitive in this study, through collaboration with the Wellcome Trust Sanger Institute (WTSI) it was possible to perform WGS in four cases from our cohort. Within the United Kingdom WGS has gained increased prominence due to the 100 000 genome project, which proposes to decode the genomes of an array of congenital and neoplastic illnesses [331]. The preliminary approach adopted to mutations in carcinomas has been to examine the WGS data at the loci of 74 'actionable' drug targets, followed by a collection of 590 'driver' genes i.e. genes implicated in carcinogenesis; this sequencing will be performed up to a depth of 30x. Although it remains to be seen whether this approach will provide substantial data which will inform the treatment of advanced cancer, the approach of only examining a specific subset of genes provides a method by which the large volume of data produced by WGS may interrogated to produce an understandable, concise dataset.

In this section, we will use the WGS data (from sequencing at 60x) to describe the overall pattern of allelic variation within a group of our cadaveric cases and examine the data at a number of driver loci.

6.2 Aims of this section

1) Identify mutational heterogeneity with a subset of ‘Gift’ cases

2) Compare the pattern of mutational change between the dMMR and pMMR ‘Gift’ cases

3) Correlate anatomical and CNA data with the pattern of mutational alteration

4) Investigate heterogeneity within a group of driver mutations from CRC and other malignancies.

6.3 Materials and methods

6.3.1 Case selection

Four cases were selected for WGS by NGS (GD1, GD3, GD4 and GD6), only four could be sequenced due to the high cost of WGS. This sequencing was performed at a depth of 60 x i.e. the sequence for each point within the genome was determined by examining 60 different overlapping but non-duplicate DNA fragments. A fifth case, GD8, underwent NGS (at a depth of 200x) for a mutational panel, targeting known cancer 'driver' genes, these loci were isolated and amplified via a technique called 'target capture' which will be described below. All library preparation and sequencing for this section was performed at the WTSI in Hinxton, Cambridgeshire by the Cancer Genome Project group under the supervision of Dr U McDermott. All DNA extraction and initial QC was performed by TP at WTBB and data analysis was also performed by TP with assistance from Dr P Tarpey (Cancer Genome Project Group).

The cases were chosen as they represented both the dMMR (GD3 and 4) and CIN/pMMR CRC (GD1, 6 and 8) within our cohort, additionally the CIN CRC selected represented the two ends of the spectrum of clonal complexity as identified through CNA analysis.

6.3.2 Autopsy procedure

The examination procedure was described in section 3 and, according to the protocol, matching FFPE and frozen tissue samples were taken from each tumour deposit sampled therefore the samples used for WGS in this section is comprised of tumour adjacent to that sequenced in section 4 and 5. The sections of tissue used for WGS were those preserved in liquid nitrogen immediately at the time of autopsy and stored at -80˚C on L4 of WTBB. FFPE material was suitable for the targeted capture panel.

6.3.3 Histological staining

Sections of frozen tissue were sectioned at 7µl using a Leica 3050 cryostat (Leica Biosystems, Nussloch, Germany), tissue was mounted using molecular grade water.

To confirm the presence of colorectal adenocarcinoma, one section from before and after the sections used for DNA extraction were stained with H and E using the following protocol:

- Sections placed in distilled water
- 2 minutes (min) Meyers haematoxylin
- Rinsed in running tap water
- 2 min Scott's tap water
- Rinsed in running tap water
- 1 min eosin
- Rinsed in running tap water
- Sections dehydrated in ethanol and xylene
- Mounted in DPX

6.3.4 DNA extraction

The regions of adenocarcinoma were outlined on the H and E sections from each tumour sample and macrodissected from 10 x 7µl sections of frozen tissue. DNA was extracted from the dissected tissue, once again, using the Qiagen QiAMP DNA micro kit according to the manufacturer's protocol (Qiagen, Crawley, UK) for fresh/frozen tissue.

6.3.5 DNA quantification

The concentration of nucleic acid within each sample was initially quantified using a Nanodrop-1000 spectrophotometer (Thermo Fischer Scientific, Loughborough, UK). This measurement was used to dilute the extracted samples for dsDNA quantification using the Quant-iT dsDNA Assay Kit (Thermo Fischer Scientific, Loughborough, UK) and Fluroskan Ascent Microplate Fluorometer (Thermo Fischer Scientific, Loughborough, UK). Further quantification was performed at the WTSI using the Quant-iT dsDNA Broad range assay-it kit (Thermo Fischer Scientific, Loughborough, UK) and Fluostar Omega BMG platereader (BMG Labtech GmbH, Ortenberg, Germany).

The quality of DNA (as represented by the fragment size of the extracted DNA) present within the samples was also assessed. This was required as the sequencing for this section was performed on the Illumina X ten NGS platform; this device requires higher quality, less fragmented DNA, as it produces reads of 350bp compared to the 78-150bp read length of the other Illumina sequencers used in section 5 [332]. The fragment size of extracted DNA was established by the running the samples on a 2% agarose gel and

the Bioline Hyperladder 1kb kit (Bioline Reagents ltd., London, UK) by the following protocol:

- Sample DNA diluted to 1 in 25 with molecular grade water
- 10 µl of each sample placed combined with 5 µl of Bioline 5x Sample loading buffer
- Mixture run on a 2% agarose gel for approximately 1 hour at 200 volts against the Bioline Hyperladder 1kb ladder
- Gel visualised using Bio-Rad Gel Doc XR+ system (Bio-Rad Laboratories ltd., Hertfordshire, UK) (Figure 39)
- Samples with unfragmented DNA (fragment sizes predominantly > 10 000 bp) selected for sequencing

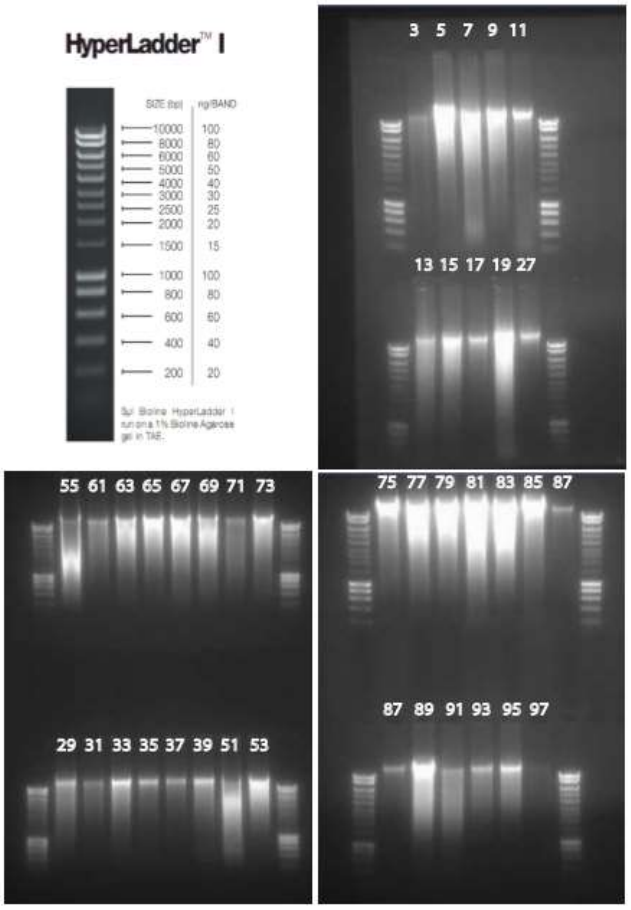


Figure 39: 2% agarose gel displaying DNA fragment sizes from cadaveric tumour DNA

6.3.6 Library preparation

6.3.6.1 WGS Library preparation

Library preparation was performed at the WTSI using Agilent SureSelect QXT Library preparation kit and the Agilent Bravo Automated Liquid Handling Platform (Agilent Technologies, Santa Clara, USA) according to the manufacturer's protocol. This protocol was performed using 50 ng of sample DNA diluted to a concentration of 10 ng/µl with molecular grade water.

As stated above the protocol was performed using an automated process and although the steps performed are similar to those used in the protocol described in section 5, the specific methodology differs slightly and comprises of the following steps:

6.3.6.1.1 Combined enzymatic fragmentation and adaptor ligation

The Agilent SureSelect kit started with an enzymatic process to fragment DNA, in contrast to the sonication method used in section 5. There is some evidence to suggest that enzymatic DNA fragmentation maybe more consistent than sonication based fragmentation [333] however the major advantage of this methodology is that it can, more easily, be incorporated into a fully automated process such as used at WTSI.

The fragmentation step was combined with adaptor ligation, which as previously described, involved the addition of a dsDNA molecule to the fragmented sample DNA. This molecule allowed binding of sample DNA to the sequencer flow cell. These processes were performed by incubation of sample DNA @ 45˚C for 10 minutes with a proprietary enzyme/reagent mix (Agilent SureSelect QXT ILM) and buffer, the fragmentation/adaptor ligation reaction was halted via addition of SureSelect QXT Stop solution.

6.3.6.1.2 Purification/size selection (using Agencourt AMPure XP beads)

As in the previous library preparation technique, this protocol used AMPure paramagnetic beads to bind large fragments of adaptor ligated DNA from the sample-reagent mix. The AMPure bead mix was added at a ratio of 1:1 (which should bind all DNA fragments larger than 200 bp), the mix was placed on a magnet, the supernatant was discarded and ethanol washes were performed followed by elution into molecular grade water.

6.3.6.1.3 Index primer annealing and PCR amplification

The third step comprised addition of primers for PCR amplification and primers for indexing or 'tagging' of individual samples; PCR amplification of the adaptor and index primer ligated DNA was also incorporated into this step. The exact method performed at WTSI diverged slightly from the manufacturers protocol as the primers ligated to the 5' and 3' prime ends of the DNA were customised, in-house molecules rather than those provided by Agilent as part of the SureSelect kit.

The purified DNA from the previous steps was combined with the customised 5' and 3' index primers, a PCR 'master mix' comprising of a nucleotide mix and DNA polymerase, along with a polymerase buffer and dimethyl sulphoxide (which provided optimal conditions when held at the temperatures provided by the manufacturer which are displayed in Table 17).

Table 17: Thermal cycler conditions for primer annealing and PCR reaction with the Agilent QXT SureSelect kit

Cycles	Temperature (˚C)	Duration
1	68	2 minutes
1	98	30 seconds
5	98	30 seconds
	56	30 seconds
	72	1 minute
1	4	Hold

6.3.6.1.4 Purification

A final purification was performed using AMPure beads at a ratio of 0.7:1 (beads:sample) binding only large fragments of adaptor and primer bound DNA, allowing used reagents, small fragments of sample DNA and adaptor dimers to be discarded. After wash steps were performed, the sample to be sequenced or 'libraries' were eluted into molecular grade water.

6.3.6.1.5 Library Quality Control

Final library preparation quality control was performed using the Agilent 2100 Bioanalyzer (Life Technologies), according to manufacturer's protocol, to confirm a mean fragment size within the sample between 600 and 1000 bp. The concentration of each library was confirmed using the Quant-iT dsDNA High-Sensitivity assay kit (Thermo

Fischer Scientific, Loughborough, UK), samples were pooled to a concentration of 3 nanomoles (nM) per litre and submitted for sequencing using the Illumina X ten sequencing platform.

6.3.7 Target capture (or 'enrichment') library preparation

This is the process by which a gene panel (listed in appendix 4 – Target capture gene panel) was captured, amplified, purified and sequenced. This methodology was preferred as it allows a greater number of targets to be sequenced as compared to traditional amplicon based methods (such as pyrosequencing) and produce a greater uniformity and reliability of sequencing than more recent array techniques incorporating large or whole exome gene panels [334].

The library preparation protocol for this method required 100 ng of DNA, which was extracted from FFPE material in case GD8 as described in section 4. The sequencing was performed on the Illumina HiSeq 2500 platform, which produces shorter reads than the HiSeq X ten platform, therefore the more fragmented DNA present in FFPE material is suitable. Additionally, as only a relatively small number of loci are sequenced at greater depth (200x compared to 60x in WGS) there are fewer issues aligning reads from fragmented DNA as small or poorly mapping reads can be discarded without a significant impact on the depth of sequencing. As DNA quality was less important to this sequencing technique, gel electrophoresis was also not required to confirm DNA fragment size.

Library preparation was performed using the Agilent SureSelect XT2 target enrichment system for Illumina paired-end multiplexed sequencing and the Agilent Bravo Automated Liquid Handling Platform (Agilent Technologies, Santa Clara, USA). Once again, the manufacturer's protocol was followed with several modifications, this process was automated, but the steps were as described below.

6.3.7.1 DNA fragmentation

As described in section 5 DNA was fragmented via a sonication method (in this instance Covaris E220 focused ultra-sonicator (Covaris, Inc., Woburn, Massachusetts, USA)) with a target fragment size of 150-200 bp. All samples were diluted to a total volume of 50 µl

(a concentration of 2 ng/µl) with 1x TE Buffer and underwent sonication in Covaris micro-TUBES under the following settings:

Table 18: Covaris E220 settings for DNA shearing for targeted capture library preparation

Setting	Value
Duty Factor	10%
Peak Incident Power	175 w
Cycles per Burst	200
Treatment Time	360 seconds
Bath Temperature	4 – 8 ˚C

6.3.7.2 End repair

The fragmented DNA was then repaired to allow ligation of adaptors; this was done by combining the fragmented sample DNA with SureSelect End Repair Enzyme and nucleotide mix within a thermal cycler @ 20 ˚C for 30 mins.

6.3.7.3 Purification / Size select

As in previous protocols AMPure beads were then used to bind the larger fragments of end repaired DNA (beads were added at a ratio of 1.8:1) and wash steps (using 70% ethanol) were performed. The sample DNA was then eluted into molecular grade water.

6.3.7.4 A' addition

An adenosine group was then added to each DNA fragment, the addition of this base provides a point of binding for index adaptor ligation. This step was performed using SureSelect dA-Tailing Master Mix in a thermal cycler @ 37 ˚C for 30 mins.

6.3.7.5 Index adaptor ligation

The next step involves the addition of index adaptor or 'tag' to the 3' end of the DNA fragments; as described in section 5 the addition of unique, short strands of DNA (or 'tags') to the DNA fragments in a sample allows multiple samples to sequenced simultaneously and separated by the presence of the 'tag' sequence within the read produced from the DNA fragment.

The sample DNA was combined with the SureSelect Ligation master mix along with an indexing adaptor unique to that sample and incubated at 20˚C for 30 mins.

6.3.7.6 Purification

AMPure beads were used to draw the indexed DNA fragments from the used reagents (at a ratio of 1.2:1 (beads:sample)) and the sample DNA was eluted into water. The larger volume of sample to bead on this occasion was required as the size of sample DNA fragment has increased following adaptor ligation.

6.3.7.7 PCR amplification

Half of the indexed sample DNA was then amplified using the SureSelect Herculase II PCR master mix, this step was combined with PCR primer addition (which allows binding of PCR primers to the sample DNA fragments) within a thermal cycler under the following conditions.

Table 19: Table displaying thermal cycler settings for pre-capture PCR amplification

Number of cycles	Tempature (˚C)	Time
1	98	2 minutes
5	98	30 seconds
	60	30 seconds
	72	1 minute
1	72	10 minutes
1	4	hold

6.3.7.8 Purification

The amplified library was then separated from the reagents via a further addition of AMPure beads at a ratio of 1:1 and eluted into molecular grade water (following wash steps using 70% ethanol).

6.3.7.9 Pre-capture quality control

The quantity and fragment size of DNA in each library was then checked using the LabChip GX Touch (PerkinElmer Inc., Waltham, MA, USA) and accompanying DNA High Sensitivity Assay according to the manufacturer's protocol. This electrophoretic technique confirmed the mean fragment size to be between 250 and 275 bp.

6.3.7.10 Hybridisation

This step involves the bonding of the sample library fragments to biotinylated RNA fragments (or 'baits') which were complimentary to specific areas within the genome (in this instance the panel listed in Appendix 4 – target capture gene panel). RNA fragments were used as baits rather than complementary DNA fragments due to the greater binding affinity of RNA [335]. Once bound, streptavidin beads were used to draw the DNA-RNA complexes out, providing a library which only contained the desired fragments of the genome.

To streamline this and subsequent steps, the samples to be sequenced together were pooled at this stage (rather than immediately before sequencing as in section 5); the protocol followed at the WTSI differs from the manufacturer's protocol as the libraries from 12 samples were pooled at this stage rather than 16 as recommended by Agilent. The lower number of samples pooled means a greater concentration of RNA 'baits' per sample (providing a greater yield from the hybridisation step) and additionally as less samples were subsequently sequenced together (or multiplexed) a greater depth of sequencing was produced. Therefore 140.62 ng of each library was pooled and the library pool was concentrated to a total volume of 7µl.

The sample DNA was combined with SureSelect RNase blocking solution, this prevents endogenous RNase from digesting the RNA 'baits' to which the sample DNA was be hydridised. The Capture Library (the biotinylated labelled RNA molecules complementary to the desired genes) and hybridisation buffer were combined with the sample DNA and held at 65 ˚C for 24 hours. During this incubation, the DNA fragments complementary to the RNA fragments became bound or hybridised.

6.3.7.11 Library capture

This step involved the use of Dynabeads MyOne Streptavidin T1 magnetic beads (Thermo Fischer Scientific, Loughborough, UK) which, due to the streptavidin coating, bound the biotinylated RNA capture library and bound sample DNA.

This was done by combining the mix of sample DNA, capture library and hybridisation buffer with the bead solution and incubating the mixture at room temperature for 30 minutes. The beads and bound library were then separated using a magnetic rack; the unbound sample DNA and reagents were then drawn off, the beads washed (using the proprietary SureSelect XT2 wash 2 solution whilst the beads are bound to the magnetic rack) and the sample DNA comprising the regions of interest were then eluted into water.

6.3.7.12 Library amplification

Each captured library was then amplified by 12 cycles of PCR under the same conditions as the pre-capture PCR reaction, the amplified libraries were then purified using AMPure beads. The libraries were then quantified and the fragment size checked using the Agilent 2100 Bioanalyzer with the DNA High Sensitivity kit as per the manufacturer's protocol; the desired fragment size was 275-300 bp and the pool (which contained 12 individually indexed samples) was diluted to a concentration of 6 nM.

6.3.8 Illumina HiSeq NGS for WGS and targeted capture sequencing

6.3.8.1 WGS

The WGS libraries were sequenced on the Illumina HiSeq X ten platform (Illumina inc., San Diego, California, USA) producing paired 350 bp reads, as stated above, the longer read length (as compared with the Illumina HiSeq 2500 or 3000) improves alignment of reads against the reference genome. The target depth for the WGS was 30x for the normal samples and 60x for the tumour samples.

Once sequenced, the bases corresponding to the adaptors were trimmed using 'cutadapt' [317] and aligned to the human genome (hg38) using 'BWA-MEM'. This algorithm is an relatively recent addition to the 'bwa' package which is required to accommodate of the longer read length of the HiSeq X ten platform [336] as compared to previous HiSeq platforms.

6.3.8.2 Targeted capture library sequencing

The targeted capture libraries were submitted for sequencing on the Illumina HiSeq 2500 producing paired 150 bp reads. These reads underwent adaptor trimming using cutadapt [317] and were aligned to the human genome (hg38) using 'bwa' [318]. The target depth for these samples was 200x.

6.3.9 Variant Calling

Both the aligned whole genome and targeted capture sequences were analysed for base substitutions, deletions and insertions using the 'CaVEMan' and 'Pindel' algorithms.

The 'CaVEMan' (Cancer Variants Through Expectation Maximisation) algorithm is an in-house application designed at the WTSI [337]. It is used to identify the presence and probability of single base, somatic mutations in a tumour by comparison with a normal sequence (which is ideally 'matched' i.e. from the same person). The read quality and copy number at the site of any detected substitution are taken into account, allowing filtering of the sequencing data to minimise interference and overcalling of mutations produced by germline variants (genomic polymorphisms present in both the normal and tumour DNA) and those arising in areas of poor coverage which may not be reliable according to the parameters devised at the WTSI. This algorithm has been used for SNP calling in the previous work describing tumoural heterogeneity in renal cell carcinoma [105].

'Pindel' is an algorithm used to detect deletions up to 16 kb in size and insertions of up to 20 bases [338]. This application is able to do so by analysing sequencing 'reads' which cannot be mapped in their entirety to one point by the genome, these fragments are then examined in both directions i.e. backwards and forwards, to identify the disparate points in the genome to which the unmappable reads can be aligned; it is then possible to identify the deleted or inserted base sequence. This algorithm has been validated on several cancer datasets collated at the WTSI [338,339].

6.3.10 Hierarchical clustering

The data produced by 'CaVEMan' and 'Pindel' amounted to over one hundred thousand genomic changes or 'calls' per sample within the dMMR cases and over five thousand 'calls' per sample for the pMMR cases. Therefore, even after filtering with variant callers, the data was not suitable for semi-automated analysis and in order to give an overview of the pattern of mutation present the variant caller data was analysed using hierarchical clustering by 'R' as described below.

For each case, a table was produced comprising of all the variants identified in each case on one axis and the sample names in the other; the loci were assigned a '0' or '1' to denote 'wild type' or 'mutant' placed for each sample. Table 20 is an example of this format.

Table 20: Table showing an example dataset used for clustering by 'hclust'. The column heading shows the case and sample number, the row heading shows a list of all variants identified in the case (the format of the data in this column is chromosome number:base number:normal sequence:variant sequence). The '1' within the grid represents the presence of this variant in the corresponding sample, '0' means the corresponding sample is wild-type for the variant.

	Sample name				
Variant	GD1_44	GD1_60	GD1_70	GD1_72	GD1_42
2:81924434:CA:C	1	1	1	1	0
3:31135429:C:CA	0	0	0	0	1
6:75755226:CA:C	0	0	0	0	0

'hclust' (a function in 'R') was then used to cluster the samples per case by the similarity in their mutational profiles; this method of hierarchical clustering is well-established in the analysis of complex, high-volume biological data [340,341]. This clustering method is a two-step process, firstly it assigns a location in a multidimensional space to each sample and then calculates a 'distance' between each sample according to the data provided (in this instance the data was entered using the 'binary' setting for the application). The algorithm then groups the samples or groups of samples according to which sample(s) are closest until they are combined within a single common cluster; this method of clustering samples in a 'bottom-up' fashion is referred to as 'agglomerative' clustering. The data is presented in a dendrogram similar to those in previous sections, although in this instance the length of the arm between each branch or 'node' is proportional to the distance between each sample/group of samples as described in Figure 40. A second important distinction between this clustering method and that used in section 5 and 6, is that this method only demonstrates the relative similarity between samples and it is not possible to infer the origin of lesions or order of events from the dendrograms.

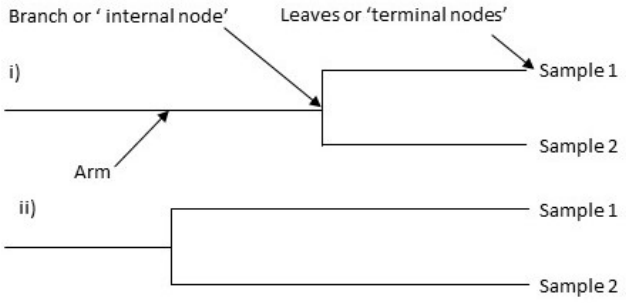


Figure 40: Example dendrograms showing the topology of a dendrogram and two example dendrograms. In i) the arms leading from the branch or 'internal node' to the leaves are smaller than those in ii) representing a greater similarity between the two samples.

6.3.11 Manual interrogation of sequencing data

To explore the presence of oncogenic/'driver' mutations within our dataset, we chose specifically examine 1034 'driver' loci (listed in Appendix 5- mpileup 'driver' loci) used as part of the solid tumour panel at the WTSI; these loci were identified as part of an as yet unpublished cohort of matched CRC primary and metastatic tumours analysed at the WTSI along with whole exome data from breast carcinomas and mixed cancer cell lines including some non-epithelial malignancies [342–345]. This method produced a more manageable volume of data and although not an exhaustive list of 'driver' loci, it provides a large number of changes by which to examine the potential emergence of new driver mutations or 'variants' during disease progression. The read depth and variant allele frequency (the percentage of bases carrying the 'abnormal' bases at a given locus (VAF)) was determined using the 'mpileup' function of 'SAMtools' [346]; this application provides the count of different bases detected at a given base position and chromosome, the VAF was then calculated as follows:

$$VAF = 100 \times (\text{no. of variant reads/total read count})$$

In order to filter out unreliable variants but to include variants shared across multiple samples, the data was filtered to include those which occurred in at least one sample in a region of sequencing depth of at least 30x with a VAF of at least 10%. These thresholds were set according to the depth to which the matched normal samples were sequenced (and therefore represented the lowest acceptable depth of coverage) and

comparable work when analysing degraded DNA from post mortem samples of breast carcinoma [97]. Variants arising in the normal sample were also excluded.

Those mutations which had a VAF of greater than 5% and were supported by more than one read were also extracted; this threshold was selected as this is the limit of detection of pyrosequencing.

6.4 Results

6.4.1 Sample numbers

In total frozen tissue was available for 111 samples, 87 of which were suitable for WGS on the Illumina X ten platform; the breakdown of samples per case is shown in Table 21. The DNA quality was poorest in GD3 and only 4 samples (including one normal sample) contained unfragmented DNA of quality suitable for WGS; the quality of material was significantly higher in the other three cases in which the majority or all samples successfully sequenced.

Table 21: The number of fresh frozen samples suitable for WGS from each case

Case Number	Number of samples passed quality control
GD1	21/34 (incl. 1 normal sample)
GD3	4/14 (incl. 1 normal sample)
GD4	35/35 (incl. 2 normal samples)
GD6	28/28 (incl. 2 normal samples)

All 77 samples submitted for the targeted capture panel (including six normal samples) passed pre-library preparation quality control checks.

6.4.2 Depth parameters

As described in Materials and methods, the target depth for the WGS samples was 30x for the normal and 60x for the tumour samples, the mean depth per case is shown in Table 22. It is acknowledged that the normal sample for GD1 fell short of the target read depth, however the normal sample only fell short by on average a single read and was essential for further analysis, therefore it was not discarded. The sequencing depth for each sample is presented in Appendix 6 -mean sequencing depth for each sample.

Table 22: Table displaying mean whole genome sequencing depth for the normal and tumour samples from GD1, 3, 4 and 6.

Case number	Sample type	Mean depth
GD1	Normal	28.6
	Tumour deposits	63.6
GD3	Normal	30.8
	Tumour deposits	67.7
GD4	Normal	30.9
	Tumour deposits	60.2
GD6	Normal	32.2
	Tumour deposits	62.5

The mean read depth for the targeted capture panel is shown in Table 23. All the normal and tumour samples were sequenced to a level greater than the desired 200x coverage. The mean read depth per sample is also in appendix 6; it is of note that the samples numbered A6-A20 were those originating from resected material, and these samples produced, on mean, more than the double read depth per sample than the cadaveric material (680.6 as compared to 308.9).

Table 23: Table displaying the mean read depth for the normal and tumour samples from the sequencing of the targeted capture libraries for GD8

Case number	Sample type	Mean depth
GD8	Normal	384.8
	Tumour	366.5

6.4.3 'CaVEMan' and 'Pindel' WGS output

The sequencing data for each case was analysed using 'CaVEMan' for the detection of single base substitutions and 'Pindel' for deletions and insertions; the number of each variant type per case is displayed in Table 24 and in Figure 41. Unfortunately, one of the three samples taken from GD3 did not produce data suitable for analysis by the variant callers.

Table 24: Table displaying the number of variant calls produced from WGS for each case. Sample numbers in parentheses.

Type		GD1 (20)	GD3 (2)	GD4 (33)	GD6 (26)
Substitutions	Per case	318303.0	150509.0	5602943.0	155965.0
	Per sample	15915.2	75254.5	160084.0	5998.7
% of calls		95.6	45	55.2	89.7
Deletions	Per case	7180.0	159130.0	3376917.0	6906.0
	Per sample	359.0	79565.0	96483.4	265.6
% of calls		2.2	47.6	33.2	4.0
Insertions	Per case	7130.0	24506.0	1174511.0	10649.0
	Per sample	356.5	12253.0	33557.5	409.6
% of calls		2.1	7.3	11.6	6.1
Complex	Per case	169.0	86.0	1785.0	288.0
	Per sample	8.5	43.0	51.0	11.1
% of calls		0.1	<0.1	<0.1	0.2
Total	Per case	332782.0	334231.0	10156156.0	173808.0
	Per sample	16639.1	167115.5	290175.9	6684.9

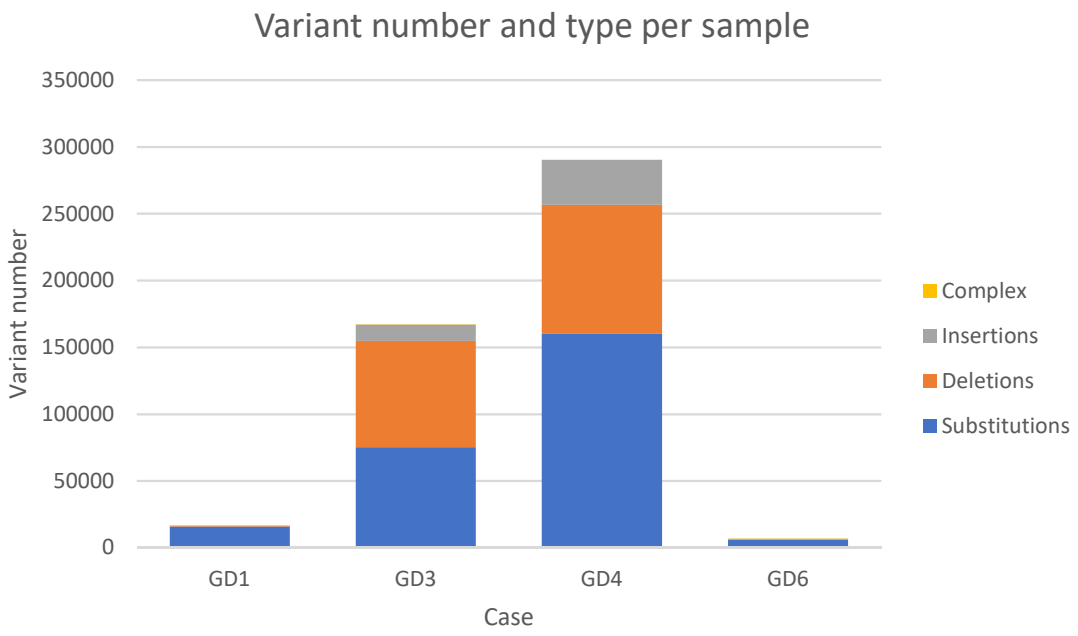


Figure 41: Histogram showing mean number and type of variants identified by 'CaVEMan' and 'Pindel' per sample for the cases submitted for WGS

In total, the two variant callers detected over ten million variants within the four cases analysed and the large majority of these were present in GD4 (one of the two CRC analysed which were dMMR according to IHC); however even in the less aberrant pMMR cases over six thousand changes were detected per sample. Within the pMMR samples over 90% of the variants identified were substitutions, whereas in the dMMR the variant type was more evenly distributed between the substitution, insertion and deletion subtypes. Within the cases there were a variable number of shared variants, variants exclusive to a subgroup of lesions or single lesion were also present; the number of these is described with each case below

The variant callers demonstrated non-synonymous mutations in genomic loci related to mismatch repair in both dMMR cases (these events were also non-germline). *MSH3* and *MSH6* mutations were present in both samples from GD3, although both samples were taken from the same tumour deposit.

All samples from GD4 contained a mutation in the gene *POLQ,* a mutation in *MSH2* was present in all but seven samples from this case. Of these seven wt samples, four originated from large deposits which were sampled and sequenced several times demonstrating *MSH2* mutation elsewhere in the respective lesions.

These changes identified in MMR genes showed a relatively high VAF as displayed in Table 25 and Table 26. No other known oncogenic mutations in *MSH, MLH, PMS, EPCAM* or *POL* genes were identified in either the dMMR or pMMR cases according to the variant callers used.

Table 25: Table presenting variants identified in 'CaVEMan' and 'Pindel' in MSH, MLH and POL genes from case GD3. The y-axis lists the gene and the resultant amino acid change from the variant, the x-axis shows the sample name. The grid shows the VAF of the variant.

		GD3_39_met_peritoneum_abdominal_wall	GD3_40_met_peritoneum_abdominal_wall
Variant		Variant Allele Frequency	
MSH3	p.K383fs*32	35.8	43.2
MSH6	p.F1088fs*2	31.7	47.4

Table 26: Table presenting variants identified in 'CaVEMan' and 'Pindel' in MSH, MLH and POL genes from case GD4. The y-axis lists the gene and the resultant amino acid change from the variant, the x-axis shows the sample name. The grid shows the VAF of the variant.

		GD4_3_met_Liver_lesion_1	GD4_5_met_Liver_lesion_2	GD4_7_met_Liver_lesion_2	GD4_9_met_Liver_lesion_2	GD4_13_met_Liver_lesion_4	GD4_27_met_Diaphragm_lesion_2	GD4_29_met_Diaphragm_lesion_3	GD4_31_met_Diaphragm_lesion_4	GD4_35_met_Diaphragm_lesion_6	GD4_37_met_Peritoneal_deposit_1_duodenum	GD4_53_met_Para_aortic_node_2	GD4_61_met_Para_aortic_node_6	GD4_63_met_Anterorectal_mass_block_1	GD4_65_met_Anterorectal_mass_block_2	GD4_67_met_Anterorectal_mass_block_3	GD4_69_met_Anterorectal_mass_block_4	GD4_71_met_Peritoneal_deposit_2_gall_bladder	GD4_73_met_Peritoneal_deposit_3_bladder_1	GD4_75_met_Peritoneal_deposit_4_bladder_2	GD4_77_met_Right_iliac_fossa_mass_block_1	GD4_79_met_Right_iliac_fossa_mass_block_2	GD4_81_met_Right_iliac_fossa_mass_block_3	GD4_83_met_Right_iliac_fossa_mass_block_4	GD4_85_met_Right_iliac_fossa_mass_block_5	GD4_87_met_Peritoneal_deposit_5_right_iliac_fossa	GD4_89_met_Mesenteric_mass_node_1	GD4_91_met_Mesenteric_mass_node_2	GD4_93_met_Splenic_flexure_mass_block_1	GD4_95_met_Splenic_flexure_mass_block_2
Variant		Variant Allele Frequency																												
MSH2	p.R389*	29.0	32.0	26.0	27.0	0.0	26.0	0.0	23.0	22.0	37.0	32.0	0.0	26.0	26.0	33.0	39.0	32.0	25.0	0.0	25.0	0.0	24.0	29.0	35.0	24.0	0.0	27.0	28.0	0.0
POLQ	p.R860Q	15.0	29.0	12.0	24.0	13.0	19.0	25.0	24.0	22.0	39.0	20.0	14.0	13.0	13.0	18.0	25.0	29.0	30.0	20.0	33.0	23.0	26.0	30.0	22.0	28.0	12.0	31.0	27.0	24.0

6.4.4 Clustering of samples by variant calls using ‘hclust’

As the number of variants per sample was prohibitively large for analysis via a semi-automated method as described in section 5 and 6, the samples in each case were clustered by the presence of variants with greater than one supporting call using an agglomerative clustering algorithm ‘hclust’. As the dendrograms produced are only informative with more than two samples, it was not possible to produce a dendrogram for GD3.

6.4.4.1 GD1 clustering

As displayed in Figure 42, the samples successfully sequenced consisted of three samples from the primary tumour, six from the peritoneal space, five from liver metastases and four from local nodal metastases, along with single samples originating from a vascular embolus and a para-aortic lymph node metastasis.

The variant callers identified variants at 225403 loci; 87% of these were private to one sample and over 95% of the variants were present in five or fewer samples. Only 58 (0.02%) of the variants were shared between all 20 samples, although a further 908 (0.04%) were present in all but one sample, sample 68, which originated in a mesenteric lymph node. This outlying deposit was a small lymph node sample, which was sequenced to the desired depth (as displayed in Appendix 6) and passed all required QC steps; although this tumour deposit may have been genuinely divergent, it is possible a low tumour cell content was present within this sample therefore variants were not identified. The remaining 5494 variants were present in five or more samples. This relatively small group of shared changes is reflected in the shape of the dendrogram displayed below; the distance between the branches is relatively short as compared to the overall length of the dendrogram and the majority of the breadth of the tree represents the differences between the samples, rather than between groups of lesions, suggesting little ongoing evolution through mutation between the groups of lesions.

The clustering of these samples forms two main groups with two outlying samples. The larger of the two main clusters consists of the samples originating from the primary tumour plus those from the liver metastases, vascular embolus and all but one of the local lymph node metastases; within this branch of the dendrogram the clustering of the samples does not appear to correlate with the anatomical location or metastatic route of the individual samples. The other main cluster comprises solely of samples originating from within the peritoneal cavity i.e. those having undergone transcoelomic spread. The two samples not falling into the two main clusters are samples originating from a para-aortic lymph node and one from a local nodal metastasis; this second sample was identified above as sharing few of the variants ubiquitous to all other samples.

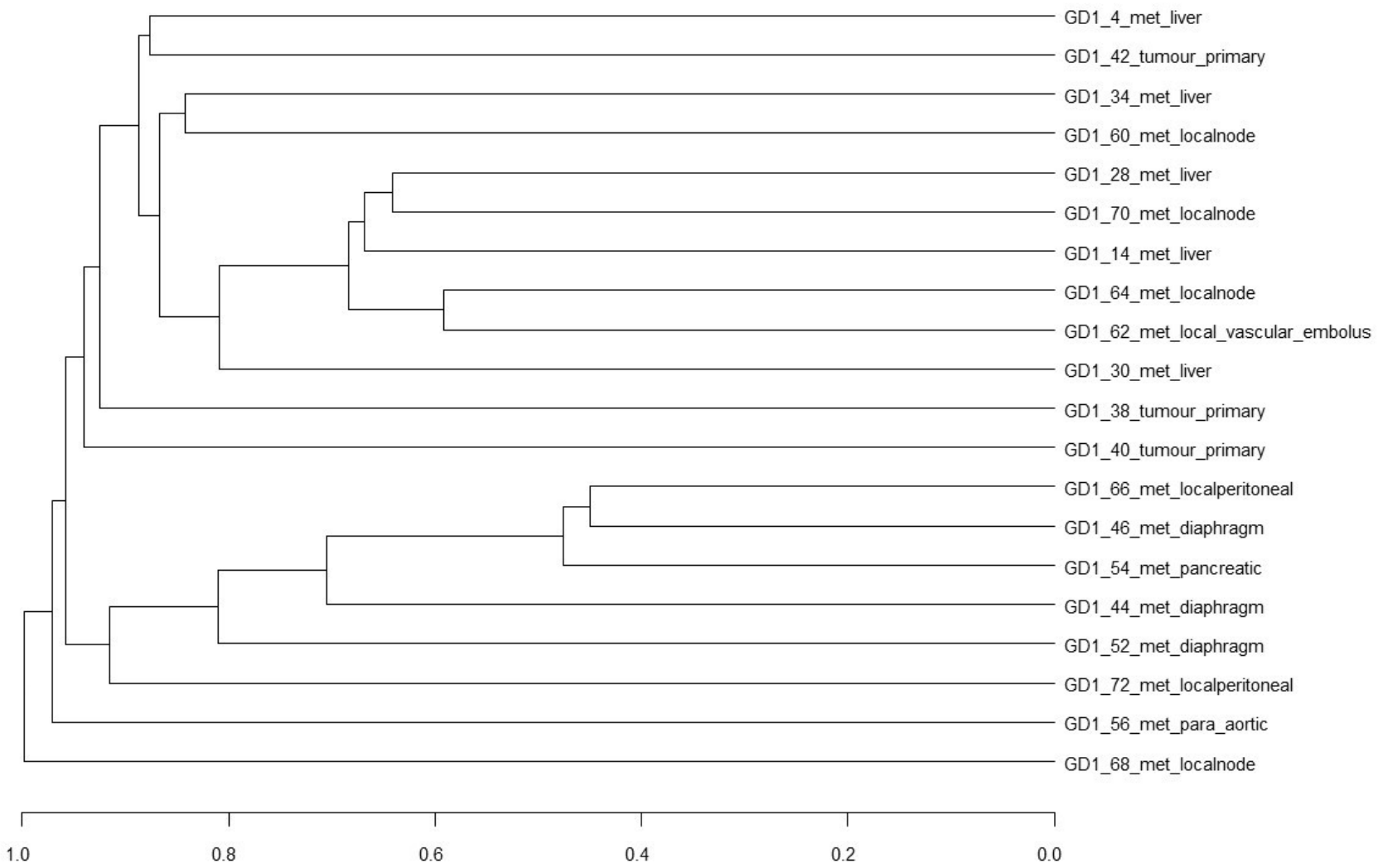


Figure 42: Agglomerative clustering dendrogram from 'CaVEMan' and 'Pindel' variants produced from GD1 WGS data

6.4.4.2 GD4 clustering

The clustering of the samples sequenced from GD4 is shown in Figure 43, this included multiple blocks from large masses in the anorectal and right iliac fossa (RIF) regions (note this is site of probable recurrence of the primary tumour), along with single samples from eleven peritoneal deposits from throughout the abdomen, five liver metastases and four nodal deposits (two local and two para-aortic).

The 10156156 variants identified within the case were present at 677928 different loci, in this instance over 17% were present in all thirty samples, therefore the shared mutational burden was significantly higher within this case as compared to GD1. A significantly smaller proportion of the variants (25%) present were private to a single deposit as compared to GD1. The remaining 58% of variants were present within a sub group of lesions; this evidence of ongoing mutational evolution between groups of samples is reflected in the relatively long arms between branches in the dendrogram below.

The clustering within this case is more complex than GD1, as might be expected with the larger number of samples and genomic events in this case. The samples from the 25 different locations sampled again fall into two main groups; one contains a small subset lesions from diverse anatomical locations and metastatic routes, the second, larger, group contains the samples from the site of probable recurrence (the right iliac fossa) and the remaining metastases. The clustering within the larger subgroup also does not appear to conform to the anatomical location of the lesions nor the potential routes of spread i.e. lesions from each anatomical location are distributed fairly evenly throughout the dendrogram. Additionally, although the majority of samples from the large tumour deposits (in the right iliac fossa and pelvis/anorectal region) do cluster together, there are also samples which appear to resemble anatomically distant disease more closely than that of the tumour immediately adjacent to it. Therefore, in this case we have a diverse group of lesions with a large number of genomic events, the pattern of which appears to suggest that all the lesions in each cluster are able to metastasise to and grow in the hepatic, nodal or peritoneal environments.

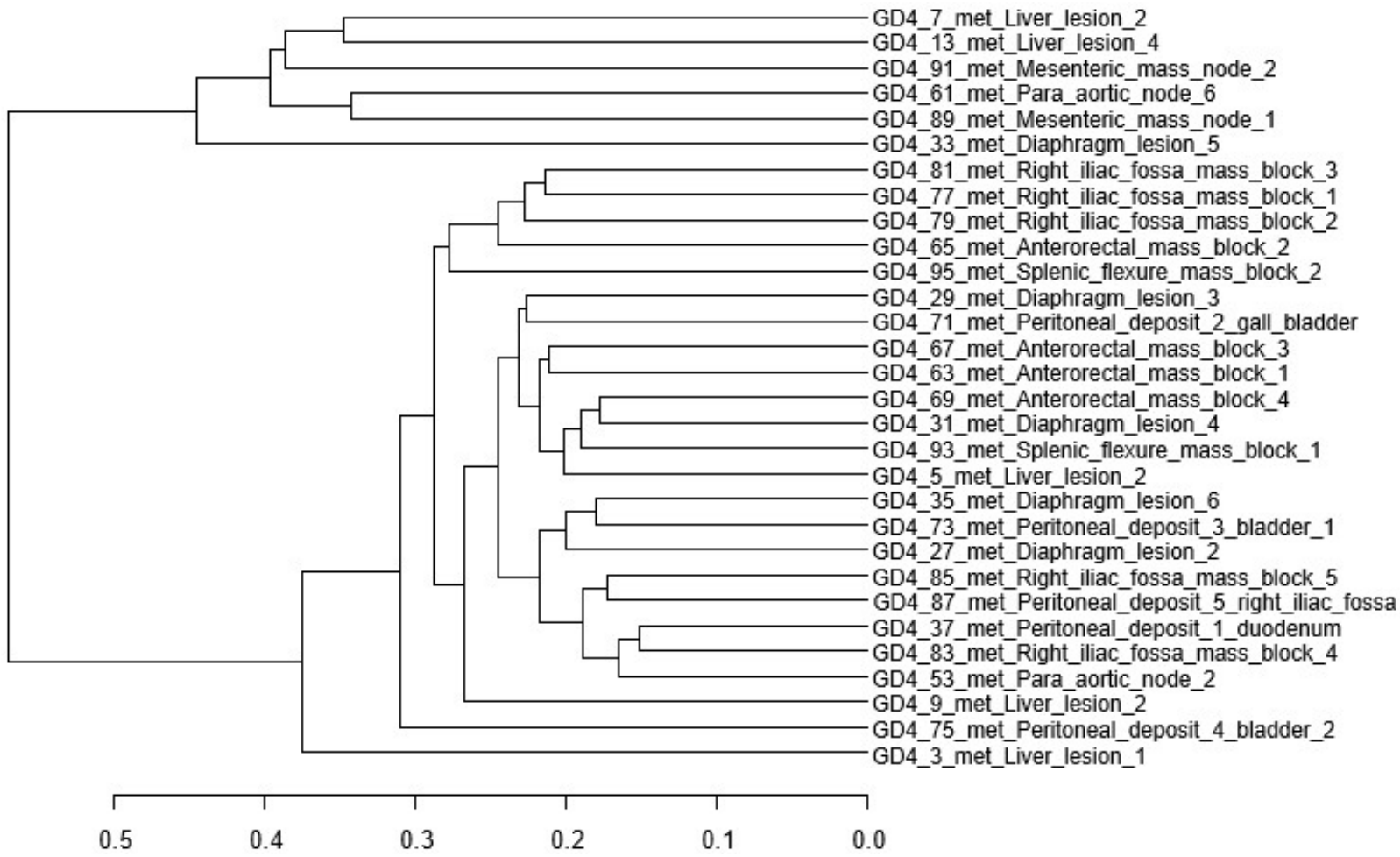


Figure 43: Agglomerative clustering dendrogram from 'CaVEMan' and 'Pindel' variants produced from GD4 WGS

6.4.4.3 GD6 clustering

The clustering for the final WGS case is displayed in Figure 44, twenty-six samples were sequenced from this case, which includes eight samples from the primary tumour, seventeen samples from different liver metastases and a single peritoneal deposit from the colonic serosa adjacent to the primary tumour.

140024 loci were identified by the variant callers in this case, over 2% of the variants were shared across all lesions, 18% were shared across subgroups of lesions and therefore the large majority were private to single samples. This case therefore lies between GD1 and 4 in terms of the degree of ongoing mutational events observed during the natural history of disease, this is illustrated by the relatively spaced branches of the dendrogram and number of genomic variants shared between subgroups of lesions.

According to the clustering algorithm used, there are three outlying samples, all of which were liver metastases, the remainder of the samples fell into two main clusters, the smaller contains exclusively samples from the primary tumour and the larger contains a single sample from the primary tumour and majority of metastases (including the peritoneal metastasis). It is of note that that the primary tumour was sampled from proximal to distal and labelled from A-U with A being the most proximal and U being the most distal, as such sample E is the most proximal sample from the primary tumour which was successfully sequenced and is the only sample from this tumour to cluster with the liver metastases rather than with the tumour adjacent to it. Therefore, in summary, this dendrogram shows an almost mutually exclusive pattern of clustering with a single sample from within the primary tumour grouping with the metastases, suggesting a fairly monoclonal group of metastases (as compared to the previous cases) most closely resembling an anatomically distinct portion of the primary tumour.

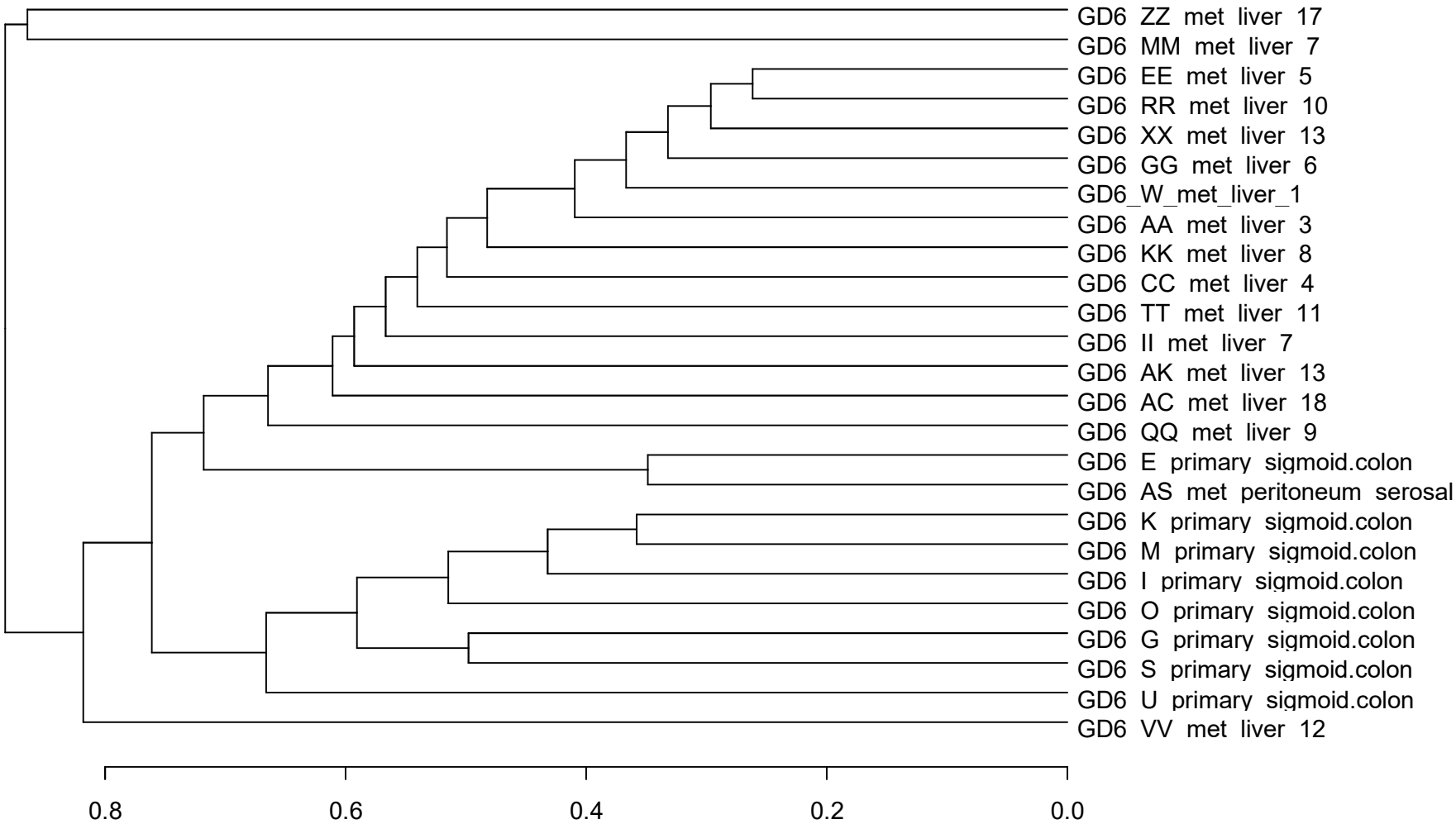


Figure 44: Agglomerative clustering dendrogram from 'CaVEMan' and 'Pindel' variants produced from GD6 WGS data

6.4.5 Driver events identified by WGS

In order to rationalise the large volume of data provided by the WGS, the data was interrogated using the SAMtools function 'mpileup' at the loci listed in appendix 5: 'mpileup' driver loci. This list comprises of 1064 locations within the genome commonly mutated in solid tumours, as identified in large datasets from published literature and unpublished data from the WTSI. The variants identified in more than 10% of reads within regions sequenced to a depth of 30x in each case are described below.

6.4.5.1 GD1 'driver' variants

The mutations identified, regions of sequencing depth greater than 30x with VAFs greater than 10%, in the twenty samples from GD1 are presented in Table 27 (those variants present below 5% are displayed in appendix 7-driver variants with VAFs greater than 5%.). *TP53* mutation was identified in all samples with a VAF suggesting it is present in all tumour cells sequenced, *BRAF* V600K was also identified in a large proportion of tumour cells in all but one tumour sample; the WT sample was only sequenced at a depth of 5x at the relevant locus, it is therefore possible that this is a false negative result. The remainder of the variants presented in the appendix include many mutations commonly seen in CRC including in *NRAS*, *APC* and *PIK3CA*. These lower frequency variants are distributed in a random fashion and additionally these variants were identified in a small subpopulation of the reads sequenced (in fact all other variants were only supported by 3 reads or less) and did not show a VAF of greater than 10%. Therefore, although the heterogeneity identified in this case may be genuine, the differences between samples may be accounted for by sequencing error due to poor depth or sample quality.

Table 27: Table displaying variants in 'driver' genes in GD1. The variants showing a variant allele frequency greater than 10%in regions of sequencing depth greater than 30X in at least one sample. The y-axis lists the gene and the resultant amino acid change from the variant, the x-axis shows the sample name. The grid shows the VAF of the variant.

		GD1_4_met_liver	GD1_14_met_liver	GD1_28_met_liver	GD1_30_met_liver	GD1_34_met_liver	GD1_38_tumour_primary	GD1_40_tumour_primary	GD1_42_tumour_primary	GD1_44_met_diaphragm	GD1_46_met_diaphragm	GD1_52_met_diaphragm	GD1_54_met_pancreatic	GD1_56_met_para-aortic	GD1_60_met_localnode	GD1_62_met_local_vascular_embolus	GD1_64_met_localnode	GD1_66_met_localperitoneal	GD1_68_met_localnode	GD1_70_met_localnode	GD1_72_met_localperitoneal
Variant		Variant allele frequency																			
TP53	V272M	44.3	39.9	33.3	28.1	26.9	20.7	19.8	31.5	23.7	45.8	39.1	23.0	28.9	42.9	46.3	37.5	49.3	37.0	26.7	31.0
BRAF	V600K	14.3	30.8	33.3	44.4	14.3	0.0	21.1	26.3	28.6	48.6	10.0	41.2	22.2	33.3	43.1	31.8	39.6	13.3	23.7	20.0

6.4.5.2 GD3 'driver' variants

Three samples were successfully sequenced from the first dMMR case; the samples sequenced originated from recurrent disease in the pelvis and two samples from the same peritoneal metastasis, which infiltrated into the anterior abdominal wall

As displayed in Table 28 three 'driver' loci were identified as being mutated between the two samples, these were in the *BRAF*, *TP53* and *PIK3CA* genes. The *BRAF* mutation was identified in all three samples in a comparable proportion of the cells sequenced, whereas the other two present were 'private' i.e. only found in one of the samples sequenced. In each case the VAF of the 'private' mutation was relatively high, comparable to the VAF of the 'shared' *BRAF* mutation, and the depth of sequencing at both loci was greater than 60x in both samples; the variants identified in *PIK3CA* and *TP53* therefore represent strong evidence of inter-tumoural heterogeneity in this case. As displayed in Appendix 7, no variants are present with a VAF between 5% and 10% in this case.

Table 28: Table displaying variants in 'driver' genes in GD3. The variants showing a variant allele frequency greater than 10% in regions of sequencing depth greater than 30X in at least one sample. The y-axis lists the gene and the resultant amino acid change from the variant, the x-axis shows the sample name. The grid shows the VAF of the variant.

		GD3_33__recurrence_pelvic	GD3_39_met_peritoneum_abdominal_wall	GD3_40_met_peritoneum_abdominal_wall
Variant		Variant Allele Frequency		
BRAF	V600K	24.4	26.8	27.6
TP53	R249T	0.0	23.6	39.8
PIK3CA	M1004I	23.4	0.0	0.0

6.4.5.3 GD4 ‘driver’ variants

The variants present in the second dMMR case, GD4, are presented in Table 29; as would be expected, due to the large number of samples and mutations identified by ‘CaVEMan’ and ‘Pindel’ and the greater number of samples, a larger number of variants (85) with VAFs greater than 5% are identified at ‘driver’ loci in the thirty samples sequenced from this case (see Appendix 7), however only nine show a VAF of greater than 10% in regions of sequencing depth greater than 30x; these variants are within *RB1*, *PIK3CA*, *SMAD2*, *SF3B1*, *USP9X*, *NF1*, *CTNNB1*, *FBXW7* and *PTEN* genes.

Two variants, those in *KRAS* and *ERBB3*, were ubiquitous across all samples and present in a sufficiently high proportion of reads to suggest that they were possibly present in all tumour cells from each sample. A mutation in *CBFB* was also ‘shared’ across all samples at low VAF; due to the homogeneity of this variant and the fact it occurred in a region of high sequencing depth, despite the low VAF, it is probable that this is a genuine sub-clonal event, but it was not present at a sufficiently high VAF to be presented in the table below. The remainder of the variants satisfying the stipulated threshold were present in only subgroups of tumour samples, however these variants did not occur in a mutually exclusive fashion and, due to the variable depth of sequencing at the same loci in different samples, it is not possible to exclude their presence in all other samples, although it is unlikely that all the WT samples were due to sequencing error. Therefore, in this case we identified ubiquitous, shared variants and a small number of reliable variants which are most likely ‘private’ or only present amongst a subgroup of samples.

Table 29: Table displaying variants in 'driver' genes in GD4. The variants showing a variant allele frequency greater than 10%in regions of sequencing depth greater than 30X in at least one sample. The y-axis lists the gene and the resultant amino acid change from the variant, the x-axis shows the sample name. The grid shows the VAF of the variant.

		GD4_3_met_Liver_lesion_1	GD4_5_met_Liver_lesion_2	GD4_7_met_Liver_lesion_2	GD4_9_met_Liver_lesion_2	GD4_13_met_Liver_lesion_4	GD4_27_met_Diaphragm_lesion_2	GD4_29_met_Diaphragm_lesion_3	GD4_31_met_Diaphragm_lesion_4	GD4_33_met_Diaphragm_lesion_5	GD4_35_met_Diaphragm_lesion_6	GD4_37_met_Peritoneal_deposit_1_duodenum	GD4_53_met_Para_aortic_node_2	GD4_61_met_Para_aortic_node_6	GD4_63_met_Anterorectal_mass	GD4_65_met_Anterorectal_mass	GD4_67_met_Anterorectal_mass	GD4_69_met_Anterorectal_mass	GD4_71_met_Peritoneal_deposit_2_gall_bladder	GD4_73_met_Peritoneal_deposit_3_bladder_1	GD4_75_met_Peritoneal_deposit_4_bladder_2	GD4_77_met_Right_iliac_fossa_mass	GD4_79_met_Right_iliac_fossa_mass	GD4_81_met_Right_iliac_fossa_mass	GD4_83_met_Right_iliac_fossa_mass	GD4_85_met_Right_iliac_fossa_mass	GD4_87_met_Peritoneal_deposit_5_right_iliac_fossa	GD4_89_met_Mesenteric_mass_node_1	GD4_91_met_Mesenteric_mass_node_2	GD4_93_met_Splenic_flexure_mass	GD4_95_met_Splenic_flexure_mass
Identifier		Variant Allele Frequency																														
ERBB3	V104L	25.4	34.1	30.1	35.0	17.3	20.8	29.8	22.1	15.3	32.3	37.8	26.5	17.3	18.5	22.4	28.2	28.3	29.6	20.5	21.3	33.6	21.1	22.6	41.5	27.4	29.7	22.5	24.1	25.3	35.5
KRAS	G13D	7.5	12.5	20.5	23.6	18.2	26.2	16.7	26.3	6.3	23.6	28.0	33.3	11.9	18.4	12.2	18.2	30.3	26.7	20.3	20.8	17.2	26.0	16.7	21.3	20.5	30.1	21.7	21.6	30.6	18.4
USP9X	?	7.0	0.0	0.0	0.0	0.0	3.2	0.0	2.2	10.6	3.1	2.4	2.6	0.0	0.0	3.1	4.8	0.0	2.1	5.6	0.0	5.1	7.3	3.3	0.0	2.1	2.4	0.0	0.0	0.0	0.0
RB1	?	12.5	0.0	2.3	0.0	0.0	0.0	0.0	0.0	1.5	1.4	2.8	1.4	0.0	1.7	0.0	0.0	0.0	0.0	2.5	1.9	0.0	0.0	0.0	1.7	0.0	2.1	0.0	0.0	0.0	0.0
SMAD2	E159*	0.0	2.9	0.0	0.0	0.0	6.4	5.3	0.0	11.5	2.6	2.1	0.0	0.0	7.7	0.0	0.0	4.3	0.0	3.3	6.3	26.1	2.4	2.7	0.0	2.7	4.7	0.0	0.0	2.4	10.0
NF1	Q589*	0.0	0.0	0.0	0.0	5.6	0.0	0.0	0.0	0.0	0.0	2.7	0.0	0.0	5.9	11.8	0.0	0.0	4.3	0.0	1.7	0.0	4.7	0.0	2.5	2.2	1.8	0.0	0.0	0.0	3.7
PTPRB	C1693*	0.0	0.0	1.9	0.0	0.0	1.5	2.2	1.4	0.0	0.0	0.0	0.0	1.3	1.8	2.7	0.0	0.0	5.4	0.0	0.0	0.0	2.5	0.0	1.3	1.3	1.5	1.7	0.0	1.8	1.9
CTNNB1	?	0.0	0.0	12.5	0.0	0.0	0.0	6.3	1.7	2.3	0.0	0.0	0.0	1.8	1.6	1.7	0.0	0.0	0.0	0.0	1.5	0.0	1.0	0.0	1.3	1.5	1.5	1.3	0.0	0.0	11.1
USP9X	R763*	0.0	0.0	0.0	0.0	0.0	1.4	11.1	0.0	0.0	0.0	1.9	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	1.5	1.8	0.0	0.0	3.6	0.0	1.5	1.8	16.7	0.0	3.3
FBXW7	R465H	0.0	0.0	0.0	9.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	6.3	0.0	12.5	0.0	0.0	0.0	0.0	0.0	0.0
PIK3CA	L748I	0.0	0.0	11.8	0.0	15.4	1.8	2.9	2.4	0.0	0.0	2.3	3.6	0.0	0.0	2.6	0.0	3.1	0.0	3.1	0.0	1.8	0.0	0.0	0.0	1.6	3.4	2.3	0.0	0.0	0.0

6.4.5.4 GD6 ‘driver’ variants

The variants present at ‘driver’ loci in the second pMMR case are shown in Table 30. A codon 61 *NRAS* mutation with a relatively high VAF was present in every sample sequenced. Although no other mutations were identified in regions of sequencing depth greater than 30x with a VAF greater than 10%, several other variants (within genes BRCA2 and ARID1A) were present in a large majority of samples at low frequency (these are list in Appendix 7). Six variants (those in *PTEN*, *SMAD2*, *FBXW7*, *ATR* and *APC*) private to a subgroup of samples are supported by a VAFs greater than 10%, however these variants are present in regions of low sequencing depth (less than 20x), the remainder of the variants are identified only with low VAFs and predominantly in regions of poor coverage, therefore there is relatively weak evidence of mutational heterogeneity in this case.

Table 30: Table displaying variants in 'driver' genes in GD6. The variants showing a variant allele frequency greater than 10%in regions of sequencing depth greater than 30X in at least one sample. The y-axis lists the gene and the resultant amino acid change from the variant, the x-axis shows the sample name. The grid shows the VAF of the variant.

		GD6_E_sigmoid colon_primary	GD6_G_sigmoid colon_primary	GD6_I_sigmoid colon_primary	GD6_K_sigmoid colon_primary	GD6_M_sigmoid colon_primary	GD6_O_sigmoid colon_primary	GD6_S_sigmoid colon_primary	GD6_U_sigmoid colon_primary	GD6_W_liver_liver_met_1	GD6_Y_liver_liver_met_2	GD6_AA_liver_liver_met_3	GD6_CC_liver_liver_met_4	GD6_EE_liver_liver_met_5	GD6_GG_liver_liver_met_6	GD6_II_liver_liver_met_7	GD6_MM_liver_liver_met_7	GD6_KK_liver_liver_met_8	GD6_QQ_liver_liver_met_9	GD6_RR_liver_liver_met_10	GD6_TT_liver_liver_met_11	GD6_VV_liver_liver_met_12	GD6_AK_liver_liver_met_13	GD6_XX_liver_liver_met_13	GD6_ZZ_liver_liver_met_17	GD6_AC_liver_liver_met_18	GD6_AS_peritoneum_serosal_met
Variant		Variant Allele Frequency																									
NRAS	Q61L	44.2	34.2	34.8	46.2	32.1	26.7	32.0	18.8	46.8	47.6	42.1	40.4	37.3	26.8	43.6	38.5	50.0	43.8	37.8	40.8	31.4	35.2	43.8	43.9	46.9	24.4

6.4.5.5 GD8 ‘CaVEMan’ and ‘Pindel’ targeted capture sequencing output

Table 31 presents the data produced by filtering the NGS sequencing data produced from the targeted capture libraries from GD8; the variants presented are those in greater than 5% of any one sample with more than one supporting read. As described in Depth parameters the sequencing data produced was of higher depth than that of the WGS cases and the variant callers used identified alterations in four genes at ‘driver’ loci, these are in *KRAS*, *APC*, *TP53* and *FBXW7*. Three of the mutations identified (those in *KRAS*, *APC* and *TP53*) were present in the majority of the samples with only occasional WT samples; no samples were WT type for more than one of the three dominant mutations. The fourth mutation, in *FBXW7*, was present exclusively in the resected primary tumour and was not identified in any of the recurrent tumour nodules; within this case the *FBXW7* mutation identified represents strong evidence of divergence between the primary, resected tumour and recurrent, distant metastases.

Table 31: Table displaying variants identified by 'CaVEMan' and 'Pindel' in GD8. The variants presented were supported by at least 2 reads and showed a variant allele frequency greater than 5% in at least one sample

		1_recurrent_Mediastinal node 1	3_recurrent_Mediastinal node 2	5_recurrent_Mediastinal node 3	7_recurrent_lung	9_recurrent_lung	11_recurrent_lung	13_recurrent_lung	15_recurrent_lung	17_recurrent_lung	19_recurrent_lung	21_recurrent_lung	25_recurrent_lung	27_recurrent_lung	29_recurrent_lung	31_recurrent_lung	33_recurrent_lung	35_recurrent_lung	37_recurrent_lung	39_recurrent_lung	41_recurrent_lung	43_recurrent_lung	46_recurrent_lung	48_recurrent_lung
Identifier		Variant Allele Frequency																						
APC	R554*	12.0	22.0	23.0	31.0	20.0	12.0	31.0	18.0	25.0	22.0	7.1	26.0	33.0	15.0	21.0	26.0	24.0	25.0	33.0	29.0	26.0	24.0	24.0
FBXW7	R465H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KRAS	G12D	8.4	72.0	60.0	31.0	33.0	27.0	43.0	33.0	31.0	33.0	35.0	46.0	60.0	51.0	50.0	38.0	41.0	38.0	27.0	47.0	39.0	42.0	58.0
TP53	G266V	19.0	78.0	70.0	54.0	48.0	31.0	50.0	35.0	53.0	50.0	52.0	37.0	49.0	39.0	48.0	46.0	59.0	57.0	57.0	49.0	41.0	44.0	46.0

		49_recurrent_lung	51_recurrent_lung	53_recurrent_lung	55_recurrent_lung	57_recurrent_lung	59_recurrent_lung	61_recurrent_lung	63_recurrent_lung	65_recurrent_lung	67_recurrent_lung	69_recurrent_lung	71_recurrent_lung	73_recurrent_lung	75_recurrent_lung	77_recurrent_lung	79_recurrent_lung	81_recurrent_lung	83_recurrent_lung	85_recurrent_lung	87_recurrent_lung	89_recurrent_liver	91_recurrent_liver	93_recurrent_liver
Identifier		Variant Allele Frequency																						
APC	R554*	46.0	27.0	38.0	27.0	25.0	22.0	35.0	15.0	18.0	21.0	29.0	28.0	15.0	16.0	26.0	23.0	24.0	30.0	31.0	26.0	21.0	20.0	22.0
FBXW7	R465H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KRAS	G12D	79.0	42.0	46.0	41.0	42.0	47.0	43.0	32.0	47.0	32.0	43.0	42.0	20.0	39.0	26.0	48.0	57.0	42.0	33.0	28.0	22.0	28.0	33.0
TP53	G266V	69.0	62.0	58.0	52.0	66.0	55.0	49.0	45.0	58.0	38.0	50.0	44.0	28.0	51.0	57.0	52.0	60.0	48.0	50.0	34.0	52.0	42.0	39.0

		95_recurrent_liver	97_recurrent_liver	99_recurrent_liver	101_recurrent_liver	103_recurrent_liver	105_recurrent_liver	107_recurrent_liver	109_recurrent_liver	111_recurrent_liver	113_recurrent_liver	115_recurrent_liver	117_recurrent_liver	119_recurrent_liver	121_recurrent_liver	A6_resected_left_colon_primary	A7_resected_left_colon_primary	A8_resected_left_colon_primary	A14i_resected_Local Lymph node 1	A14ii_resected_Local Lymph node 2	A15_resected_Local Lymph node 4	A17_resected_Local Lymph node 5	A18_resected_Local Lymph node 6	A19_resected_Local Lymph node 7
Identifier		Variant Allele Frequency																						
APC	R554*	26.0	40.0	25.0	27.0	9.2	22.0	25.0	20.0	8.1	31.0	16.0	0.0	27.0	29.0	19.0	20.0	12.0	11.0	0.0	0.0	9.3	7.1	5.2
FBXW7	R465H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	14.0	13.0	12.0	0.0	6.6	5.1	7.8	5.5
KRAS	G12D	24.0	26.0	26.0	30.0	28.0	0.0	31.0	15.0	18.0	40.0	39.0	19.0	22.0	22.0	21.0	21.0	19.0	23.0	4.8	5.7	9.1	7.7	5.5
TP53	G266V	33.0	41.0	55.0	44.0	28.0	36.0	35.0	41.0	27.0	56.0	32.0	39.0	39.0	41.0	19.0	26.0	22.0	29.0	5.8	11.0	12.0	8.5	7.0

6.5 Discussion

Within this section WGS was performed on 87 DNA samples, which were extracted from fresh frozen material taken at post mortem examination. This data was analysed at the allelic level using automated variant callers, to demonstrate the number and pattern of small genomic aberrations, and using a function of SAMtools [346] to interrogate specific driver loci to investigate the possibility of heterogeneity at 'driver' loci. NGS was also performed upon libraries generated by 'target capture' library preparation from DNA samples derived from surgical and cadaveric FFPE material from a fifth case (GD8); this more limited sequencing provided a higher depth and quality of sequencing in specific 'driver' genes. The QC, clustering and targeted analyses highlighted differences between and within the cases.

6.5.1 DNA QC

The initial QC steps, preceding the library preparation for WGS, showed a considerable difference in the quality of the DNA between the first and second pairs of cases; all of the samples submitted from GD4 and 6 contained DNA of sufficient quality for WGS, whereas only two thirds and one third of the samples from GD1 and GD3 respectively were acceptable. It is not entirely clear why the disparity between the two sets of cases was present; all donors were refrigerated within 6 hours of death and post-mortems were performed within 48 hours. The same sampling protocol was also followed during the post mortem examination (tissue was 'snap-frozen' in liquid nitrogen at the time of the examination following documentation and photography of each lesion) and all samples were sectioned and extracted by the same individual under the same conditions. The pathological characteristics were also similar between the cases; all tumours were moderate to poorly differentiated adenocarcinomas and all cases comprised of a mixture of small and large tumour deposits. GD1 and 6 both had bulky primary tumours in-situ at the time of examination, GD3 and 4 also had large local recurrences therefore, the potential impact upon DNA quality of large deposits becoming centrally necrotic, either before or after death, is present within all cases. The most probable cause for the disparity between the cases is due to increased experience of tissue sampling during PM examinations; the examinations performed on donors GD4 and 6 occurred with the experience gained from the GD1 and 3, and tissue will likely have been removed and

frozen more quickly, allowing less time for tissue degradation at room temperature during PM examination.

Due to the limited use of the PM as a tool for research in neoplasia there is little literature with which to compare this data. The closest comparable data from adult solid organ malignancy has been published studying breast, pancreatic and prostatic carcinoma, unfortunately QC and sampling data was not included as part of the published data for the latter two studies [87,97,106]. The 'CASCADE' research autopsy program performed "rapid" post mortem examinations in four patients with advanced breast carcinoma and although the precise post mortem interval and pre-library preparation DNA QC is not described, 4 of the 52 samples either failed QC or produced NGS whole exome sequencing data of inadequate quality for analysis [97]. Autopsy material gathered in the setting of paedi*ATR*ic glioma demonstrated that, even with a short post mortem interval (less than 8 hours), DNA degradation was demonstrable by gel electrophoresis in approximately 10% of samples, although it was felt that mildly degraded post mortem DNA was suitable for SNV analysis by an array technique [347]. In summary from the small volume of published data, it appears that even with rapid post mortem harvesting of tumour deposits, a significant minority of samples will exhibit DNA degradation although the significance of this will be dependent on the sequencing type, depth and analytic method used for further analysis.

6.5.2 Variant calling

The number of variants described in the cases clearly correlated with the mismatch repair status identified by immunohistochemistry; GD1 and 6 were thought to be pMMR and GD3 and 4 were dMMR. There was almost a 30-fold difference in the number of variants identified per sample between these two pairs of cases. Within the pMMR cases, using 'CaVEMan' and 'Pindel', we demonstrated a mean of approximately 11 000 variants per sample (consisting predominantly of substitutions), whereas the dMMR cases showed a mean of slightly under 300 000 variants per sample (with a larger proportion of deletions). Both the dMMR and pMMR cases showed a comparable number of mutations as compared to the comprehensive data published by TCGA [65]; the TCGA work demonstrated mutation rates of up to 500 per million bases in hypermutator CRC as compared to GD3 and 4 which showed a mean mutation rate of 93.4 per million bases, the pMMR cases showed a mean of 3.4 mutations per million bases which falls well within the rate described in this literature and other literature [50].

Within GD1 the number of variants private to individual samples is striking, over 95% of variants were present in only one sample. It is probable that at least a reasonable proportion of the private events were artefactual (likely due to alignment, sequencing or caller error due to DNA degradation) as GD1 displayed a degree of DNA degradation (as represented by the number of samples failing pre-library preparation QC) and when genomic loci were examined manually (using the driver loci panel) it was not possible to reliably identify evidence of mutational heterogeneity at loci with good sequencing depth. Additionally, it is counter intuitive that GD1, which was pMMR (according to IHC), should become mutationally diverse at a late stage in the disease process, when the hypermutator/dMMR cases (GD3 and GD4) displayed a more gradual accumulation of genomic diversity with strong evidence of inter-sample heterogeneity. Realistically it is probable that a proportion of the private variants across all cases were false positives and ideally loci demonstrating heterogeneity would be selectively re-sequenced at high depth. However, as there is a large number of samples and variants from each case and the variant callers will have provided a degree of filtering, the overall pattern of variants (especially that produced by shared variants) is likely to be a reasonably accurate method by which to assess the genomic similarity between the samples.

6.5.3 Clustering

The pattern of clustering, as derived from the automated variant calls, displayed further differences between the pMMR cases and the dMMR case suitable for this analysis method. The clustering dendrograms (Figure 42 and Figure 44) from GD1 and 6 displayed a degree of correlation between the grouping of samples and the anatomical origin of the samples. In GD6 only one of the samples originating from the primary tumour more closely resembled the liver metastases than the other samples taken from the primary tumour. The primary tumour in GD1 displayed a greater diversity of mutations, samples originating from this lesion grouped with liver and nodal metastases, although the samples taken from within the peritoneal cavity resembled one another more closely than the primary tumour or metastases at other sites. The dendrogram for GD4 (Figure 43) did not display an appreciable relationship with the location of the lesions with the samples from different locations grouping apparently at random with samples from other sites.

The anatomical association of the clustering in this section is similar to that described in section 3, although on this occasion association has been highlighted by the overall similarity in the genomic events identified in the samples rather than the pattern of mutually exclusive events; this distinction is important as we cannot infer any sequence of events from the clustering of samples.

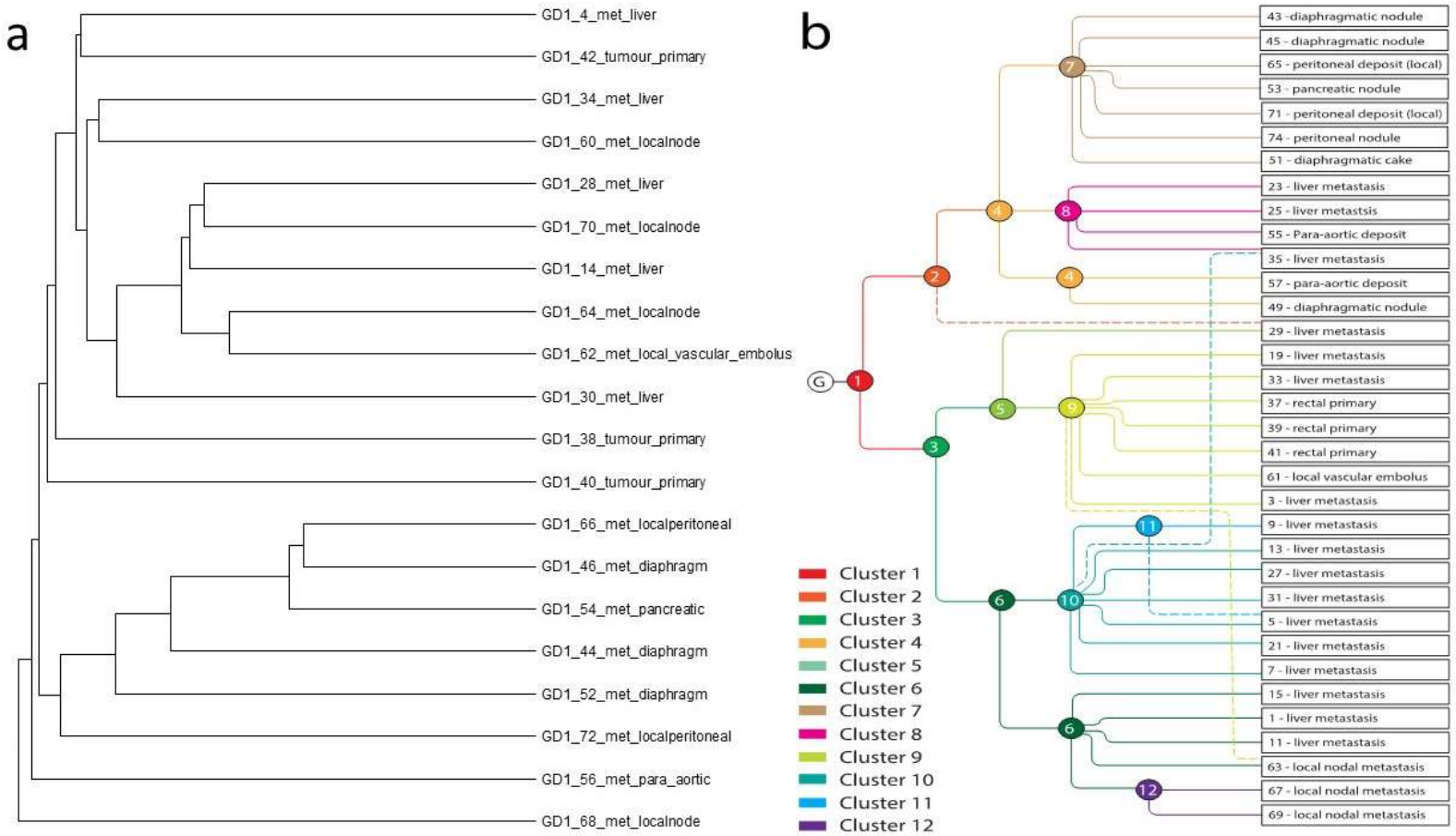


Figure 45: Comparison of dendrograms produced from the case GD1 using a) 'hclust' and b) Dirichletian semi-automated method. Note the lesions sequenced are matched, with the FFPE samples (in plot (b)) given odd numbers and the matched fresh samples (in plot (a)) labelled with the following even number i.e. sample 1 and 2 originate from adjacent tissue within the same tumour deposit

6.5.3.1 Clustering analyses in GD1

Within GD1 the overall pattern of mutational events is comparable to that identified within section 3 and the similarity between the peritoneal deposits is once again identified (a comparison of the two dendrograms is present in Figure 45). The disease originating within the peritoneum (samples 43/44, 45/46, 53/54, 51/52, 65/66, 71/72) clustered together according to both the CNA and mutational analysis. As stated in this previous section, copy number similarities between unmatched peritoneal deposits were identified in an early study [91], although peritoneal disease has not been found to demonstrate characteristic genomic changes in larger series or more detailed studies performed with matched samples [83,92]. However, the more recent data is not directly comparable to the data from this work as they did not contain cases with multiple peritoneal deposits and, at least within this case, there is strong evidence of a relative mutual exclusivity between the peritoneal disease and disease elsewhere in the body.

It is not clear why this pattern of genomic events has occurred, it is probable that the peritoneal disease originated from the same portion of the primary tumour, as if the peritoneal disease originated from multiple portions of the primary tumour it would be expected that the peritoneal disease would reflect the intratumoural heterogeneity observed within the primary tumour from GD1. Additionally one might expected the pattern of mutual exclusivity to be disturbed by the phenomenon of metastasis to metastasis seeding [87,88,103]. A possible explanation for the absence of this disruptive phenomenon would be that the peritoneum produces an enclosed cavity and is not exposed to tumour cells circulating in haematogenous or lymphatic vessels and therefore is not vulnerable to further seeding events. Significantly, peritoneal disease is often refractory to systemic chemotherapy but more responsive to similar agents delivered directly into the peritoneal cavity [348–351] suggesting the peritoneal disease is less exposed to the haematogenous and lymphatic circulation than other sites of metastases and therefore provide specific clinical challenges and opportunities.

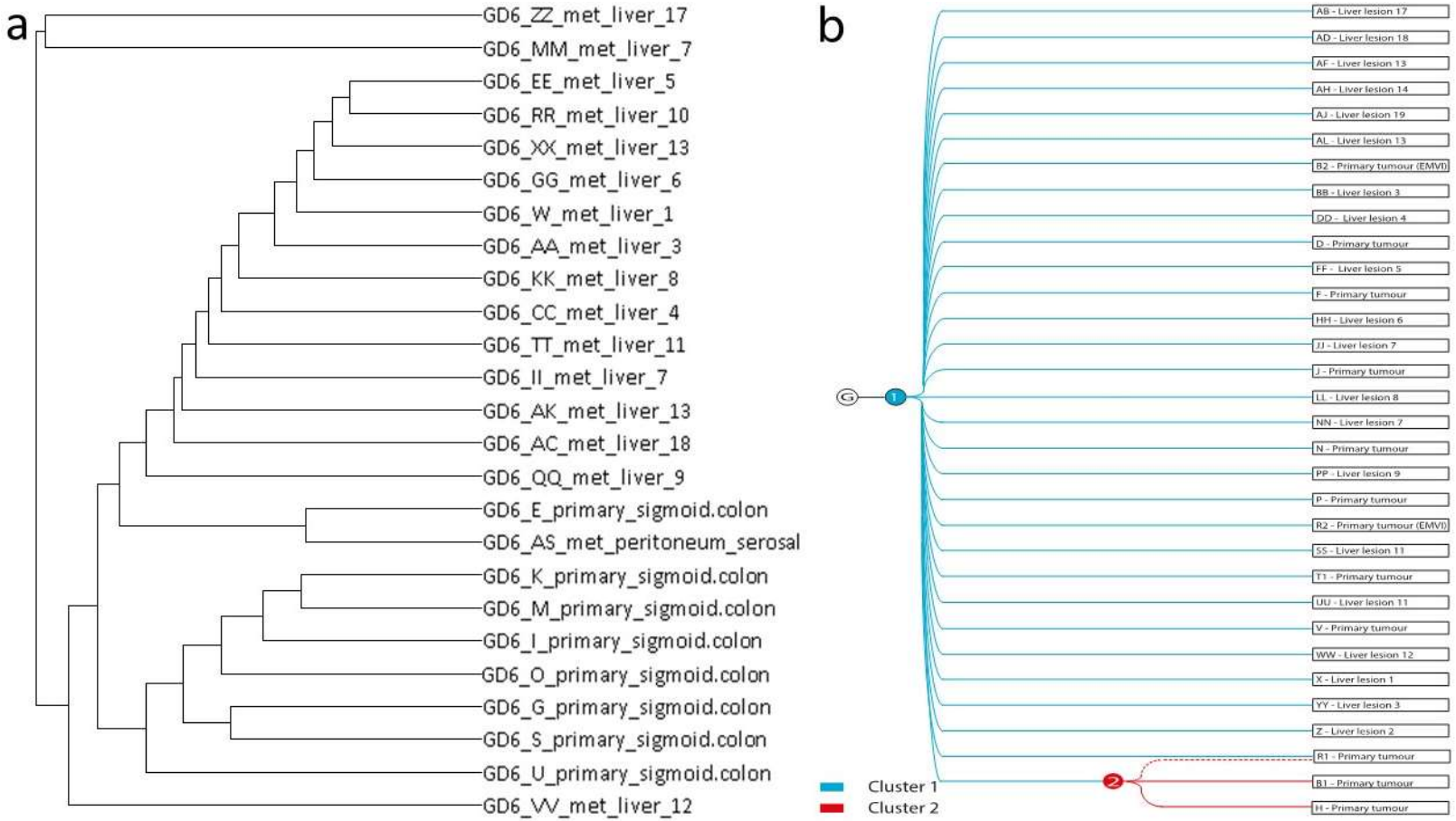


Figure 46: Comparison of dendrograms produced from the case GD6 using a) 'hclust' and b) Dirichletian semi-automated method. Note the lesions sequenced are matched, with the fresh samples (in plot (b)) given a letter and the matched FFPE samples (in plot (a)) labelled with the following letter i.e. sample A and B originate from adjacent tissue within the same tumour deposit

6.5.3.2 Clustering analyses in GD6

As displayed in Figure 46 the pattern of clustering for the WGS data via 'hclust' and that of the CNA analysis differed in GD6. First of all, clustering according to the mutational changes identified divergence and patterns of clustering between samples which was not identifiable from the CNA data; the pattern of clustering conformed broadly to the anatomical location of the metastases as described in section 0.

The only sample from the primary tumour which more closely resembled the metastases via the mutational changes was sample 'E', the matched FFPE sample (sample 'F' in Figure 46b) does not appear to diverge from the either the metastases or primary tumour according to the CNA analysis. Due to the clustering method used, it is not possible to determine the mechanism by which the relationship between sample E and the metastases arose i.e. it is possible that the metastases arose from the region of tumour sampled in sample E or that a clone of tumour cells from a liver metastasis re-circulated to this portion of the primary tumour and produced the observed result. As cited in previous sections tumour to tumour seeding is an established phenomenon in animal and human studies of tumour evolution [87,88,97,103].

Of the samples originating within the primary tumour, according to CNA analysis, two slightly diverged from the remainder of the lesion ('B1' and 'H'), only one of the matched fresh samples was available for WGS (sample 'G'); this sample did not appear to lie out with the primary tumour according to the pattern of mutations. It would appear unlikely that the association identified by CNA analysis is not genuine, as two identical CNAs were identified within the two samples in section 5. It is therefore probable that the differences between the two clustering plots for these samples reflects genuine differences between two adjacent regions of tumour or an artefact of sequencing or clustering method; the former explanation is favoured due to the large number of genomic changes used in the analysis. However, to reach a conclusive explanation would require resequencing of samples either at greater depth or manual interrogation of WGS data, neither of which was possible within the time frame available.

Although minor differences exist between the two analysis methods the overall pattern within case GD6 is of a disease process which appears to have a combination of CNA and mutational changes. The CNA appears to have been an early event in tumour development with ongoing mutational changes later in disease evolution.

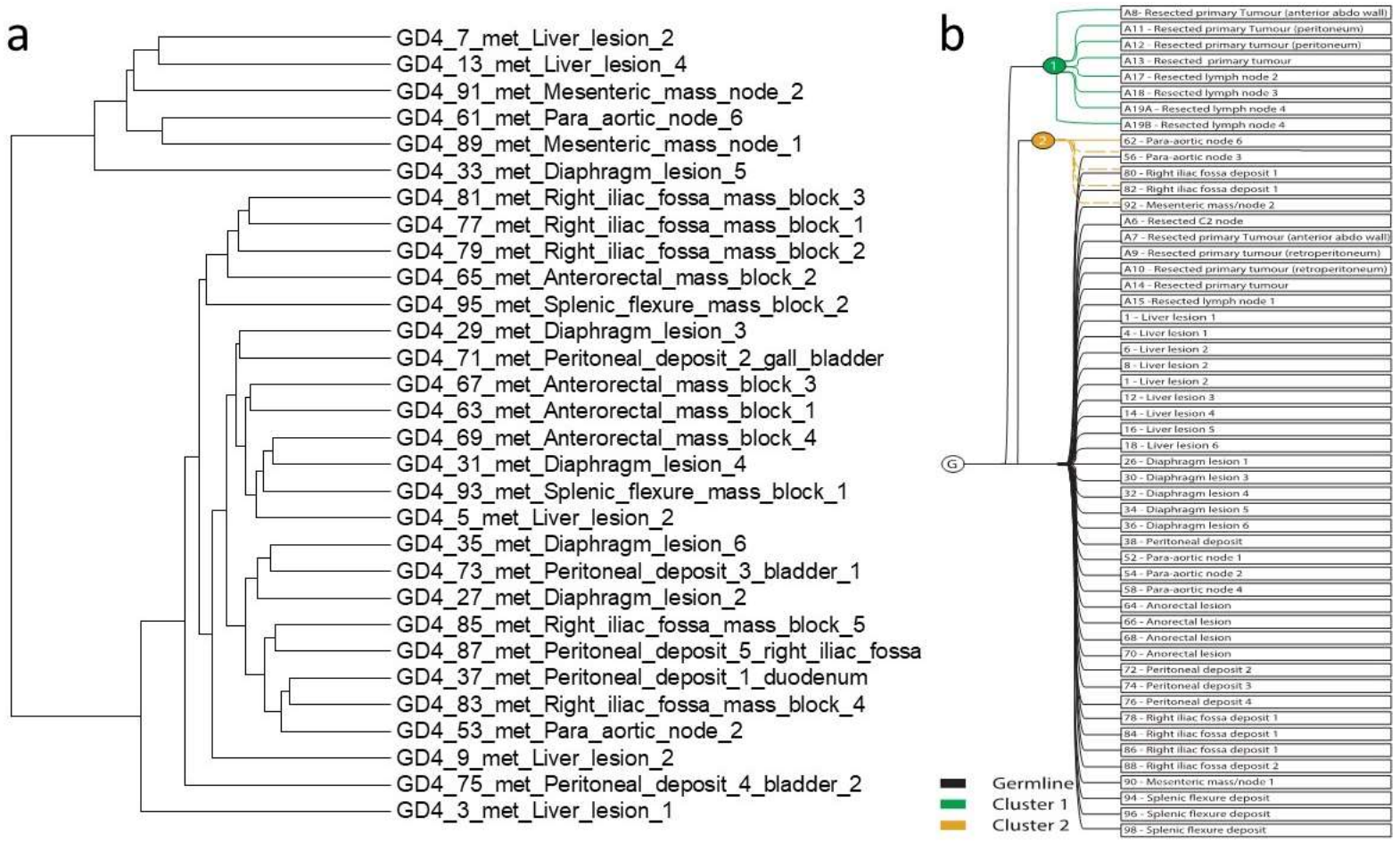


Figure 47: Comparison of dendrograms produced from the case GD4 using a) 'hclust' and b) Dirichletian semi-automated method. Note the lesions sequenced are matched, with the FFPE samples (in plot (b)) given odd numbers and the matched fresh samples (in plot (a)) labelled with the following even number i.e. sample 1 and 2 originate from adjacent tissue within the same tumour deposit

6.5.3.3 Clustering analyses in GD4

In contrast to the pMMR cases, the samples from GD4 clustered in a fairly random fashion and similarity between samples from similar anatomical locations is not apparent in the dendrogram for this case (as displayed in Figure 47). It is possible that the relatively pattern-less clustering is a result of a high burden of circulating tumour cells and multiple large intraperitoneal deposits breaching the peritoneum overcame any functional or biological barriers which dictated the anatomical pattern of clustering observed in the pMMR cases. The data from this case is, however, extremely complex and, even after the attempt to parse the data with variant calling algorithms, it comprises of over ten million variant calls split over thirty-five samples. Whilst it is possible that a more complex analytical method incorporating other sequencing parameters and aspects of the WGS data (such as described in Naxerova *et al* [258]) would produce a more comprehensible result; unfortunately, the current automated methods such as used in the cited literature have been devised for use with a smaller volume of data (usually whole exome or SNP array) produced by this case, and the quantity of data was also in excess of the capacity of the methods available at WTSI at the time of writing.

6.5.4 MMR gene alterations

As identification of genomic lesions at loci related to MMR would be central to the interpretation of the other WGS data, any data relating to *MLH, MSH, PMS, EPCAM* and *POL* genes was extracted from the variant caller data, as this would provide a more comprehensive list of changes than an enquiry at specific loci using 'mpileup'. Variants in MMR related genes were present in both dMMR cases but the pMMR cases did not display any variants; mutations in MSH3 and 6 were present in both samples from GD3 and alterations in MSH 2 were present in GD4. Mutations in MMR related genes were not identified in the germline samples.

The presence of mutations in mismatch repair genes in both dMMR cases was surprising as sporadic mutations in MMR are relatively rare and the majority of dMMR tumours occur as a result of epigenetic silencing of MMR proteins [61,352]. However, in GD4 the MSH2 mutations correlated with the loss of MSH2 and 6 expression identified with IHC (as MSH2/MSH6 and MSH2/3 dimers are unable to form to effect mismatch repair), but the MSH3 and 6 mutations in GD3 was discordant with the immunohistochemical findings, which showed loss of MLH1/PMS2 function. It is difficult to ascertain why this discordance has occurred and it is improbable that this result is a false positive or artefactual as the IHC in section 1 was repeated and the MSH mutations were present at high VAF in areas of good coverage across multiple samples. The specific MSH3 mutation in GD3 is non-synonymous, it is not within the COSMIC database [29] and therefore should be viewed with caution, however the MSH6 mutation is a well-established alteration related to dMMR status [353] and therefore should have been reflected in the IHC findings. The relative rarity of these mutations means that the rate of discordance between mutation and IHC status is relatively unexplored and the discordance in this case requires further confirmation through translational and/or transcriptional analyses, but in essence the presence of a mutation in an MMR related gene is in keeping with the hypermutator genotype identified in this case. It is possible that the initial event producing the hypermutator phenotype in this case is in fact methylation of *MLH1* and the variants described above are subsequent events, and methylation analysis could be included in further study of this cohort to to exclude this possibility.

An additional notable mutation demonstrated in GD4 is a missense mutation in POLQ (also described as POL θ). This is a DNA polymerase molecule and, although less well recognised in CRC than other molecules of the same class, such as POLE and POLD

[65,66], the presence of POLQ mutation has been associated with poor outcome in CRC and other carcinoma types [354–356]. This molecule is thought to have a role in genomic stability, preventing large deletions within the genome, along with a possible secondary role in single base repair [357,358]. The variant identified within this case lies in exon 16 of the gene, which is translated to form part of the central portion of the protein [19,357,359]. The precise role of this region of the protein is unclear, although portions of this region may be important to efficacy of polymerase activity [360]. GD4 therefore showed two mutations (in MSH2 and POLQ) in the majority of samples with high VAF, which play a role in genomic repair and it is possible that this combination of events had a cumulative effect, producing the vast number of substitutions and deletions identified in this case.

6.5.5 'Driver' loci analysis

The 'mpileup' function of 'SAMtools' was used to interrogate the data for mutations at over one thousand loci and identified common driver mutations in all four cases. This more focused approach demonstrated evidence of intertumoural heterogeneity in both the pMMR and dMMR cases; the variation identified within the pMMR cases was present in regions of poor sequencing coverage and was supported by only a small number of reads, whereas the dMMR cases showed better quality evidence of heterogeneity.

It is acknowledged that the list of loci used in the 'mpileup' interrogation is not exhaustive and it is probable a significant number of driver events have been missed, however the list was the best available option at the time of analysis and allowed for an exploration of the principal of mutational heterogeneity rather than a definitive conclusion, which would require a greater depth of sequencing and more cases than currently available. An alternative methodology attempted was to filter the variant caller data from the four cases for variants in 'driver' genes; this would be preferable as it would allow all changes in driver genes to be identified rather than just the specific loci as examined with 'mpileup'. It was not possible to perform this relatively simple bioinformatic process as, due to the size of the dataset produced by the variant callers, the computing capacity required was in excess of that available at the WTSI, although as described above the variant caller data was manually filtered to detect MMR related mutations.

A superiority of using 'mpileup' over variant caller data is that the data returned by 'mpileup' comprises of the raw base counts at a desired locus rather than a positive or negative call such as produced by 'CaVEMan' and 'PINDEL'. Within the context of identifying shared variants between samples this is useful, as a mutated 'driver' loci in

one sample might, by chance fall into a region of low coverage and will be dismissed by an automated caller but manual interrogation using 'mpileup' will allow identification of changes in regions of poor coverage if they are supported by the same change with more supporting data in other samples.

The ubiquitous 'driver' mutations identified by 'mpileup' did not differentiate between the dMMR and pMMR cases; GD1 (pMMR) and GD3 (dMMR) showed mutations in *BRAF* in all samples and GD6 (pMMR) and 4 (dMMR) both showed mutations in *RAS* genes (*NRAS* and *KRAS* respectively). These widespread changes are felt to be reliable as they are present at high VAF and in all or a large majority of samples in each case; they are also almost identical findings to those of the pyrosequencing assay performed in section 4, the similarities will be further discussed in the following section. The pattern of mutations is slightly discordant with the large sequencing series in the literature as *BRAF* mutations tend to be more common within the dMMR/hypermutator tumours, whereas *KRAS* mutations tend to be over represented in the non-hypermutator cases [8,11,50,361], however the small number of cases within this study are likely to represent slight outliers to the overall pattern.

A second point for discussion is that, whilst mutations 'private' to a subgroup of lesions were identified in the pMMR tumours, the data supporting these changes was poor (either due to poor sequencing coverage or low VAF), whereas both dMMR cases showed heterogeneity supported by a greater proportion of variant reads in regions of better sequencing depth; in fact adopting a threshold of VAF greater than 10% and in areas of sequencing depth greater than 30x completely excludes all potential 'private' mutations from the pMMR cases. It is acknowledged that similar work within the literature has performed deep or ultra-deep sequencing to confirm the presence of 'private' mutations [83,97,105,288,362], however the facility to perform this additional sequencing was not available for this work and therefore VAF and sequencing depth thresholds were adopted for the detection of mutations to reflect the limitations of the sequencing modality and similar work using potentially degraded DNA [97]. The mutations satisfying the VAF and depth threshold in the dMMR cases included several changes which may potentially affect targeted therapy. One of the deposits in GD3 showed a 'driver' mutation in *PIK3CA* as did 14 of the samples from GD4 (although many of the majority of these calls are supported by low VAFs); *PTEN* mutation was also isolated in several samples from GD4. Both of these lesions have been related to resistance to EGFR therapy in meta-analytical work [363] although large trials have included an insufficient number of samples to produce

statistically significant results. *PIK3CA* and *PTEN* mutations have not been found to have prognostic significance in the absence of EFGR therapy and therefore may have remained 'private' to a subgroup of samples, as they may not confer an evolutionary advantage over surrounding WT tumour cells [364] without the selective pressure of targeted treatment.

Within a group of samples from GD4 a mutation in *FBXW7* was identified, mutations in the same gene were also identified in a group of samples from GD8. GD8 was the fifth case sequenced at the WTSI, which underwent target capture library preparation and NGS to a depth of 200x and comprised of FFPE material from both a resected primary tumour and recurrent disease sampled at autopsy. This case also showed *KRAS*, *APC* and *TP53* mutations in almost all samples sequenced, the samples which were WT for one of these three variants (no samples were 'double' WT), included two liver metastases and two local lymph nodes resected with the primary tumour. The two local lymph nodes showed low VAFs for other mutations, therefore it is possible that a mutation was missed due to low tumour content. The liver metastases appeared to display reasonable tumour content (via high VAFs in other mutations) and therefore these samples are likely to be genuinely WT which may have arisen due to loss of heterozygosity at the loci of the relevant gene. It is of note that a *KRAS* codon 12 mutation was identified in sample A20, this sample was wt for this mutation according to the pyrosequencing described in section 4. it is probable therefore that the negative result from pyrosequencing for this sample is reflective of the inferior sensitivity of this sequencing technique.

FBXW7 is a tumour suppressor gene, which is implicated in multiple tumour types [29] with a role in tumour initiation and growth; in the context of CRC, *FBXW7* is important to regulation of the WNT pathway [65], a pathway felt to be important to both dMMR and pMMR CRC. The presence of loss of function mutations in this gene have been associated with poor outcome as compared with *FBXW7* wt tumours [365] but the mutation status of this gene is not yet known to be therapeutically significant. It is reasonable to state therefore that *FBXW7* mutation provides a biological advantage to tumour cells and if a 'selective-sweep' type evolution were occurring within the tumour population of GD4 and 8 it would be expected that the *FBXW7* mutant cells would be more widespread. It is possible that the *FBXW7* mutation has been missed in several samples from GD4, due to the relatively low depth of sequencing, but it is highly unlikely that such changes were missed in GD8 due to the higher depth, more focused sequencing.

It is also striking that the *FBXW7* mutation was present only in the resected primary tumour and regional lymph nodes and not in the recurrent disease sampled at autopsy. This suggests that either the development of *FBXW7* mutation was either an event occurring after the initial seeding events leading to the development of disseminated disease but before the resection of the primary tumour or the region of *FBXW7* was lost in the precursor clone to the recurrent disease; the former situation would appear more probable as *FBXW7* mutation appears to be biologically advantageous to tumour cells and would likely be retained in metastatic disease. Therefore, the distribution of this *FBXW7* mutation provides evidence that nodal metastases occurred after dissemination of disease to distant sites. This divergence between the resected and recurrent disease was also observed within the CNA analysis in section 5.

This observation that 'late' driver events remain subclonal within these cases (especially in the instance of the *FBXW7* mutation) is further evidence supporting the theory that CRC are 'born to be bad' [323,325] with tumours developing the biological repertoire required to disseminate early in their evolution with later events playing a minor role in the natural evolution of disease; as previously discussed this theory does not incorporate the effect of targeted therapy, therefore these subclonal driver events may become important to therapeutic resistance. Further study of tumour dynamics in treated tumours is required to explore and potential impact of these subclones.

In summary, within this section WGS was performed on four cases of advanced CRC which included both pMMR and dMMR CRC, a fifth case underwent higher depth NGS on a panel of genes isolated by target capture library preparation. Initial QC steps excluded a significant minority of samples from the first two cases however all samples from the other three cases were adequate for the relevant sequencing technique. Clustering analysis of the data demonstrated an anatomical pattern to the mutational diversity within the pMMR cases which was not present in the dMMR case suitable for this approach. Analysis of 'driver' loci demonstrated variation across all four cases sequenced with WGS, however only the variation present within the dMMR cases was supported by reasonable VAFs and sequencing depth at the relevant loci. The genes showing evidence of heterogeneity in the dMMR cases included therapeutically significant genes suggesting that therapeutic resistance may arise via a mutational mechanism in these two cases. Finally, the fifth case (GD8) sequenced at high depth showed a homogenous population of distant metastases which appeared to arise before the emergence of local nodal metastases.

7 Genomic heterogeneity in locally advanced colorectal cancer

7.1 Introduction

As demonstrated in section 4, it is possible to demonstrate intertumoural genomic heterogeneity within cases of disseminated colorectal cancer by phylogenetic analysis of CNA data generated by NGS. The degree to which each case demonstrated CNA correlated with mismatch repair status, those which were dMMR showed minimal copy number change. The cases which showed evidence of CIN (resulting in CNA), showed a range of heterogeneity and, via the identification of shared breakpoints between deposits, we inferred a phylogenetic tree for each case comprising of 'truncal' events present in all samples and 'branch' events exclusive to a subgroup of lesions which represent the emergence of new tumour cell clones. Of the CIN CRC within the cohort, one case (GD6) showed minimal heterogeneity, with almost all CNAs identifiable in all deposits from the individual, whereas the remainder showed at least a handful of changes exclusive to a subgroup of lesions. This latter group of cases showed considerable divergence between primary tumours and their metastases, in addition to the grouping of tumour cell clones by anatomical location. From this data it is unclear at which point the cases developed heterogeneity in copy number, i.e. did it emerge within the primary tumour or as a result of time and possibly factors present within the microenvironment of the metastasis with clones recirculating to the primary tumour.

As previously described (in the introductory section) the presence of intra and intertumoural heterogeneity is established in CRC and other carcinoma types, however there has been little examination of the heterogeneity associated with different modes of metastasis identified within resected primary CRC. As described in section 2 and in the RCPath minimum dataset for the reporting of colorectal cancer [196], the established microscopic prognostic features include vascular invasion, perforation of the tumour onto the serosal surface of the bowel and involvement of regional lymph nodes; these features correspond to the main routes of metastasis for CRC, haematological, transcoelemic and lymphatic.

Traditionally only venous invasion present outwith the muscular wall of the bowel (extramural venous invasion (EMVI)) was thought to be of clinical relevance [366], however more recent data (including meta-analytical data) now also suggests intramural

lymphovascular invasion (LVI) is also prognostically significant when correctly identified [245,261].

The presence of regional nodal metastases in CRC and other carcinoma are amongst the most established prognostic factors in cancer; prognosis is also linked to the number of positive lymph nodes identified on histological examination of a resected tumour [367]. Local nodal metastases are the most easily sampled examples of local metastatic spread and as such data has been accumulated in comparing CNA profiles of primary CRC and limited numbers of nodal metastases has been accumulated [83,92,94,291,303,312,368,369]. Some of this evidence suggests that haematogenous and lymphatic metastases [369] represent different metastatic processes, however the limited sampling within these studies and the small number of samples sequenced does not rule out the possibility that the changes observed are incidental, neutral evolutionary events. A recent (previously cited) study of hyper-mutational regions within CRC also identified divergence of hepatic (presumably haematogenous) and lymph node metastases [85].

Extension of the tumour through the bowel wall, visceral peritoneum and to the serosal surface carries a poor prognosis as this may result in inoperable (stage M1c) disease [211], as such it is a staging criterion for CRC. As peritoneal metastases usually represent inoperable, and therefore unsampled disease, the current literature contains limited work examining heterogeneity between primary and peritoneal deposits in CRC. Early published work examining small number of peritoneal lesions suggested a characteristic CNA in the short arms of chromosomes 5 and 12 [303], however further meta-analytic work (representing a larger number of lesions) did not substantiate this observation [92] and concluded that there are not CNAs characteristic of peritoneal disease identifiable between individuals. The data presented in section 3 did demonstrate relative homogeneity in one case of disseminated peritoneal disease, suggesting that either a specific tumour clone either was biologically well suited to the micro-environment of the peritoneum and/or only a small portion of the tumour cell population had access to the peritoneum.

Correlation of genomic heterogeneity and the anatomical location of tumour in advanced CRC is worthy of exploration, as if the ability to invade vasculature, the peritoneum or spread to lymph nodes were restricted to specific subclones within a tumour or if clonal evolution were demonstrable with invasion into one of these anatomical regions, it would suggest ongoing tumour evolution in established CRC is an important process in the

progression of disease (as has been postulated in a recent papers examining rectal cancer [83,288] and prostate cancer [87]). Conversely if heterogeneity is not identifiable and multiple regions of the same tumour are capable of vascular, peritoneal or nodal spread, it would suggest that any evolution identified is incidental (or 'neutral') and CRC acquire metastatic capability early in their development [323] and metastases may arise simply due to access to vascular channels or the peritoneum granted by increasing tumour size and invasion of surrounding structures.

Therefore, in this section we explored the patterns of CNA in CIN CRC in relation to tumour location. DNA from intravascular, intraperitoneal and intranodal tumour was extracted and sequenced alongside that from the main tumour mass in a cohort of tumours with multiple poor prognostic features. The sequencing data from each region was analysed for CNA which allowed was used to explore the patterns of CNA alteration in relation to anatomical location of tumour.

7.2 Aims

1) Identify the presence of CNA within a cohort of resected CIN CRC

2) Identify the emergence of genomic divergence or ‘clonal evolution’ within these lesions

3) Correlate the anatomical position of tumour with the clonal evolution of the lesion and explore if:
    a. Clonal divergence coincides with direct vascular or peritoneal invasion
    b. Intraperitoneal and intravascular disease resembles the adjacent tumour
    c. Multiple foci of vascular invasion within the same tumour share similar CNAs
    d. Local nodal disease is restricted to specific subclones of tumour

7.3 Materials and methods

7.3.1 Case selection

The cases selected for analysis were 10 colorectal cancers resected pre-2006 (under ethics 08-H0903-62) which were identified being stage pT4b N2 (according to TNM5 [211]) along with exhibiting extramural vascular invasion. These cases were identified form a database held as part of the multidisciplinary team meeting record. As the material available for analysis was resected and reported prior to the recognition of IMVI as a prognostic indicator, tumours with IMVI only were not identifiable for selection.

7.3.2 Histological staining

All tumour blocks for each case were retrieved from the LTH archive, sectioned at 5μm and stained with haematoxylin and eosin (H and E).

The presence of conventional-type colorectal adenocarcinoma was thereby confirmed along with the presence of nodal metastases, serosal perforation by tumour and extramural vascular invasion. H and E staining was performed as per the method described in section 2. As IMVI did not form part of the standard reporting protocol at the time the cases were originally reported and dissection of tumour from small intramural vessels was unlikely to yield sufficient DNA for analysis, IMVI was not recorded.

The presence of extramural venous invasion was defined as tumour present within an extramural endothelium-lined space that is either surrounded by a rim of muscle or contains red blood cells as per the RCPath reporting guidelines [196]. To ease the identification of EMVI within this cohort Miller's elastin stain as performed to highlight the elastic lamina present within the walls of medium to large vessels within the mesenteric fat. This stain was also performed to confirm perforation of the peritoneal elastin in blocks suspicious for serosal perforation.

Elastin staining was performed as per the following protocol:

- Sections de-waxed

- 5 mins in potassium permanganate 5 mins
- Rinsed in distilled water
- Decolourised with oxalic acid 1 mins

- Rinsed in distilled water
- Rinsed in 95% alcohol
- 1.5 hours in Miller's stain
- Washed in 95% alcohol for 1 min
- Rinsed in distilled water
- 2 mins in haematoxylin

Several serpiginous extensions of tumour were identified on H and E staining, which were equivocal for elastin staining. These are denoted as ?vascular invasion in the results section.

For the first 5 cases, foci of vascular invasion and tumour extending beyond the peritoneal elastin layer to the surface of the bowel wall were dissected using laser capture microdissection (LCMD), in addition to macrodissection of nodal metastases. The LCMD protocol is as described in section 7.3.4.

The second 5 cases of the cohort underwent macrodissection of whole sections of tumour and their nodal metastases. The macrodissection protocol was as described in section 2.

7.3.3 MMR IHC

As described in section 3 only CIN CRC appear to evolve via CNA, therefore the MMR status of each case was confirmed to exclude any dMMR/chromosomally unstable cases. To confirm the tumours mismatch repair status, IHC was performed for MLH1, MSH2, MSH6 and PMS2 on two blocks from each case, one from the primary tumour and one from a metastasis.

The protocol used was as described in section 3.

7.3.4 Laser Capture Microdissection (LCMD) Protocol

For each section with EMVI or serosal perforation within the first 5 cases, 10 7 µm sections were taken on Applied Biosystems Arcturus Polyethylene Naphthalate (PEN) Membrane Glass Slides (Thermo Fisher Scientific, Foster City, California, United States). These slides bear a plastic membrane onto which the tissue sections are placed, this allows sections to be dissected using a laser with microscopic accuracy. Once cut the membrane and overlying tissue can be lifted manually or catapulted into a collection tube using an in-built function of the PALM MicroBeam (P.A.L.M. Microlaser Technologies, Carl Zeiss Microscopy GmbH, Jena, Germany) LCMD platform.

To allow the areas for LCMD to be identified the tissue was counterstained using instant haematoxylin according to the following steps:

- Sections dewaxed

- 2 mins in instant haematoxylin

- Rinsed in distilled water

- Placed in 70% ethanol

Slides are then mounted on the stage of the PALM Microbeam (version 1104) LCMD platform and viewed via the PALM Robo software. The microscope platform was moved using the PALM Robomover (version 0804).

Areas of intravascular (IV) and intraperitoneal (IP) tumour were marked out and dissected with the laser settings as follows:

Cut: Energy: 49% Focus: 63%

Slides were then transferred to a dissecting microscope and the dissected areas were transferred into Eppendorph tubes containing 180 µl of buffer ATL (contained within the Qiagen QiAMP DNA micro kit). The slides were then viewed again confirming the correct area has been removed.

Figure 48: Photomicrograph displaying a region of intravascular tumour (A) marked out in the PALM Robo software and (B) following dissection with the region of vascular invasion excised.

The remaining tumour was then dissected and placed in 180 µl buffer ATL.

7.3.5 DNA extraction

DNA was extracted from the dissected tissue using the Qiagen QiAMP DNA micro kit according to the manufacturer's protocol (Qiagen, Crawley, UK). The sections of PEN membrane are removed from the sample before transfer into the elution columns.

7.3.6 DNA Quantification

The concentration of nucleic acid within each sample was initially quantified using a Nanodrop-1000 spectrophotometer (Thermo Fischer Scientific, Loughborough, UK). This measurement was used to dilute the extracted samples for dsDNA quanitification using the Quant-iT dsDNA Assay Kit (Thermo Fischer Scientific, Loughborough, UK) and Fluroskan Ascent Microplate Fluorometer (Thermo Fischer Scientific, Loughborough, UK).

7.3.7 Copy number library preparation

The library preparation for CNV was performed using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, Massachusetts, USA).

Due to lower DNA yield from LCMD a minimum of 100 ng of dsDNA was submitted for library preparation. The protocol used in this section is identical to that described in section 5except for the DNA shearing and the PCR step of the library preparation, these are described as follows:

7.3.7.1 DNA shearing

100 ng of extracted DNA was sheared into 200 bp fragments using the Episonic multi-functional bioprocessor 1100 (Epigentek Group inc, Famingdale, United States). This instrument operates in the same way as the Covaris S220 Focused-ultrasonicator, i.e. by producing high frequency sound, which fragments the DNA within the sample. This instrument was chosen as it can be used with a standard 96 well plate (rather than shearing samples individually in Covaris micro-tubes as described previously), it was therefore quicker and more affordable than the Covaris sonicator. The settings used for the sonication procedure were:

Sample volume:	30 µl	Amplitude:	18-19
Target power output:	170-190 watts	Process time:	15 secs
Pulse-ON Time:	15 secs	Pulse-OFF Time:	30 secs
Duty Cycles:	60		

25.5 µl of Buffer EB was used to make the samples up to 55.5 µl, the samples were then run using the protocol described in section 5.

7.3.7.2 Library preparation PCR

The NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, Massachusetts, USA) used for this section includes the NEBnext Q5 Hotstart High Fidelity 2x PCR Master Mix (rather than the NEBNext High Fidelity 2X PCR Master Mix included in older iterations of the kit). This enzyme produces higher yields of PCR product (allowing lower starting DNA concentrations) with greater accuracy than the previous reagent [370]. The modified PCR amplification step was as follows:

Reagent	ng/µl For 1 well
NEBNext Q5 2X PCR Master Mix	25
Universal PCR Primer	5
Total volume to add per sample	30
Volume of unique index primer per well	2
Volume of dH20 per well	3
Total volume of sample	*50*

The thermal cycler conditions were:

Step	Temp (°C)	Time	Cycles
Initial denaturation	98	30 sec	1
Denaturation	98	10 sec	6-15
Annealing	65	30 sec	
Extension	72	30 sec	
Final extension	72	5 minutes	1
Hold	4	∞	

All other steps in library preparation were identical and 30 ng of each sample was submitted for sequencing.

7.3.8 CNA sequencing

DNA sequencing was performed using the Illumina HiSeq 3000 platform producing paired 150 base pair reads; the increased fragment length of produced by this newer platform (compared to the Illumina HiSeq2500 used in section 5) improves sequencing quality by increasing the quality of alignment to the reference genome.

The post sequencing bioinformatic pipeline used for this section was identical to the described in section 5.

7.3.9 Mapping of tumour heterogeneity

The methodology used to identify tumour evolution was that described in section 5.

7.4 Results

7.4.1 Histological characteristics of archived cases

All ten cases sampled when viewed on H and E staining were confirmed as moderate to poorly differentiated adenocarcinoma arising within the colon, confirming the stated diagnosis in the original histopathological report.

Within the five cases dissected using LCMD, 40 and 35 regions of intravascular and intranodal tumour were identified and sampled. Within SC1 and SC2 multifocal extramural vascular invasion was identified (as described in the materials and methods section), three further areas highly suspicious for EMVI and 13 and 10 involved lymph nodes were also sampled in SC1 and SC2 respectively. Unfortunately blocks containing intraperitoneal tumour were unsuitable for the first two cases. In the third, fourth and fifth cases which underwent LCMD more prominent peritoneal invasion was present within the blocks available, along with occasional foci of intranodal and intravascular tumour.

The whole sections of tumour extracted from SC6-10 contained 8 regions of extramural vascular invasion and 17 involved lymph nodes. The foci of EMVI sampled in these cases were all either from blocks taken from the primary tumour which comprised only of intravascular tumour or regions sampled and described as ‘involved lymph nodes’ in the clinical histopathology reports, which appeared to be vascular involvement on elastic staining. All five cases comprised of a similar number of foci of EMVI, nodal involvement and peritoneal invasion. The sections arising from the primary tumour were extracted en-bloc and areas of pT4 were not dissected separately unless they were a macroscopically distinct lesion.

As shown in Table 32, all samples extracted from the macrodissected cases (SC6-10) yielded a sufficient quantity of DNA for library preparation, whereas at least one sample from each laser-microdissected case failed to produce 100 ng of DNA. The inadequate samples all originated from foci of vascular invasion, except for one small nodal metastasis in SC2.

Table 32: Table displaying the number of samples successfully extracted/areas dissected from foci extramural venous invasion (EMVI), probable foci of vascular invasion (?VI), nodal tumour deposits and foci of peritoneal invasion in the laser dissected and macro dissected surgically resected tumours examined.

	EMVI	?VI	Node	pT4 tumour	Dissection method
SC1	5/7	2/2	13/13	0/0	Laser
SC2	3/4	0/1	9/10	0/0	Laser
SC3	6/8	1/1	2/2	3/3	Laser
SC4	6/7	0/0	5/5	2/2	Laser
SC5	10/14	1/1	5/5	2/2	Laser
SC6	2/2	0/0	3/3	3/3	Macro
SC7	4/4	0/0	2/2	2/2	Macro
SC8	0/0	2/2	5/5	2/2	Macro
SC9	2/2	0/0	2/2	2/2	Macro
SC10	1/1	1/1	5/5	1/1	Macro

7.4.2 MMR IHC

All ten cases examined were mismatched repair proficient, according to IHC for MLH1, MSH2, MSH6 and PMS2, the results for each case are listed in Table 33. An example of the IHC staining from SC1 is shown in figure 49.

Table 33: Table displaying the results of the IHC staining for the MMR antibodies for each case examined in this section.

	MLH1	MSH2	MSH6	PMS2
SC1	+	+	+	+
SC2	+	+	+	+
SC3	+	+	+	+
SC4	+	+	+	+
SC5	+	+	+	+
SC6	+	+	+	+
SC7	+	+	+	+
SC8	+	+	+	+
SC9	+	+	+	+
SC10	+	+	+	+

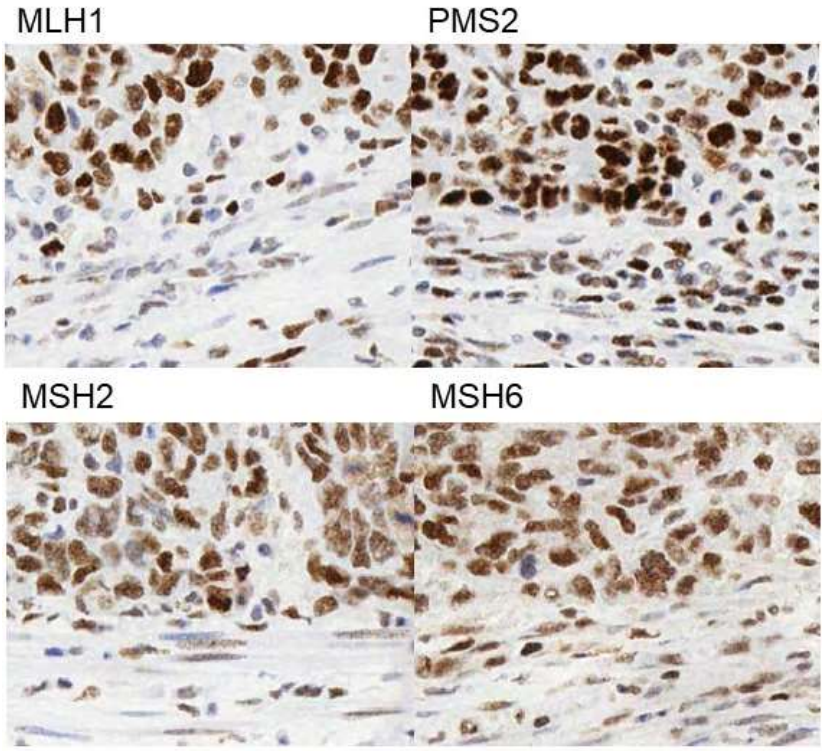


Figure 49: Photomicrographs showing immunohistochemistry from SC1 demonstrating retention of MSH2/6 and PMS2/MLH1 staining

7.4.3 CNA analysis

All samples, which produced more than 100 ng of DNA, were successfully sequenced and analysed according the method described in section 5. As shown in Figure 51 - Figure 60, similar patterns of CNA were observed in this cohort to those observed in the CIN CRC 'Gift' cases described in section 5. In all ten cases a group of CNAs were observed as being ubiquitous, 'truncal' changes with a variable number of subsequent CNAs which were exclusive to a subgroup of lesions; in the majority of cases the truncal cluster of CNAs contained a larger number of CNAs than any subsequent step in the phylogenetic tree for that case. This latter group of CNAs were considered to represent the emergence of a new subclone cells within the tumour and a branch in the phylogenetic tree for the respective disease process. Once again, several samples were found to contain CNAs characteristic of two or more branches of the phylogenetic tree and were considered polyclonal.

Across the ten cases analysed the CNAs present were also similar to those described in section 5 and by the TCGA [16]. These CNAs are presented in the figures for each case and include losses in 1p, 4q, 5q, 8q, 15q, chromosomes 17, 18 and 20 and gains in 1q, 7p, 8q, 12q, 13q and 20.

Another characteristic shared across all ten cases examined, is the observation that regional metastases (either nodal metastases, vascular emboli or peritoneal seedlings) appear to have arisen from multiple different clones within the primary tumour; this finding will be described in greater detail in the context of the individual cases.

A final general finding is that regions of intravascular spread which were contiguous with the primary tumour appeared to be of the same clone as the adjacent tumour.

In each case analysed and described below the numbers prefixing each sample correspond to the FFPE block from which the samples were extracted. In several cases, numerous samples have originated from the same FFPE tissue block. Those samples which originate from a block containing primary tumour all originate from the same piece of tissue and are spatially close to one another i.e. the tissue block contains more than one tumour deposit separated by normal tissue. Those samples originating from blocks which did not contain primary tumour, are from blocks containing spatially distinct, macroscopically suspicious areas combined into one FFPE block by the reporting

pathologist at the time of clinical pathological examination. An example of these two scenarios is presented in Figure 50.

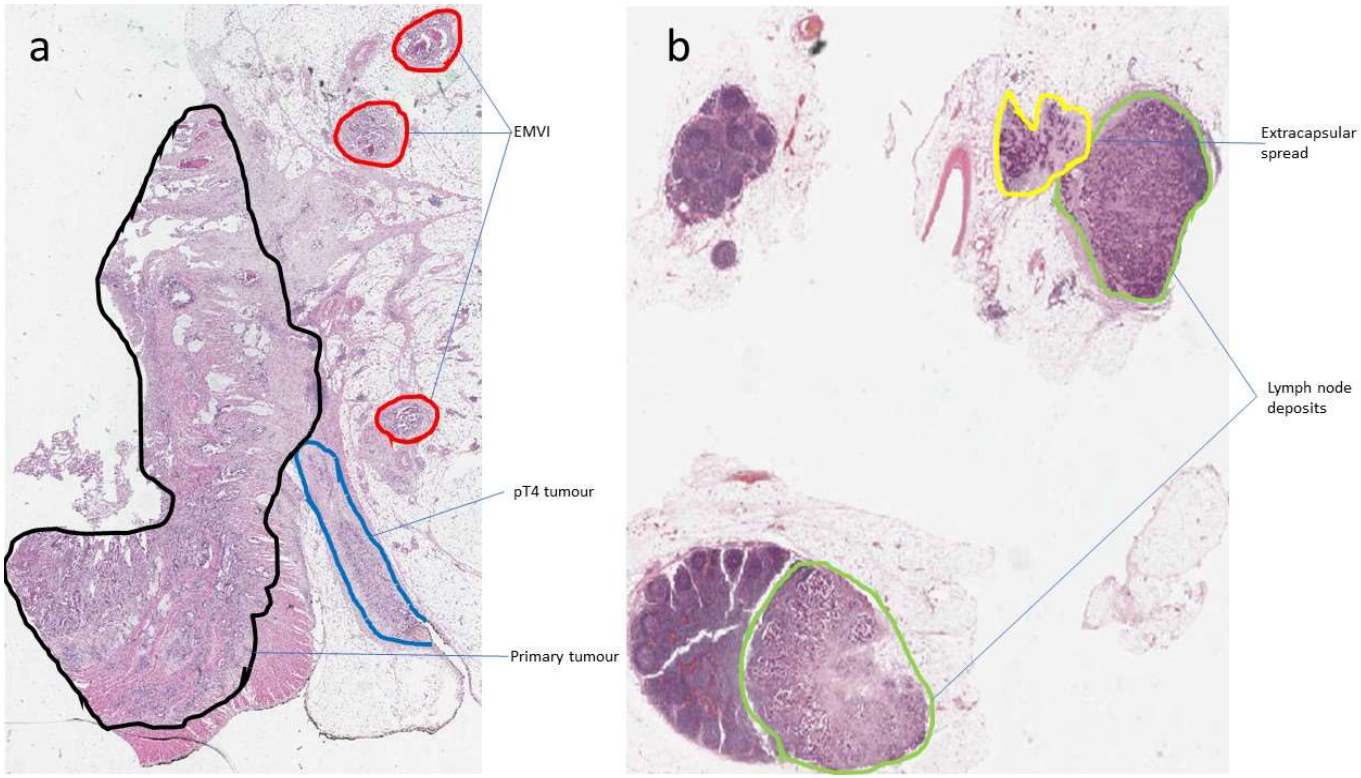


Figure 50:This figure demonstrates the regions dissected from surgically resected tumours in this section. a) shows a haematoxylin and eosin stained section from a FFPE block containing a single portion of tissue comprising primary tumour with a contiguous portion of intraperitoneal tumour ('pT4 tumour') and three foci of extramural vascular invasion ('EMVI'); this tissue block therefore contains tumour in distinct anatomical compartments which are closely spatially related. All the tissue blocks containing primary tumour were single pieces of tissue. b) shows a section of tissue from a FFPE block containing multiple tissue fragments which comprise two lymph nodes both partially replaced by metastatic carcinoma, as these two tumour deposits are within separate fragments the anatomical relationship between the two and with the primary tumour is unknown.

7.4.3.1 SC1

This case, as described above, was dissected using LCMD and comprised a large primary tumour with prominent vascular invasion, nodal metastases and peritoneal perforation. CNA analysis, shown in Figure 51, yielded a number of changes, the largest group, or cluster, was shared or 'truncal' and included aberrations in chromosomes 8, 13 and 18. Five subsequent clusters were identified, which were present in subgroups of lesions and occurred in a mutually exclusive fashion apart from six samples, which appeared to be polyclonal. The only CNA present within the 'branch' clusters also identified in the TCGA publication is a loss in 4q identified in cluster 2 and 5 (although the precise position of the breakpoint in 4q differs between the two clusters). The polyclonal samples were two samples taken from the primary tumour, one sample taken from a region of vascular invasion adjacent to the primary tumour and three lymph nodes.

Within this case there was association of the anatomical location and the phylogenetic clustering of each sample. As shown in the figure below those samples originating from a block containing primary tumour (which we know to be closely related anatomically as they have been sampled in continuity within the same FFPE tissue block) cluster together. There does not appear to be an association between the mode or type of spread (nodal, vascular or peritoneal) and the position of the sample within the phylogenetic tree.

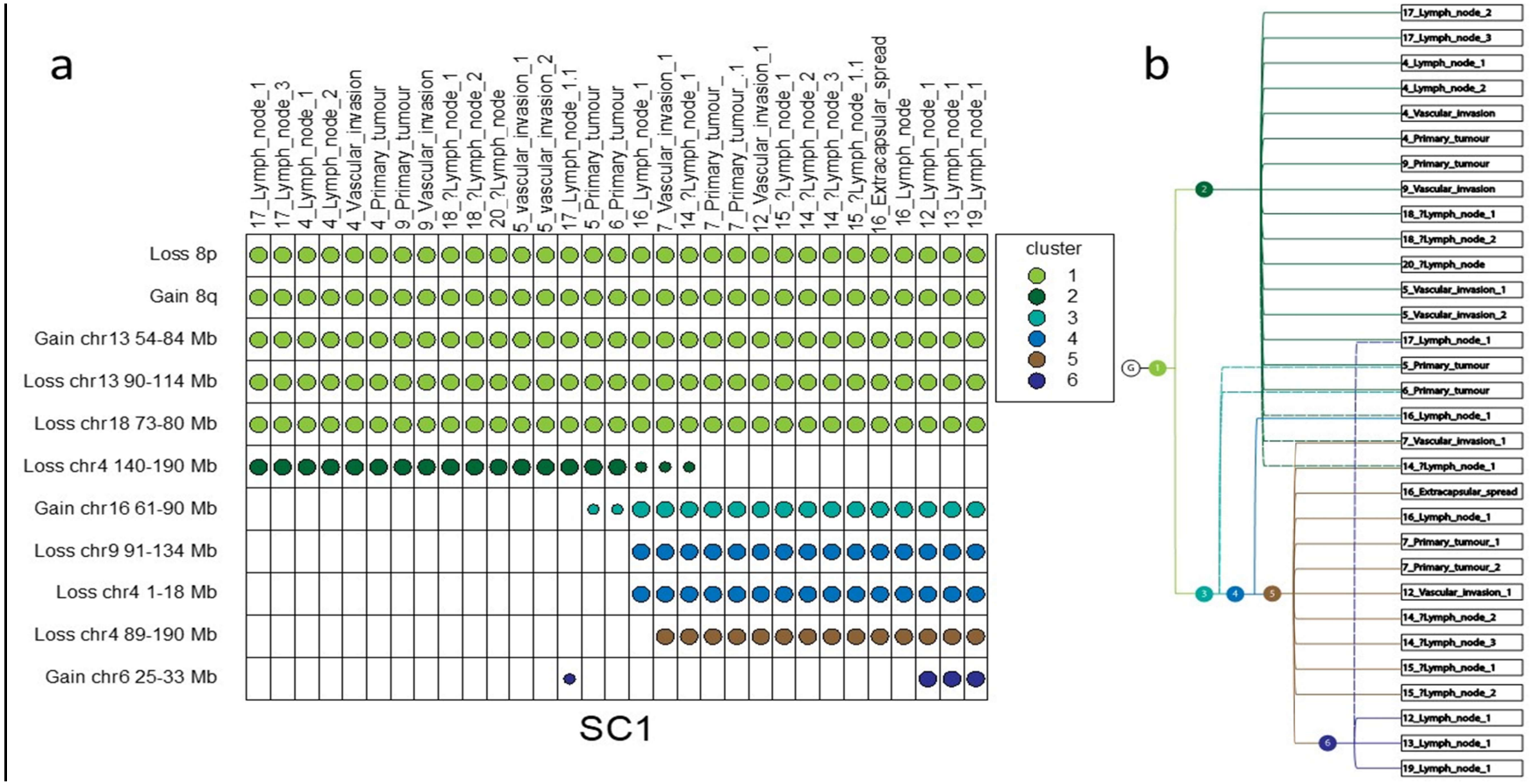


Figure 51: This figure displays the pattern of CNAs in SC1, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC1 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.2 SC2

This is the second LCMD case and comprises of samples from the primary tumour, vascular invasion and nodal metastases. Once again, as displayed in Figure 52, the largest cluster of CNAs was shared across all samples taken from this tumour, and included several CNAs recurrently identified in CRC such as gain of 1q, chromosome 7, 19q and chromosome 20, with several subsequent evolutionary steps, which occurred in a largely mutually exclusive fashion with occasional polyclonal deposits which included nodal metastases and nodal emboli. Within this case the samples do not appear to cluster by mode of metastasis nor by anatomical location, however the samples arising from metastatic sites are again distributed throughout the phylogenetic tree as described in SC1, in fact a vascular embolus from block 15 and a nodal metastasis from block 12 exhibited only the 'truncal' CNAs. The only CNA present within the list of 'branch' CNAs in this case also observed within the TCGA analysis of CRC was loss in 22q.

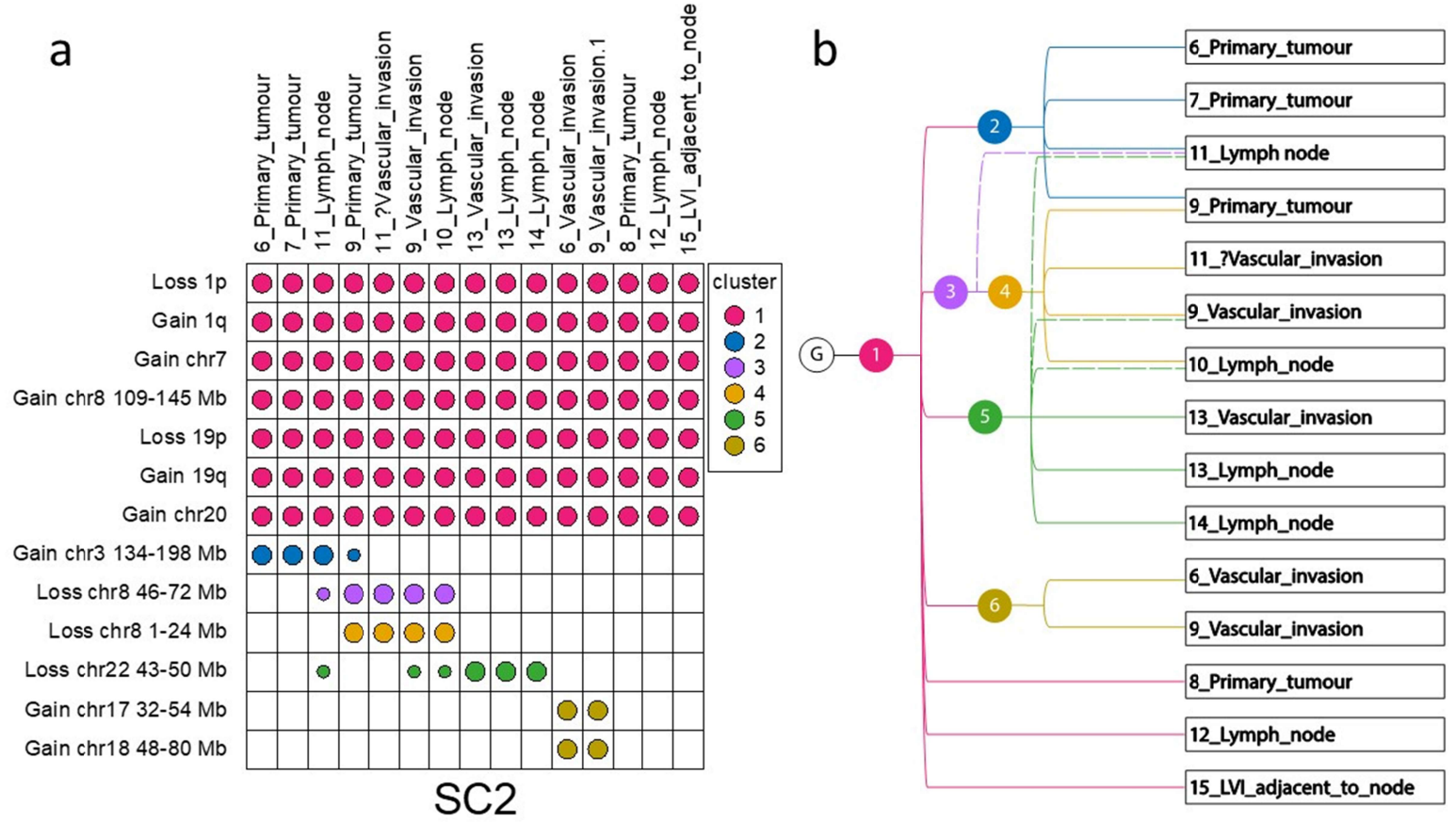


Figure 52: This figure displays the pattern of CNAs in SC2, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC2 on the yaxis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.3 SC3

This third LCMD case is subtly different from the initial two cases. As observed in the previous examples, there is a group of 6 'truncal' CNAs, which include several common CNAs observed in CRC (such as those detailed below in chromosome 8, 18 and 20), however almost half of the samples (those falling into cluster 4) share 8 'branch' CNAs. As samples of angio-invasive tumour are present in both major branches of the phylogenetic tree for this case (clusters 4 and 9), there appears to have been a significant change in CNA prolife which developed after the lesion has developed the capacity to invade blood vessels and bowel wall. Apart from this distinction the pattern of evolution with this case is again that of an iterative process with metastases emerging at each step. 5 polyclonal samples are identified, one from the primary tumour, one from a lymph node and 3 from foci of vascular invasion which were contiguous with the primary tumour (rather than vascular emboli anatomically distinct from the primary tumour). The similarity between those samples taken from the same anatomical location, as represented by those samples originating from the blocks containing primary tumour, is once again observed in this case. These findings are shown in Figure 53.

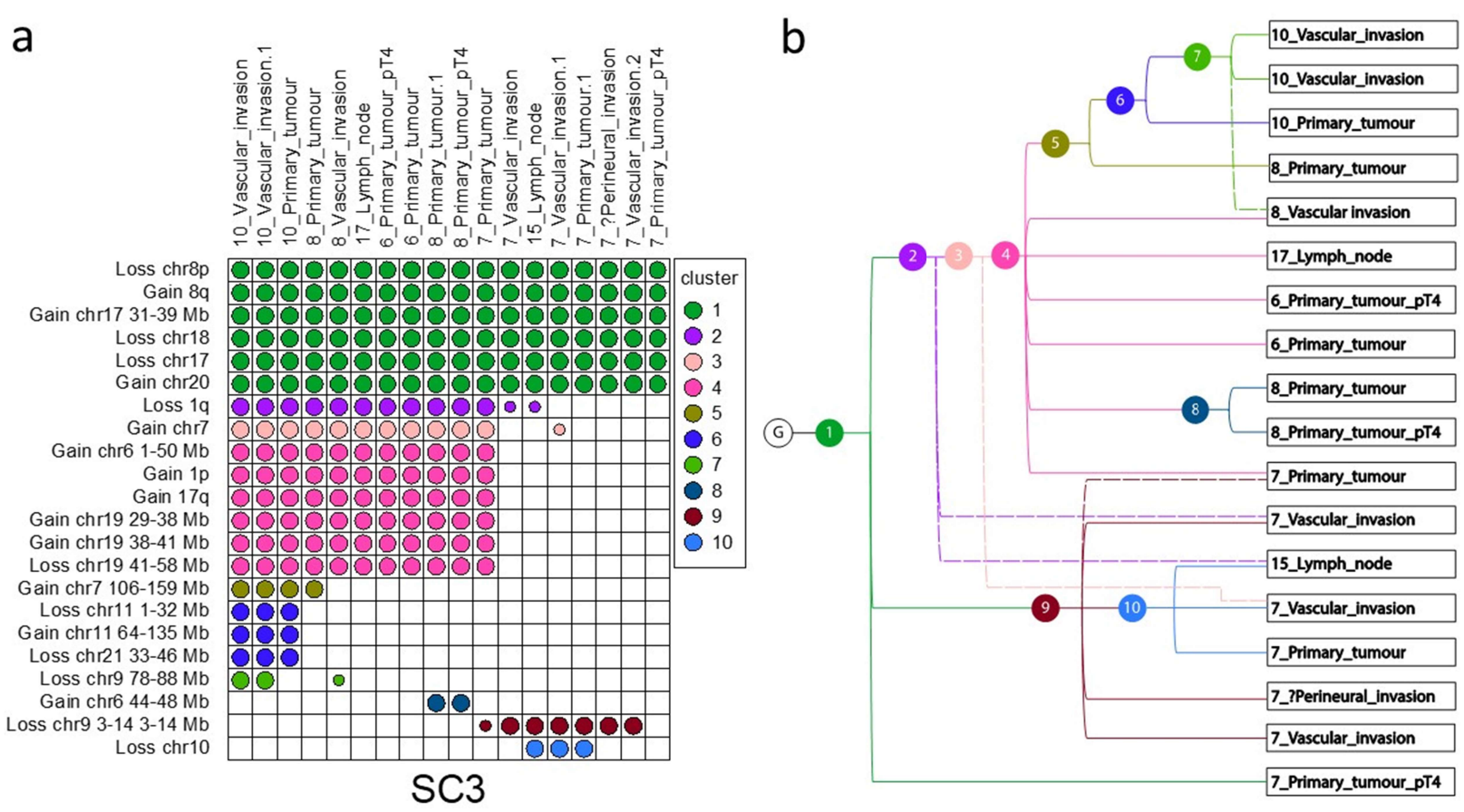


Figure 53: This figure displays the pattern of CNAs in SC3, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC3 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.4 SC4

The fourth LCMD case shows similar patterns of CNA to the SC1 and 2 (as shown in Figure 54), with a large group of 'truncal' CNAs similar to those described in the cited literature (gain of 1q, 7p, losses in 5q and 18q), with subsequent 'branch' CNAs and a broad correlation between those samples known to be spatially related and the occurrence of cluster defining CNAs. One exception to this correlation in the case is a vascular embolus originating from block 3 which has diverged from the other samples in this tissue block. It has however clustered with other samples originating from blocks containing primary tumour and adjacent intravascular tumour from block 5; as all blocks taken from the primary tumour will be relatively close anatomically, it is probable that these samples represent a clone present in the primary tumour in block 5 which is beginning to extend into block 3 via vascular spread. This intermingling of tumour cell clones within the primary tumour is again represented by the polyclonality identified in the block 3 and 12 which originate from the primary tumour. Additionally, once more a polyclonal metastasis within a lymph node was identified and metastases arose from different regions and from different stages of tumour evolution (as represented by the emergence of new clones).

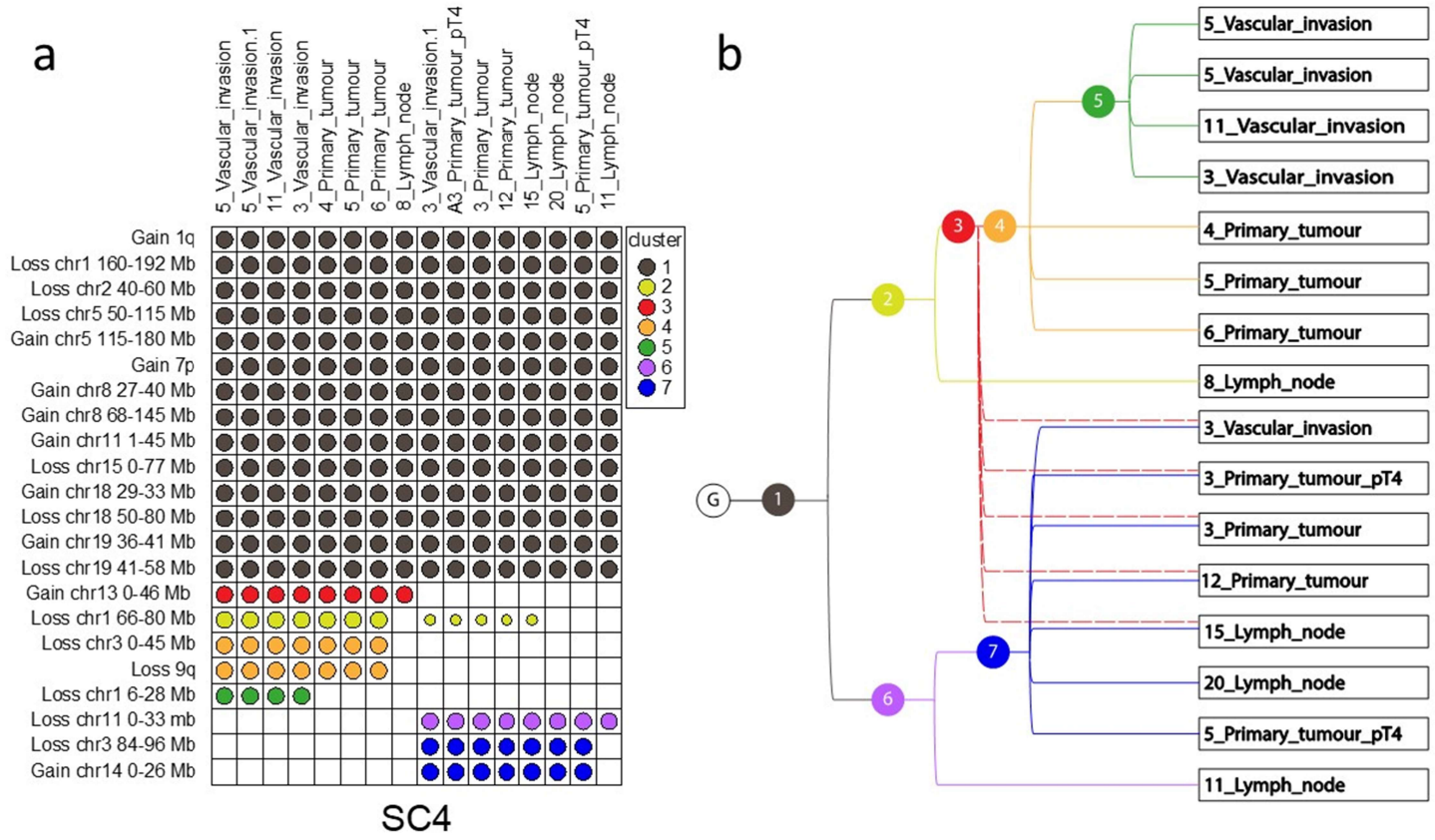


Figure 54: This figure displays the pattern of CNAs in SC4, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC4 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.5 SC5

This, final case which underwent LCMD, displays a large number of CNAs and a complex pattern of evolution as displayed in Figure 55. As was the case in the majority of other cases the truncal, ubiquitous cluster of CNAs is the largest identified in this case this cluster contains losses of 5q, 8p and chromosome 18 and gain of chromosome 13, and there are then multiple, divergent CNAs, the majority of which occurred in a markedly mutually exclusive fashion with only two polyclonal deposits. In this case, as observed in others, nodal metastases and samples arising from intravascular tumour (either probable emboli initially sampled as a probable involved lymph node, as in one of the samples from block 10 or vascular invasion in direct continuity with the primary tumour, in block 5) are present within multiple branches of the phylogenetic tree, with some metastases sharing only the truncal CNAs with samples taken from the primary tumour; once again the metastases appear to have arisen at multiple stages of tumour development. Within this case the relationship between samples in direct continuity with one another e.g. those taken from block 5 and those from the primary tumour (which all originate from the same tumour mass) fall into two branches of the phylogenetic tree which are markedly divergent, sharing only cluster 1 and having acquired multiple mutually exclusive CNAs; this divergence is particularly striking between two of the foci of VI in block 5 and the primary tumour sampled from the same tissue block, these divergent samples do, however cluster with the sample taken from the primary tumour from block 8 (which is extending through the peritoneal elastin) and may represent extension of the clone present in block 8 into block 5 via infiltration of the vasculature. Therefore, in this case we have identified considerable diversity of CNA, both within the primary and the between the metastases, this diversity suggests the occurrence of metastases at multiple stages of evolution by CNA and appears to correlate with the location of the samples.

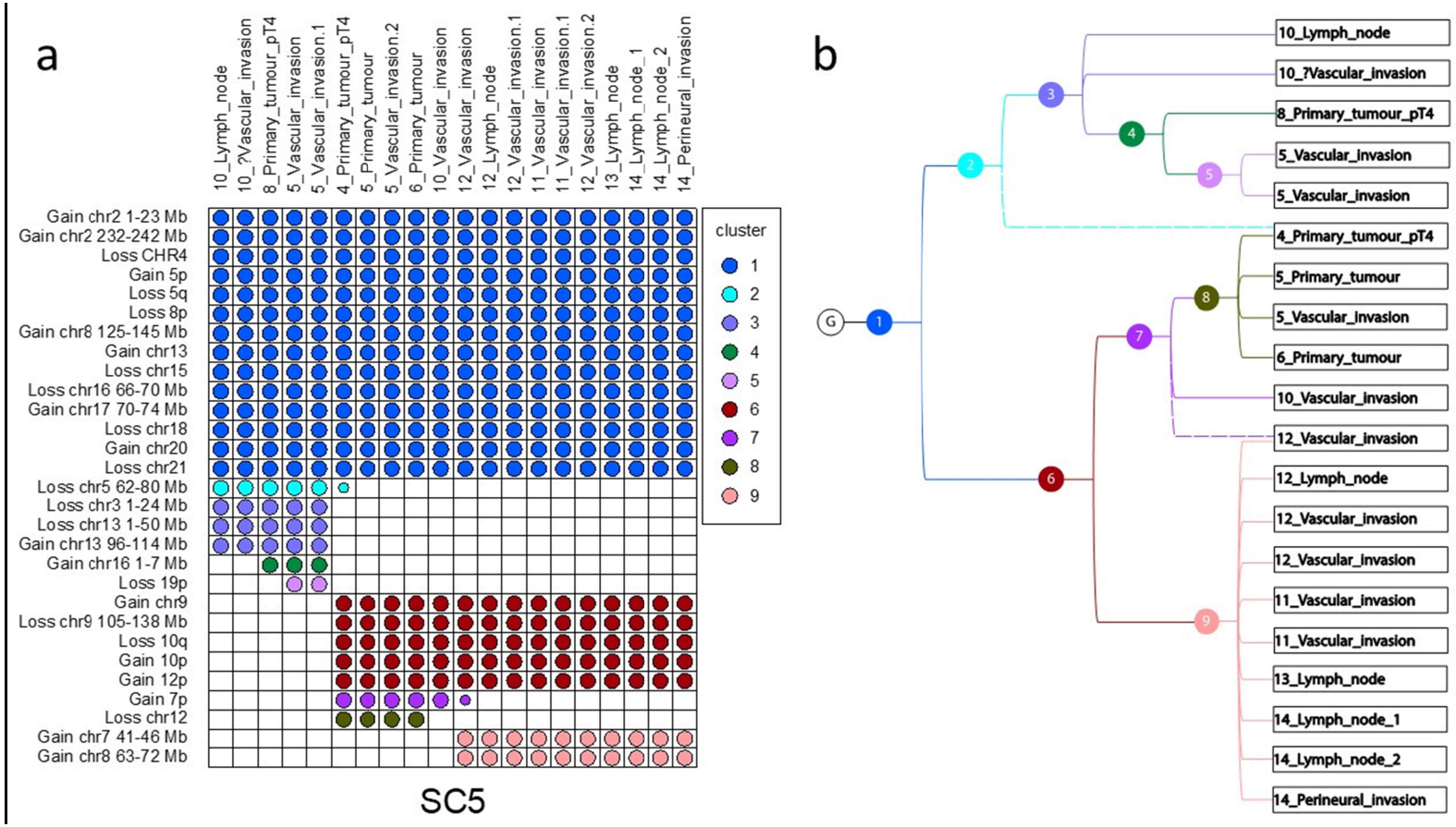


Figure 55: This figure displays the pattern of CNAs in SC5, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC5 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.6 SC6

This is the first within this cohort which did not undergo LCMD. As such any DNA from intravascular or intraperitoneal (pT4) tumour present within the primary tumour blocks (samples numbered 6, 7, 8, 10, 13 and 14) was macrodissected and extracted with the rest of the tumour from the section; the blocks which showed areas of EMVI and peritoneal invasion (6, 7 and 8) are highlighted as such in Figure 56. Despite the less exact method of dissection used in this and the subsequent 4 cases, similar patterns of clonal evolution are observed when compared to the LCMD cases. As previously demonstrated a relatively large number of shared, truncal CNAs are present in all samples with the emergence of branch CNAs taken to represent clonal divergence and evolution within the samples; metastases once again appear to emerge at different stages of disease evolution. The branch CNAs in this case were relatively numerous and the degree of mutual exclusivity of these changes were less marked than in some of the laser dissected cases; a degree of polyclonality was observed in the majority of blocks taken from the primary tumour.

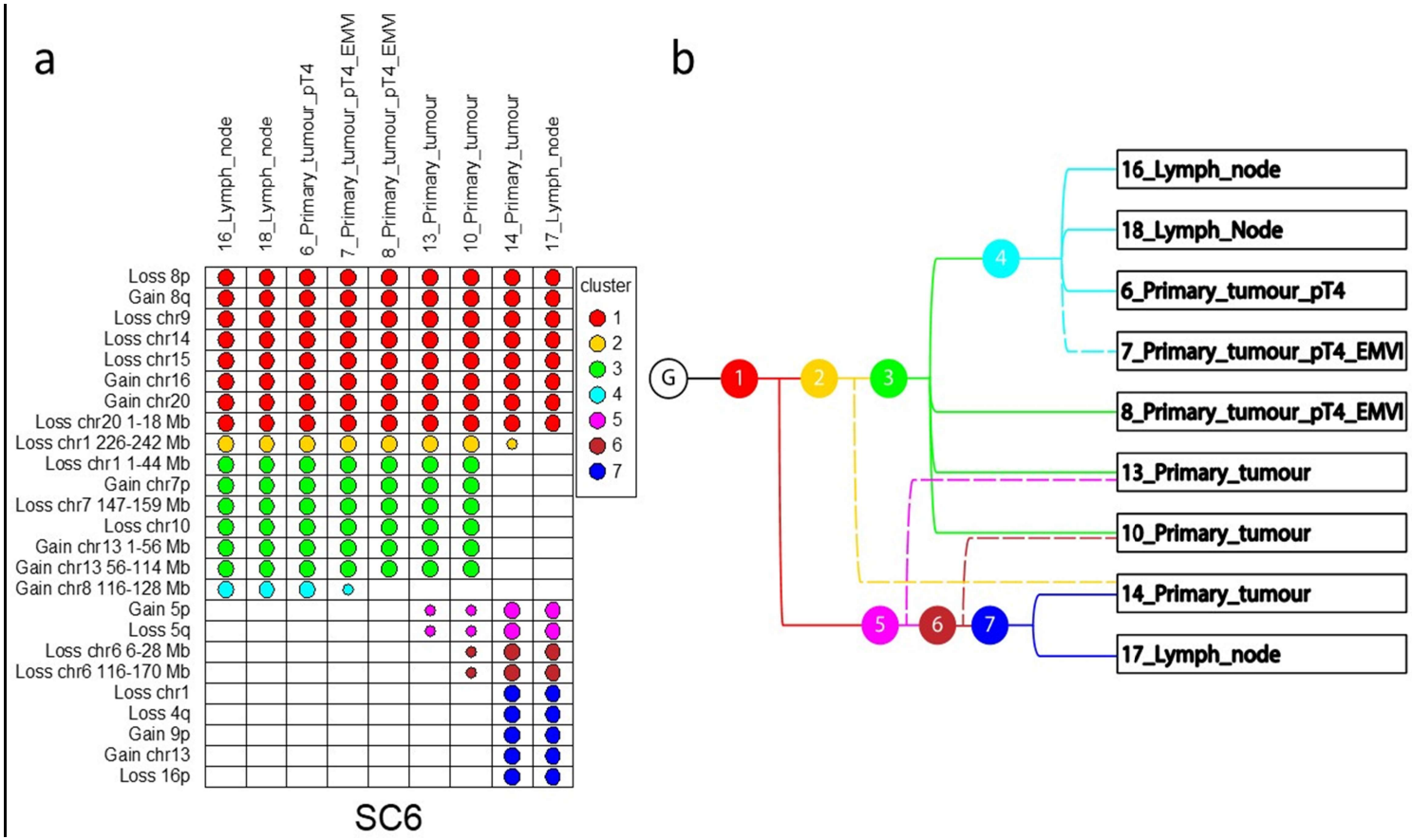


Figure 56: This figure displays the pattern of CNAs in SC6, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC6 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.7 SC7

This second, macrodissected case is unique (within the cohort) in displaying a small number of truncal CNAs (although significantly they include canonical CRC CNAs such as loss of 18q) with much more numerous CNAs within subgroups of lesions. Otherwise once again divergence within the primary tumour and metastases is present, another shared observation is metastases appear to arise at different points in the evolution of the tumour. The lymph node metastases sampled from blocks 11 and 12, share only the main truncal CNAs and appear to resemble different sections of the primary tumour rather than those taken from deposits arising from the same metastatic route. The 'branch' CNAs in this case were mutually exclusive to different subgroups of samples and therefore no polyclonal samples were identified. These findings are displayed in Figure 57

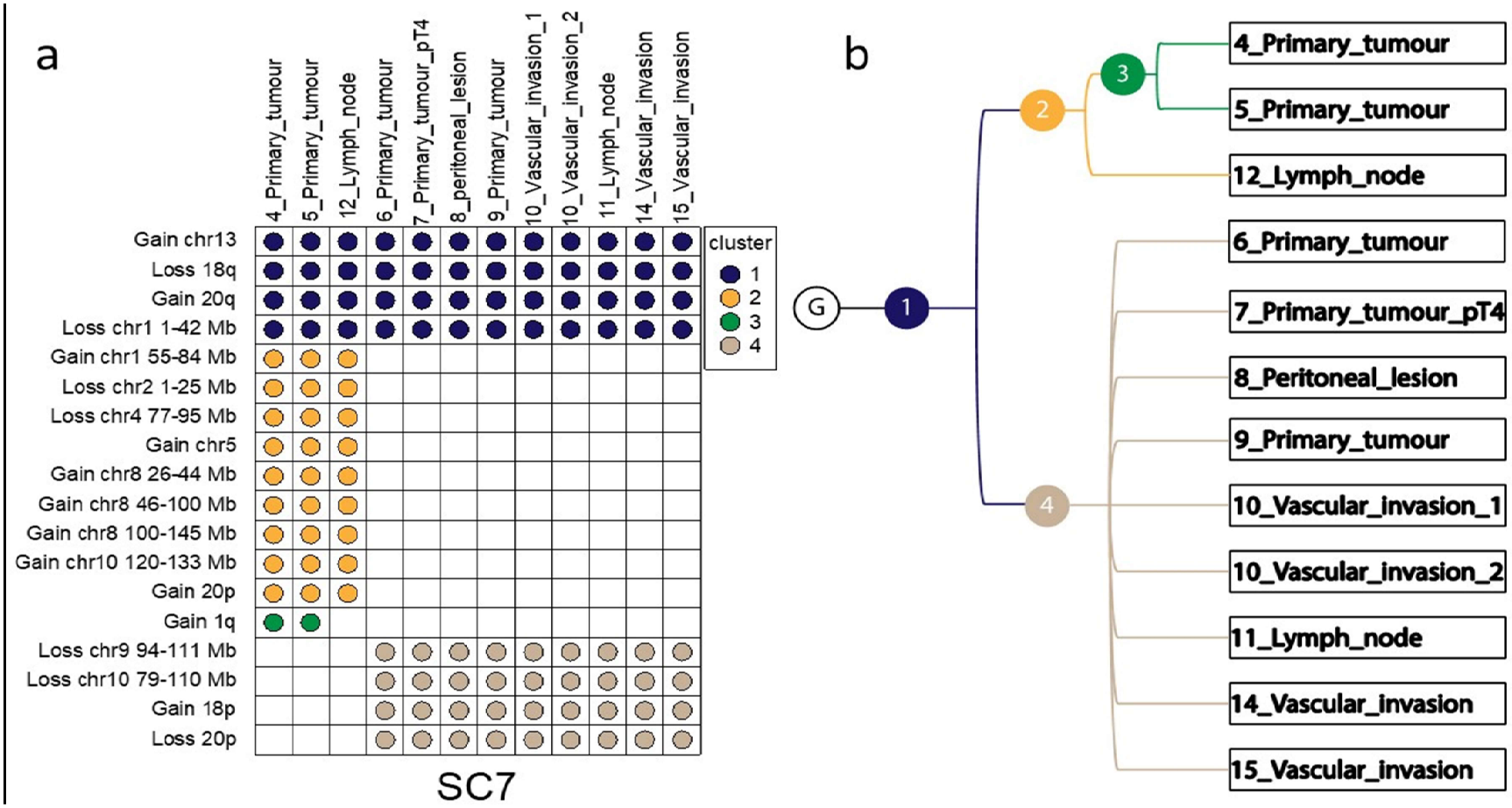


Figure 57: This figure displays the pattern of CNAs in SC7, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC7 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.8 SC8

This case (as shown in Figure 58) comprised a well sampled primary tumour and a large number of local metastatic deposits in addition to a tumour deposit which had perforated the peritoneal elastin (in block 7). As might be expected, with an abundance of sampled material, a relatively large amount of diversity was identified within the case, although few 'truncal' CNAs were identified in comparison with the majority of cases previously discussed. The four truncal CNAs do not include the loss of 18q as observed in many of the previous cases (and in CIN CRC as a whole) but all samples from this case did display loss of 1p, 4q and 5q (common events in CRC). Apart this subtle distinction this case reiterates the previously described branching pattern of evolution with metastases arising from different branches of the phylogenetic tree. This case harboured several polyclonal samples, which were mainly taken from the primary tumour, and the samples taken from the primary tumour were well represented throughout the phylogenetic tree for the case.

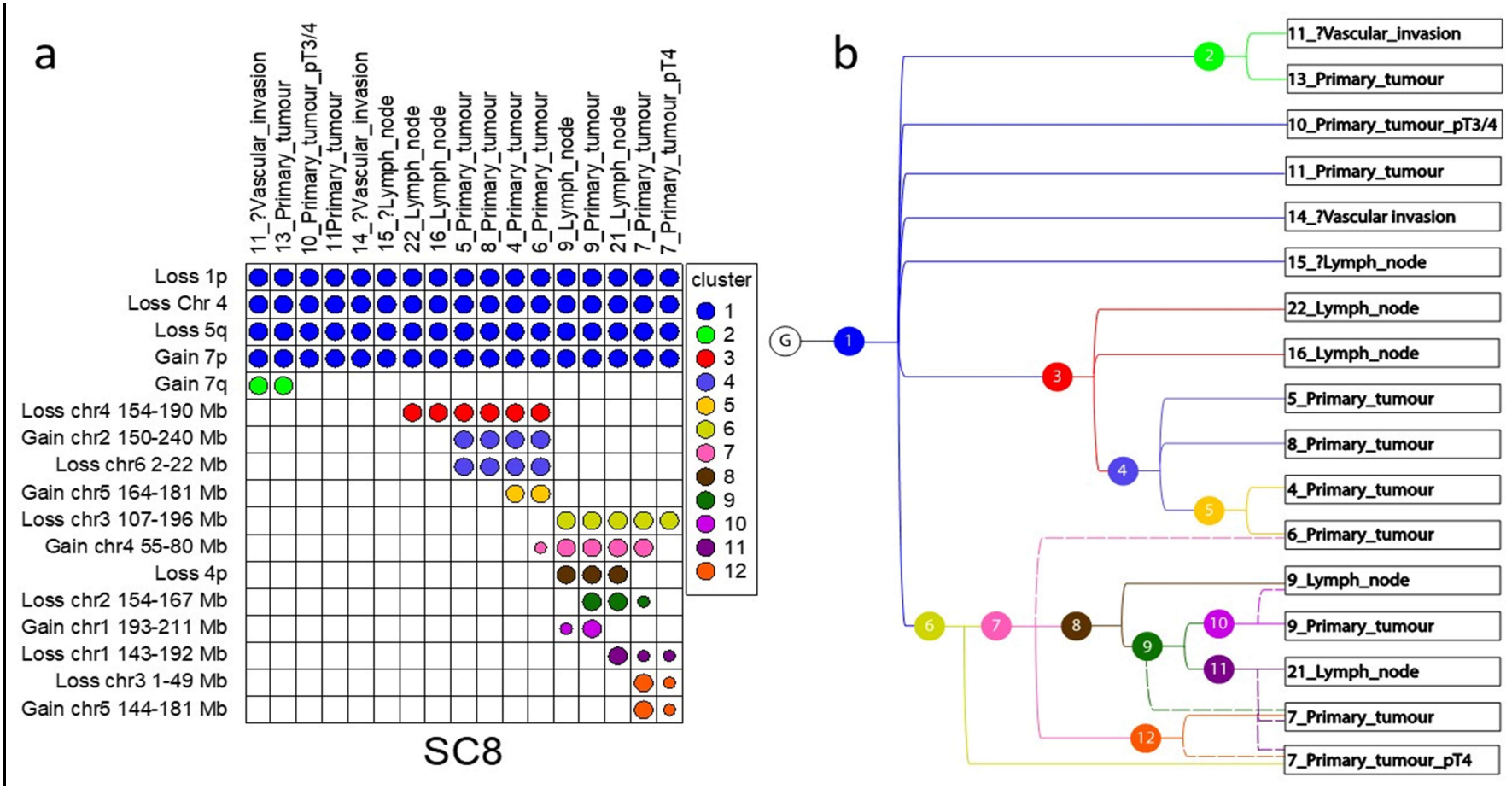


Figure 58: This figure displays the pattern of CNAs in SC8, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC8 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.9 SC9

This case, in contrast to the last, was less well sampled on examination by the reporting pathologist (at the time of resection) and contained fewer lymph nodes, one of which did not produce a sufficient amount of DNA for library preparation, therefore fewer samples were available for analysis. As shown in Figure 59, despite the paucity of material, a diversity of CNA identified with a large number of shared, truncal CNAs including many canonical CRC CNAs (including loss of 8p, 17p and 18q) with the emergence of branch CNAs containing lymph node metastases genomically divergent from the primary tumour and one another. The samples taken from the primary tumour in this case displayed a range of complexity of CNA and polyclonality in two of the three samples; note that these primary tumour samples were large and contained foci of peritoneal involvement and EMVI (as confirmed by histochemical staining for elastin).

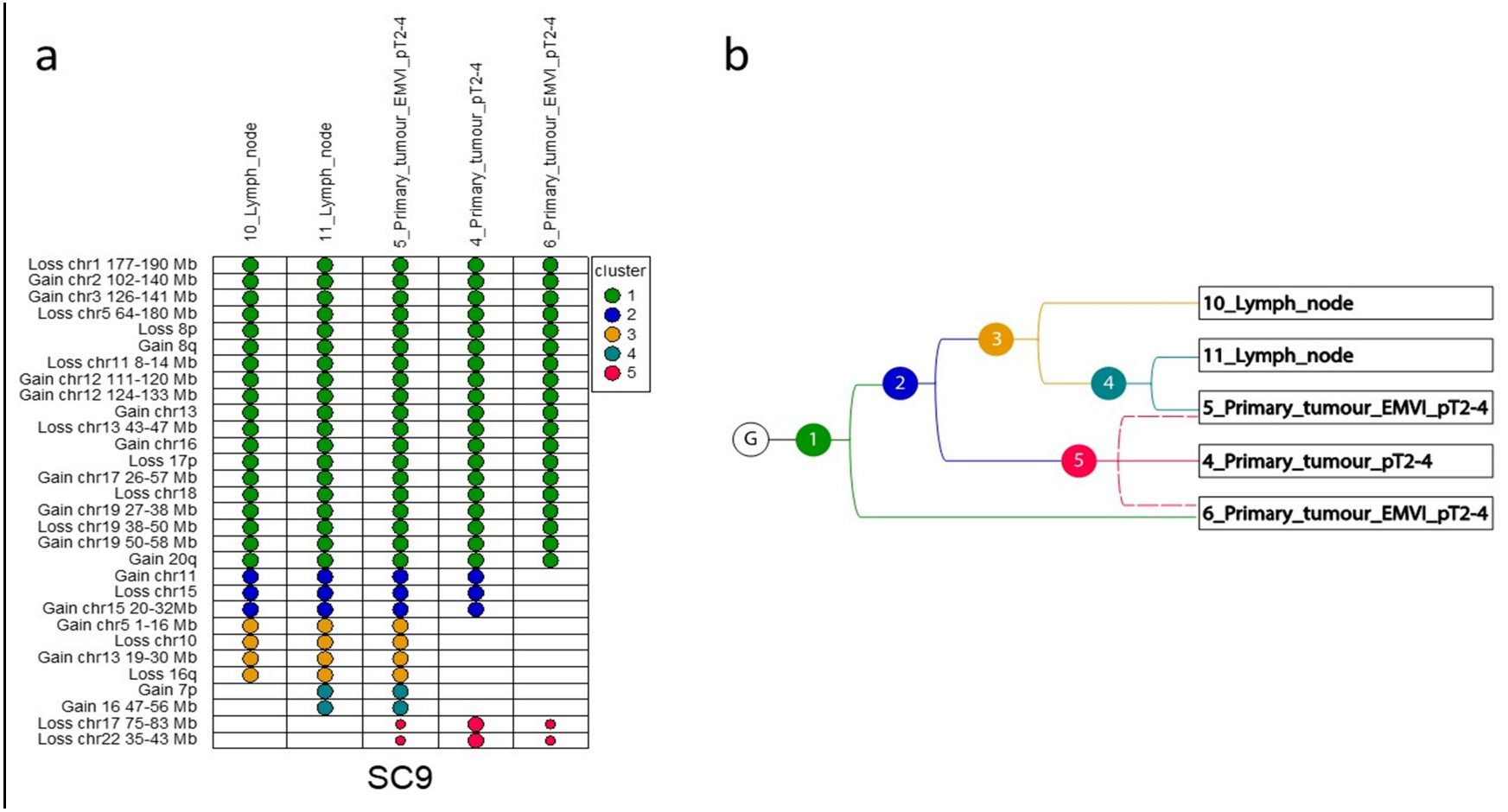


Figure 59: This figure displays the pattern of CNAs in SC9, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC9 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.4.3.10 SC10

This final case (shown in Figure 60) also contained a small number of samples, however it was considered a useful case as the available material from the case comprised a relatively large number of lymph nodes and a primary tumour block containing tumour within extramural veins and perforating through into the peritoneum. ; The case also included one metastatic deposit (number '11') which did not show definitive histological evidence of being either a lymph node or vascular embolus. The clonal diversity in this case was less pronounced than in other cases in this series (most likely reflecting the limited material sequenced) with similar truncal CNAs to those previously described. Once again there was divergence between the primary tumour and the majority of regional metastases. One sample originating from a lymph node metastasis showed a combination of the two 'branch' clusters and was therefore considered polyclonal. The histologically indeterminate lesion (sample 11) grouped with the nodal metastases rather than with the sample from the primary tumour.

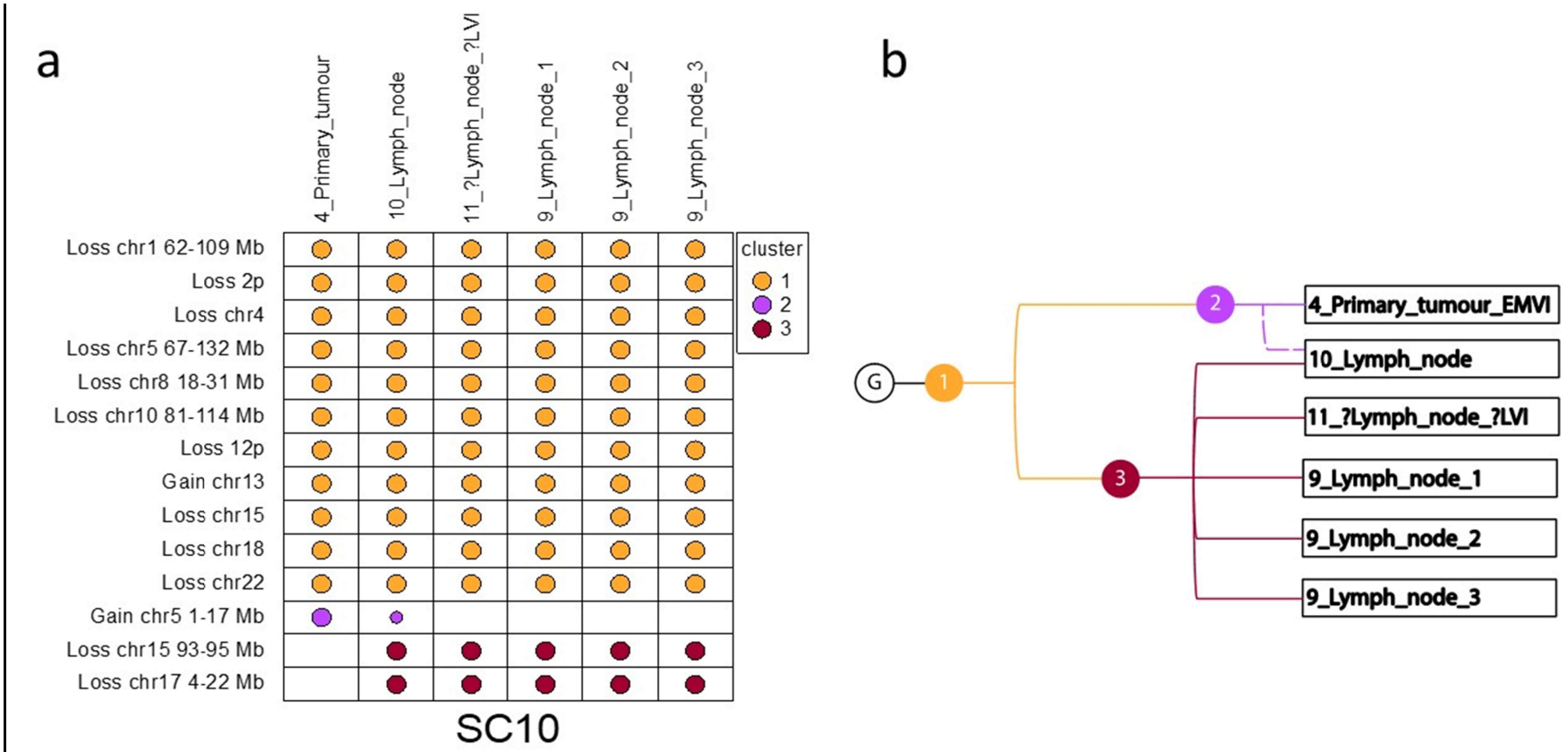


Figure 60: This figure displays the pattern of CNAs in SC10, a) lists all shared, 'truncal' or 'branch' CNAs identified in all the lesions from SC10 on the y-axis, the coloured circles identify the samples in which the CNA is present, the large circles represent those changes which were present in the majority of tumour cells and the small circles are those events which were interpreted as only occurring in a subclone of tumour cells. b) Shows the phylogenetic tree inferred from the distribution of CNAs shown in a), the solid lines are those events present in the majority of cells, the dotted lines are those events interpreted as being 'subclonal'. The G in this figure represents the germline state for the individual.

7.5 Discussion

In this section, we have compared the pattern of CNA within multiple sections of ten locally advanced colorectal adenocarcinomas, with other regions of the primary tumour and tumour sampled from foci of intraperitoneal, intravascular and nodal spread. These ten cases were selected from the archive at LTH as they all exhibited the principal microscopic prognostic features as described in the RCPath minimum dataset for colorectal cancer and the literature cited therein [196]. As such, they represent a cohort of cases most likely to disseminate throughout the body, and examination of these cases may further substantiate or refute observations made in the autopsy cases examined in section 5 and 6. Although the archive at LTH contains many cases exhibiting these features, due to blocks missing from archive and the use of 'mega-blocks' (extra-large FFPE blocks used to demonstrate a whole cross section of tumour which are, due to their size, incompatible with the LCMD platform available), the ten cases selected represent a consecutive sampling of appropriate cases with suitable blocks between 2003-2006. The cases selected were all pMMR as dMMR CRC are unlikely to have been informative, as it has been shown in section 5and in other literature, that dMMR CRC are unlikely to show a significant number of CNAs [52,65] and therefore are unlikely to show evidence of clonal evolution via CIN.

Tumour from the first five cases was dissected from the surrounding normal tissue using LCMD, whereas the second five were macrodissected. The two different methodologies were employed due to constraints of time (as LCMD takes approximately half-a-day per block) and expense (as LCMD produces greater numbers of samples for sequencing and at least ten PEN membrane slides, costing £4 per side, were required for each block microdissected). As the latter five cases contained only occasional foci of tumour requiring LCMD and the same patterns of CNA were identified between the two groups of cases, it is unlikely a significant quantity of data has been lost by the use of macrodissection rather than LCMD.

Within this cohort, by examining the occurrence of shared CNAs, it was possible to identify several features present across all ten individuals. These included the presence of truncal changes shared by all samples from a case followed by the divergence of samples by CNA, the presence of canonical CNAs within the truncal group of CNAs and the emergence of metastases at multiple stages of evolution in each case. Additionally, in the majority of cases the pattern of CNA correlated with those samples which were

from similar regions of the tumour rather than any relation to any mode of metastasis i.e. within the same case, samples originating from the same piece of tissue were more alike than intranodal or intravascular deposits.

The truncal, ubiquitous changes identified in the cases were similar to those described by the Cancer Genome Atlas Network classification [371]. CNAs characteristic of CRC in chromosomes 1, 5, 7, 8, 12, 13, 15, 17, 18, 20 and 22 were all identified on more than one occasion within the truncal changes present in the cohort. Loss of 18q was the most common of these 'canonical' changes identified and was identified in 7 of the ten cases, this frequency is in keeping with that reported in the previously cited literature. Although the purpose of the CNA analysis in this study was not to identify the 'driver' events within this group CRC, the identification of canonical CNAs within the cohort suggests that the methodology and depth of sequencing is adequate to identify CNA and that the cohort examined is at least comparable to CIN CRC within the general population.

In addition to the 'truncal' changes described above, 'branch' CNAs are also identified which were exclusive to a subgroup of samples, suggesting the emergence of new subclones within the tumour cell population. Once again, several canonical CRC CNAs are identified within some of the cases examined (including loss of 4q, 7p, 12q and 20p) raising the possibility of the acquisition of new biologically significant events within new subclones of tumour cells but, to reiterate, this experiment was undertaken to identify the emergence of new clones rather than the identification of 'driver' events.

The emergence of new tumour clones within each case appears to correlate with the anatomical position within the tumour rather than vascular or peritoneal invasion. Furthermore, it was consistently observed that multiple clones emerging from the same tumour were capable of vascular and peritoneal invasion and lymph node metastasis. This pattern of clonality suggests two possible scenarios, the first being 'convergent evolution', by which multiple different subclones within one tumour have converged upon the same biological characteristic. 'Convergent evolution' has been described within metastatic renal cell carcinoma by Gerlinger *et al* [105], however within this work the observed phenomenon was the emergence of the mutations within the same genes, whereas in this series a number of genomically divergent clones exhibited the same biological behaviour. As the process of local and lymphovascular invasion is most likely reliant upon the co-existence of several genetic, epigenetic and local factors it seems improbable that different clones within the same tumour would arrive at appropriate

combination of factors. An alternative, more probable, explanation would be that these tumours developed the capacity for vascular, peritoneal and lymphatic involvement at an early stage of development (within the common 'truncal' stage) and the subsequent divergence is incidental to the progression of the disease.

The development of multiple co-existent clones during tumour evolution (rather than a Darwinian 'selective sweep') has been established within early and [326,372] more advanced primary CRC [288,373], although the biological significance of intratumoural heterogeneity in established tumours is uncertain. Hardiman *et al*[288] performed an in-depth examination of mutational and CNA heterogeneity within six rectal tumours and confirmed a wide range of observable heterogeneity in this small cohort and suggested that this heterogeneity may aid tumour survival and progression. Mamlouk *et al* [83] performed WGS of 68 regions within a single locally advanced CRC and demonstrated a degree of divergence by CNA, some of which corresponded to the position of cancer 'driver' genes, once more suggesting that intratumoural heterogeneity may confer biologically advantageous characteristics. The conclusion drawn within these papers (which represent the most detailed published analysis of locally advanced CRC) is to a degree at odds with data derived from in-depth analysis of colorectal adenoma and carcinoma which suggests that the majority of subclonal events in CRC are, in fact, 'neutral' or passenger events [323,325] and that the majority of CRC acquire biologically significant genomic changes at an early stage. The data presented here, although performed with lower depth sequencing contributes to the understanding of intratumoural heterogeneity within primary CRC as the cited papers only analysed the respective tumours as per the area of tumour, e.g. luminal, lateral and invasive front, and did not examine tumour within specific anatomical compartments relating to metastatic spread as was performed in this work. The patterns present with this cohort is more supportive of the 'neutral evolution' as multiple subclones within the same tumour were capable of intravascular, peritoneal and lymphatic involvement and it is most probable that the capacity for local aggressive behaviour and regional metastases was acquired at an early stage in tumourigenesis before the emergence of the observed clonal diversity. It would therefore follow that any heterogeneity observed within the metastases is simply reflects origin within different anatomical portions of the tumour. Additionally, as blood and lymphatic channels are not distributed homogenously [366,374] throughout the bowel wall, it is possible that the variation between lymph node metastases and those at other sites (as observed in Naxerova *et al* [85]) reflects the vascular anatomy of the bowel rather than a biological characteristic of the tumour.

The theory that the observed clonality within a portion of a tumour is a reflection of the anatomical location (rather than a specific biological characteristic) is relevant to the observed homogeneity within the peritoneal disease documented in case GD1 in section 5 and 6. In this section we had suggested that a group of genomically homogenous lesions within the peritoneum suggests either a single seeding event or that these lesions are exhibiting a specific biological characteristic enabling peritoneal seeding. With the benefit of the data from this section it would suggest that the former explanation is more likely as the capacity for peritoneal and vascular invasion would appear to be ubiquitous across polyclonal tumours.

A central controversy with the argument that tumours are 'born to be bad', acquiring all major biological characteristics early in tumourigenesis is that the presence of polyclonality has been demonstrated to be associated with poor outcome in CRC [375] and new clones have been demonstrated to emerge with the use of targetted therapy [100]. This apparent contradiction is possibly explained by the fact the born to bad hypothesis was generated by sequencing treatment-naive, adenomas and carcinomas and therefore the interaction between tumour heterogeneity, tumour micro-environment and drug effect may explain the adverse prognosis conferred by the presence of multiclonal tumours.

In summary within this section we have demonstrated polyclonality within a cohort of local advanced tumours, and whilst this clonality is associated with anatomical location within a tumour, the ability to invade the peritoneum, vascular structures and produce lymph node metastases is almost ubiquitous across polyclonal tumours. This observation suggests that at least in treatment naive locally advanced CRC heterogeneity is not important to local progression. The potential for aggressive behaviour across all clones within the tumour underlines the importance of accurate radiological and pathological staging, as identification of adverse pathological findings even in small, early tumours, may allow stratification and proactive management of aggressive lesions.

8 Discussion

8.1 The clinical context

mCRC continues to be a significant source of cancer related mortality within the UK and globally [2,199]. The introduction of highly specific therapies producing inhibition of EGFR signaling has produced only small improvements in progression-free and overall survival [376] . Several novel therapies such as VEGF and HER blockade have not been shown to be clinically effective either as monotherapy or in combination with conventional chemotherapy [136,377] or EGFR inhibition [378], although dual HER2 blockade plus EGFR inhibition has shown some efficacy in early trials [379,380]. Even within the subgroup of patients predicted to be responsive to EGFR blockade (i.e. those individuals WT for *N/KRAS*, *BRAF* and *PIK3CA* mutation) disease progression is inevitable and the mechanism of treatment resistance has been observed to be correlated with the emergence of new tumour cell clones bearing either mutations at the aforementioned loci or a more recently described amplification of MET [100,152]. A major issue requiring resolution is whether these new clones arise de-novo on the initiation of therapy or do they exist pre-treatment as part of a heterogenous tumour cell population across the body and are clonally advantaged by treatment [100] ; the data presented in this thesis has attempted to provide insights as to the nature and timing of heterogeneity in advanced CRC.

8.2 Clinical and histological characteristics of metastatic disease

A substantial portion of the data presented within this project was produced from material sampled at autopsy examination of 8 individuals with mCRC. This approach allowed extensive sampling of disease within each patient; in total 375 samples from 8 individuals were taken. This dataset, provided by the 'Gift' research program, provided a significantly greater number of samples than in previous and subsequent autopsy based work in renal cell [105], breast [104], pancreatic [99,106,381] and prostatic carcinoma [87].

The group of patients ('donors') recruited to this work had either contacted Leeds Teaching Hospitals NHS tissue services expressing a desire to donate their body to research after death or via oncology outpatient clinics where a poster was displayed advertising the 'Gift' project. Consent was gained from both the donor and next of kin once the donor had passed away. The recruitment methodology employed for autopsy-based study of advanced malignancy either falls into a largely unselected self-referral method, which has been employed in breast [104] and pancreatic carcinoma [99,106,381], or cohorts of patients who have all received similar treatment modalities either due to standard clinical approach (such as androgen deprivation in prostate carcinoma [87]) or if patients have been recruited from a drug trial cohort [105]; the work present in this thesis falls into the former category. This unselected self-referral method can produce a heterogenous dataset which might be difficult to interpret; although the cohort provided by the 'Gift' program was largely a relatively unique treatment-naive set of tumours and metastases.

The cohort within this study contained 5 individuals that presented with inoperable metastatic disease, two of the three resected cases experienced disseminated recurrent disease and the third died with only bulky local recurrence. The distribution of disease broadly reflected that within the general population; the majority of primary tumours were left sided disease [2], the distant metastases were predominantly within the liver, although pulmonary metastases were identified in several cases and represented the major burden of metastatic disease in one case [382]. Peritoneal disease was also examined. The histopathological characteristics of the tumours were also those that would be expected in a group of biologically aggressive lesions [211,222,223,235–238,246] i.e. all eight cases showed at least areas of poor differentiation plus vascular, perineural and peritoneal invasion; all but one case showed lymph node metastases and all three

resected tumours showed regional lymph node metastases at the time of resection. As only two of the cases from this group had received chemotherapy, it is conceded that this cohort may not be ideal to demonstrate the effect of treatment on tumoural heterogeneity. However, 108 samples from 104 lesions were available from the two post-treatment cases, therefore a significant number of metastatic deposits within a background of previous chemotherapy were analysed.

In summary the autopsy cohort used for this study broadly represented the anatomical distribution and histological characteristics of advanced colorectal cancer within the wider population, the majority were treatment naive and the cohort of cases represented the most extensively sampled autopsy series of advanced carcinoma studied to date.

8.3 Genomic characteristics

The genomic analysis of disseminated carcinoma requires in-depth analysis of each sample from an individual and a comparison of the genomic features within each case. The large number of samples provided by the autopsy series in this project, whilst providing a comprehensive catalogue of lesions within disseminated CRC, was also problematic as the DNA from each sample had to be extracted, sequenced and analysed. There were, therefore, considerable constraints as regards reagent costs and sequencing depth per sample along with the bioinformatic challenge of reliably identifying genomic events across large numbers of samples with variable DNA quality and tumour cellularity. Due to these constraints a pragmatic approach was adopted using a focused initial mutational panel plus low coverage CNA of every sample, followed by whole genome sequencing of a subgroup of cases selected on the data provided by the initial two assays.

8.3.1 Pyrosequencing

The initial mutational panel described in section 4 comprised pyrosequencing for the variants in *N/KRAS*, *BRAF* and *PIK3CA* which are the most powerful predictive mutational biomarkers in CRC [8,32,163,363,383] . Pyrosequencing was chosen for this initial assay as, at the time of analysis, it was the most commonly used method for clinical mutational testing in CRC within the UK [273,384]. It was the modality of choice as it provided a balance of specificity and sensitivity superior to the alternate sequencing modalities available prior to the widespread deployment of next generation sequencing for clinical genetic testing [275–278].

Each case within the 'Gift' autopsy cohort demonstrated a mutation or variant within the pyrosequencing panel and only one mutation was observed within each case; *KRAS* exon 2 (codon 12 or 13) mutations were present in 4 of the 8 cases, 2 cases showed *BRAF* codon 600 mutations, the remaining cases showed *KRAS* codon 147 and *NRAS* codon 16 respectively. Most significantly, as regards tumoural heterogeneity, within the 367 tumour samples sequenced only three tumour samples were shown to be WT for the entire panel of mutations. Two of these three samples originated from small nodal deposits raising the possibility of a false negative result on the grounds of low tumour cellularity; further sequencing using higher depth NGS showed this to be the case in one of the two wt nodal samples. Interesting, meta-analytical data has suggested that there may be a higher rate of mutational discordance between primary tumours and regional

lymph node deposits as compared to primary tumours and distant metastases [86] however, these studies would face a similar issue of the smaller amount of tumour available from lymph node metastases than primary or resected liver secondaries. The third WT sample originated from one of several samples taken from the primary tumour in the second autopsy case (GD2); the remainder of the samples in this case yielding a *KRAS* codon 12 mutation. This outlying sample microscopically appeared to contain a significant volume of tumour and produced good quality NA according to the QC steps within the protocol. As such it was felt that this tumour sample was genuinely WT either as this region of the tumour from which the sample had originated had not acquired the tumour or the genomic material bearing the *KRAS* mutation exhibited by the other sample within the case had been lost through a LOH-type event.

The occurrence of intratumoural heterogeneity within primary CRC is fairly well established within the literature and several studies, using a variety of assays, have made similar observations [287,288]. Similarly the striking intertumoural homogeneity in disseminated disease within the 'Gift' cohort confirms the observations of previous studies examining matched, resected primary and metastatic CRC [81,82,385,386]. It is of note that the presence of lung metastases within this cohort correlated with the presence of *KRAS* mutation, this concordance has also been documented in the published literature [387]. In conclusion, despite exhaustive sampling, it was not possible to demonstrate marked intertumoural heterogeneity at key predictive loci within the 'Gift' autopsy cohort. It would appear, therefore, that in an untreated population, biopsy and mutational testing of distant metastases would not alter the management of patients with mCRC.

8.3.2 Low coverage WGS for CNA

In contrast to the pyrosequencing panel, low coverage WGS for CNA analysis did highlight extensive genomic heterogeneity suggestive of clonal diversity within the 'Gift' cohort and provided a dataset from which the phylogenetic relationship of each case could be inferred. A litany of methods for the determination of phylogeny within tumour cell populations have been published, many of which are automated probabilistic algorithms [117,121–123,125,126] such as TuMuLT [125] and Pyclone [321]. Whilst these methods may be quick and robust when handling small numbers of samples with fairly parsed data, it was felt they were inappropriate for analysis of the 'Gift' cohort data as they were either not designed to manage NGS CNA data or were only capable of incorporating a limited number of samples into their inferred evolutionary relationships [126] . As a solution

to this methodological issue a semi-automated approach was adopted to infer the phylogenetic relationship between each sample within an individual. This method involved an automated bioinformatic application to identify the CNAs in each sample [321], which were then interpreted manually using Dirichletian principals of phylogeny in a similar manner to comparable work in breast carcinoma [117]. The application of this semi-automated Dirichletian method of phylogenetic analysis to the 'GIFT' autopsy CNA data highlighted the patterns of clonal diversity across the eight cases.

Firstly, the cases fell into the two major categories of CRC, six (GD1, GD2, GD5-GD8) which displayed evidence of chromosomal instability and two (GD3 and GD4) which appeared to be chromosomally stable. This finding correlated with the immunohistochemical staining performed in section 4 which suggested that the latter two cases were dMMR and unlikely to show CIN [61,371]. The two dMMR/non-CIN cases appeared to retain this genomic characteristic throughout the progression of disease, demonstrating minimal evidence of clonal evolution through CNA in the 79 samples sequenced and suggesting that if significant clonal divergence were occurring it was most likely at the mutational or epigenetic level. This supposition was explored in section 6 and will be described below.

Within those cases which displayed evidence of CIN (GD1, GD2 and GD5-GD8), the presence of shared CNAs allowed a phylogenetic tree to be inferred for each case which illustrated the pattern of clonal diversity and evolution in each case. This phylogenetic analysis demonstrated genomic divergence between deposits in all but one case (GD6); several samples within the cohort were considered to be 'polyclonal' showing CNAs characteristic of different branches of the phylogenetic tree for the relevant case. The outlying case (GD6) showed evidence of CIN but the CNAs identified within the case were almost ubiquitous across all tumour deposits; it would appear most likely that this tumour showed an aggressive phenotype/genotype from an early stage and neither required any new 'driver' events from the early stages of disease to disseminate and/or progressed so quickly that new events did not accumulate over time. An alternate explanation would be that the tumour cell population within the individual was evolving through a different means i.e. mutational or epigenetic events, and whilst this is a possibility it appears unlikely as the tumour did not demonstrate the acquisition of new mutational 'driver' events on WGS (as described in section 6) and the clinical history (as described in section 3) was that of rapidly progressive disease.

Therefore only 5 of the 8 cases showed evidence of evolution through CIN/CNA; within these cases the character and degree of clonal evolution varied significantly. At one end of the spectrum there were two cases, GD2 and GD5, which showed divergence between different sites of metastasis (liver versus lung disease and lung versus lymph node disease respectively). This observation is similar to those made by Naxerova *et al* [85], who described clonal divergence between lymph node and liver metastases within a cohort of resected CRC by analysing hypermutable regions of the genome. In contrast the three remaining cases showed a far more complex pattern of phylogeny with significant overlap between lung, nodal and liver metastases; although GD1 showed a relatively divergent population of peritoneal metastases. This range of complexity within the cases is a difficult picture to dissect with only the snapshot of information provided at post mortem examination. The two possible scenarios are that there are a group of CRCs within which individual clones exist which are most suited to disseminating to a specific metastatic site or the presence of polyclonality is entirely incidental to disease progression and the observed divergence is a random event.

To further explore the timing and emergence of metastatic clones, in section 7 phylogenetic analysis was performed on archived material from 10 locally advanced CRC. The selected cohort comprised tumours showing vascular and peritoneal invasion along with local nodal metastases; a combination of LCMD and macrodissection was used to isolate regions of the primary tumour including intravascular and intraperitoneal tumour in each case. The DNA from the dissected portions of the primary tumour and regional lymph node deposits was then sequenced by low coverage WGS. Semi-automated phylogenetic analysis was performed on each case in order to characterise the emergence of tumour cell clones within locally advanced CRC and to correlate the clonal phylogeny of each case with the anatomical location of tumour. As observed in the 'Gift' autopsy cohort and the previously cited literature [83,288,326,372,373], the locally advanced CRC within this cohort showed evidence of heterogeneity and clonal evolution by CNA. However, the detailed microscopic analysis and LCMD provided the additional insight that, although the pattern of clonality correlated with anatomical location within the tumour (as per the observation in Mamlouk at al [83]), the capacity to produce lymph node deposits and infiltrate vascular and peritoneal structures was present across multiple subclones within a tumour. It is felt that whilst this phenomenon may have arisen via 'convergent evolution' [105], it was more probable that these biological characteristics arose early within the development of the respective lesions and were retained during tumour growth and the development of CNA polyclonality. This conclusion was drawn

first upon the work performed by Sottoriva *et al* [323] which suggested that the majority of 'driver' events occur early in the development of CRC and secondly due to the improbability of the acquisition of biologically complex characteristics independently within the same cancer cell population [119].

Thus, within this thesis low coverage WGS of both disseminated and locally advanced tumours has shown evidence of clonal evolution through CNA. However, the pattern of phylogeny is not convincingly associated with either the establishment of metastasis or aggressive local behaviour; hence it could be suggested that clonal evolution and heterogeneity via CNA is not important to the progression and behaviour of CRC. This conclusion underlines the importance of accurate radiological and pathological reporting of CRC as, outwith the core genomic features of a tumour, it would appear that anatomical access to routes of metastasis is of greater significance than subclonal genomic changes.

A significant caveat to this conclusion is that the 'branch' CNAs identified within the 'Gift' cohort often included events thought to be important in the establishment of CRC and other neoplasms [371] and within the wider literature tumour heterogeneity is shown to be linked to poor prognosis [375]; furthermore clonal evolution is almost certainly central to treatment resistance in highly specific therapies [100,388]. Within this thesis only two tumour cell populations have been exposed to systemic chemotherapy and neither of these two cases completed a course of EGFR inhibition, the treatment most strikingly associated with the emergence of clonal resistance and therefore perhaps in the absence of targeted therapy clonal evolution does progress in a neutral fashion. Alternatively, the key evolutionary events may have occurred at a mutational level in a wider range of loci than examined in section 4, therefore WGS was performed at the WTSI on 4 of the 'Gift' cases. These four cases were the two predicted to be dMMR by immunohistochemistry (GD3 and 4) and two cases at either end of the observed spectrum of CNA heterogeneity (GD1 and GD6). A targeted high depth panel of 'driver' genes was also performed on GD8.

8.3.3 WGS for mutational analyses

The WGS sequencing performed at a depth of 60x at WTSI produced a large volume of sequencing data, from this the 'CaVEMan' [105] and 'Pindel' [338] algorithms further illustrated the essential genomic differences between the pairs of CIN (GD1 and GD6) and dMMR (GD3 and GD4) CRC submitted for sequencing. Firstly, non-synonymous somatic mutations were present within genes related to MMR within both CRC predicted to be dMMR on IHC. The variant caller data also showed a 30-fold greater mutational burden within the dMMR cases, the magnitude of this disparity was within the spectrum described in the literature [52,371].

Hierarchical clustering using 'hclust' was performed using a matrix produced from variant caller data for GD1, GD4 and GD6; only 2 samples from GD3 were suitable for variant caller analysis and therefore clustering analysis would have been uninformative. This coarse method of comparison demonstrated that the peritoneal disease in GD1 was relatively genomically homogenous, thus confirming the observation made in section 5 and supporting the validity of the phylogenetic analysis from this section. The 'hclust' algorithm provided some evidence of divergence between the primary tumour and metastases in GD6 but was largely uninformative in GD4 with no apparent correlation between the anatomical location and clustering pattern. The lack of a comprehensible clustering pattern in GD4 was possibly reflective of a genuinely pattern-less path of evolution, however the volume of data from this case along with variation in the DNA quality and sequencing depth may have confounded the method of comparison.

In an attempt to parse the large volume of data, manual interrogation was performed at a number of 'driver' loci. By using the 'mpileup' function of the SAMtools [346] application it was possible to confirm the presence of the ubiquitous mutations identified by pyrosequencing (in section 4) and suggested the presence of mutational heterogeneity within the CIN and dMMR cases. However, on examination of the sequencing data it was apparent that many variants across all four cases where either supported by low VAFs or were present in areas of low coverage, arbitrary cut-offs of 5% and 10% were adopted to further improve the robustness of variant identification. Within these parameters it appeared that the heterogenous variants within the CIN cases (GD1 and GD6) were present in regions of poor sequencing coverage and were supported by only a small number of reads; the dMMR (GD3 and GD4) cases showed better quality evidence of heterogeneity at several 'driver' loci between samples in each case.

The CIN CRC case (GD8) submitted for target capture sequencing with 200x coverage for a panel of 200 driver genes also demonstrated a ubiquity of several key canonical CRC mutations. An additional putative 'driver' mutation in *FBXW7* [19] was also identified within the samples from this case, however it was only present in the DNA extracted from the resected primary tumour and local metastases and all of the post mortem samples of disseminated disease were WT for this mutation. No CNAs were identified within the region of *FBXW7* on low coverage WGS, therefore it is probable distant metastasis had occurred prior to the formation of local lymph node metastasis and attempted cure by local resection (although LOH in the relevant genomic region remains a possibility). Once again this observation lends weight to the theory that CRC develop the repertoire of genomic changes required for dissemination at an early stage and subsequent evolution may be neutral [325] .

The first dMMR case (GD3) provided DNA suitable for WGS (at 60x) from two distinct anatomical locations, an abdominal wall/peritoneal metastasis and recurrent pelvic disease. These two sites shared a *BRAF* V600E mutation but also showed variants in *TP53* and *PIK3CA* with relatively high VAFs that occurred in a mutually exclusive fashion between the two sites. The second dMMR case (GD4) showed a far more complex pattern of heterogeneity, variants private to a subgroup or a single sample were present at 9 of the loci examined; this group of variants included alterations in *TP53, PIK3CA*, *RB1*, *SMAD2*, *SF3B1, USP9X*, *NF1*, *CTNNB1*, *FBXW7* and *PTEN*. These variants did not occur in a mutually exclusive fashion between samples and the allelic frequencies of many of the variants were suggestive that the majority of heterogenous variants only occurred in a subgroup of tumour cells within a sample. Although 60x coverage is inadequate to completely exclude the presence of a variant within any sample, it is unlikely that all of the heterogeneity identified is simply due to variation in sequencing depth and DNA quality between samples. Therefore, although further sequencing at greater depth (at perhaps a limited number of loci) would be required to determine a pattern of phylogeny in GD4, the overall picture within both dMMR CRC was of highly aberrant genomes with a complex mutational patterns, both ubiquitous and private, which included several variants in genes related to chemotherapeutic resistance [363] and oncogenesis [342–345].

Thus, the tumour cell population within disseminated CRC is a group of different clones sharing common progenitor mutations, evolving in parallel and developing a spectrum of genomic changes some of which may provide survival advantage given the correct

stimulus e.g. EGFR blockade. The natural evolution of tumour cell clones appears to concur with the hypothesis for the importance of the biological characteristics driving early tumour development, and it is likely that those genomic changes occurring early in tumour development are most important to a tumour's capacity for subsequent local or disseminated spread.

Due to the absence of detectable clonal evolution via new mutational events within the cohort of CIN CRC examined in this thesis and paucity of such events within the cited literature [83,84], it would appear most likely that those new *RAS* mutant clones detected by Diaz *et al* [100] were *do novo* mutations occurring after the initiation of EGFR therapy. It is conceivable that extremely low frequency variants are present within the samples in this thesis, detection of such variants would require ultra-high depth sequencing but still may be impossible to rule out due to the limitations of sequencing sensitivity and post mortem DNA degradation. Due to the presence of such large mutational burdens in dMMR tumours, it is possible that *KRAS* mutated clones may be present in a pre-treatment tumour cell population. This possibility would correlate with the lower rates of response in right sided CRC to EGFR blockade, as dMMR CRC tend to arise within the proximal bowel [389,390], although this genomic characteristic is likely one of several factors producing this trend [391].

8.4 Further work

The post mortem material gathered as part of this study provides an opportunity for further study of both the genomic and epigenetic aspects of clonal evolution in mCRC. As acknowledged in section 6 the WGS sequencing performed at the WTSI was of insufficient depth to exclude low frequency mutational events therefore higher depth target capture sequencing (such as performed on case GD8) could be used to further establish the presence mutational heterogeneity within the GIFT cohort. Alternate methods of determining phylogeny, such as the analysis of hypermutatable regions of the genome (as described in Naxerova *et al* [258]), could be performed to validate the observations made by CNA analysis although as identical CNAs were identified across multiple samples in the CIN CRC GIFT cases the observations made in section 5 are likely to be reliable. An examination of the epigenetic variation within the GIFT cohort would give additional detail as to the overall pattern of heterogeneity within CRC. DNA methylation is the most extensively studied source of epigenetic variation within CRC as it characterises a subcategory of CRC [50] and detection of methylation at specific loci may have value as predictive markers for conventional chemotherapy, radiotherapy and EGFR blockade [392–395]. A further piece of the heterogeneity puzzle would also be highlighted by examination of gene expression, for example expression of EGFR ligands is known to be predictive of EGFR blockade response [396]. In summary, to produce a comprehensive assessment of heterogeneity regarding factors impacting prognosis and treatment-effect would require extensive examination of translational and epigenetic factors in addition to genetic assays. This examination may be highly complex as both methylation and gene expression are relatively transient compared to the genomic features of the malignant cells [93,397–399] but this analysis would provide valuable information as to the process of metastasis and disease resistance.

As the sequencing data presented in this thesis and the published literature suggests that the genomic events conferring to the biological characteristics of a tumour arise early in tumour development, the significance of genomic heterogeneity occurring at metastatic sites is therefore most relevant to therapeutic resistance rather than the natural progression of disease. Based upon this premise and using this cohort as a basis for comparison study of a second cohort patients treated with targeted therapies would address several issues raised in this thesis and within the literature and could inform clinical data if incorporated into a clinical trial such as FOCUS-4 [171].

Firstly, it appears that heterogeneity for *RAS* mutation is unusual in EGFR blockade naive CIN CRC and an autopsy-based study of fatal *RAS* WT tumours treated with EGFR blockade would be valuable in examining the emergence of treatment resistant tumour cell clones. Whilst this phenomenon is identifiable and documented by tissue biopsy and via the analysis of CTDNA [148–153], the use of exhaustive post mortem sampling would complement the current data in several ways. Post-mortem sampling would document the pattern of disease resistant clones across the body and describe whether new RAS mutations arise at a single or multiple sites thereby confirming the presence of 'convergent evolution' described by Gerlinger *et al* [105]. Additionally, the anatomical distribution of treatment resistant clones provide evidence that certain anatomical sites are less likely to produce disease resistant tumour cells, possibly suggesting lower exposure to systemic therapies. Autopsy based study of treated mCRC would also provide a greater understanding of the molecular characteristics of treatment resistant disease i.e. do CIN CRC show a greater degree of mutational heterogeneity across the genome when treated with EGFR blockade than would be expected in treatment naive disease and therefore is there a shift in the molecular characteristics of tumours receiving EGFR. Finally, the current published comparisons of matched ctDNA versus limited tissue biopsy in mCRC treated with EGFR blockade have shown significant discordance in mutation detection [152]; correlation between ctDNA with tissue sampled in the antemortem and post-mortem setting may provide a powerful validation tool for ctDNA sequencing.

9 Summary of conclusions

9.1 Section 3: the 'Gift' autopsy project

- The distribution of disease within the 'Gift' autopsy cohort was in-keeping with that within the cited literature

- In total 375 samples from 266 tumour deposits were taken during the consent autopsy procedure, this depth of the sampling was more extensive than comparable studies within the current published literature.

- The eight cases within the 'Gift' cohort demonstrated only minimal variation according to the histopathological features predictive of outcome in CRC. The histological examination incorporated H and E, histochemical and immunohistochemical staining to characterise each case.

- The 'Gift' cohort contained two cases dMMR CRC according to immunohistochemistry.

9.2 Section 4: Targeted mutational sequencing of therapeutically significant loci in disseminated colorectal cancer

- Pyrosequencing at 12 therapeutically significant loci within *KRAS, NRAS, PIK3CA* and *BRAF* demonstrated mutations in each of the eight 'Gift' cases

- Only a single locus was mutated in each case, no double mutants were identified.

- Within each case the mutation identified was almost ubiquitous across every sample; with only 3 wt samples present in 367 samples successfully sequenced.

- Two of the three wt samples originated from small regional lymph node deposits, one of which was subsequently shown to have a low frequency mutation. The third was one of several samples from a large primary tumour.

- This work demonstrates that the mutation status at the major therapeutically predictive loci is almost entirely homogenous and, whilst this has been shown in

studies of resected disease, the data within this thesis is unique as it incorporates extensive analysis of the disseminated disease state.

- Therefore, in the context of current predictive genomic testing in CRC, analysis of the primary tumour will give an accurate representation of the mutation status across the entire bodily tumour burden.

9.3 Section 5: Copy number alteration analysis in disseminated CRC by next generation sequencing

- CNA data generated via low coverage WGS provided evidence of genomic heterogeneity and clonal evolution with the non-dMMR/CIN cases in the 'Gift' cohort.

- The cases dMMR CRC (according to IHC) do not show evidence of clonal evolution via CNA within the disseminated disease state

- The majority of CNA within the six cases of CIN CRC were shared across all samples within each case. This suggests the events arose at an early stage of tumour evolution; only small numbers of CNA characterised further clonal evolution.

- CIN CRC cases displayed a wide range of clonal complexity according to the pattern of CNA

- The pattern of clonal evolution broadly conformed to the anatomical location of tumour deposits, this trend was most pronounced the less aneuploid tumours.

- The pattern of evolutionary divergence between distant and locoregional metastases suggested that distant metastases may arise prior to the emergence of regional lymph node metastases.

- Thus, the data in this section demonstrates clonal evolution within the 'Gift cohort', although the majority of genomic alterations are early events. The association of clonal evolution and site of metastasis may suggest either a biological predilection by a tumour cell clone or passenger events arising in regions of the tumour with anatomical access to a site or mode of spread.

9.4 Section 6: Cadaveric material mutational sequencing

- WGS at a depth of 60x was performed on four cases in the 'Gift' cohort, showing a significantly great mutational burden within the dMMR CRC; somatic mutations in MMR-related genes were identified in all deposits from the dMMR CRC cases sequenced.

- According to variant caller-filtered WGS data, the degree of clonal evolution via mutational events was greater in dMMR than CIN CRC

- Hierarchical clustering of variant caller data correlated with the Dirichletian phylogenetic analysis of CNA data performed in section 5 supporting the validity of the latter methodology.

- dMMR CRC demonstrate reliable evidence of mutational heterogeneity at 'driver' loci whereas the two cases of CIN CRC sequenced at 60x depth showed only unreliable, low frequency heterogenous mutations.

- Higher depth 'target capture' sequencing was performed on one case of CIN CRC which did identify mutational heterogeneity between resected disease and subsequent recurrent disseminated disease. This comprised a putative 'driver' mutation in FBXW7 within the resected disease which was absent from any of the metastatic lesions sampled at post mortem; suggesting early dissemination of disease prior to locoregional metastasis.

9.5 Section 7: Genomic heterogeneity in locally advanced colorectal cancer

- Heterogeneity and clonal evolution via CNA were demonstrable within a cohort of resected, locally advanced CIN CRC.

- Evidence of CNA was demonstrable via low coverage WGS in intravascular, intranodal and perineural tumour dissected using LCMD and macrodissection.

- The pattern of CNA correlated with anatomical location within the tumour rather than a specific mode of metastasis.

- The capacity to invade blood vessels and produce local metastatic disease was exhibited by multiple clones within the same tumour, suggesting either convergent evolution or more likely these capacities were ubiquitous across the tumour.

- It is probable therefore that the association the anatomical distribution of clones in section 5 most likely represents anatomical access to a metastatic site for a tumour subclone rather than genuine biological heterogeneity.

- The pattern of CNA therefore underlines the importance of accurate identification of high risk histological and radiological features within CRC.

10 Appendix 1: 'Gift' programme posters

GIFT RESEARCH PROGRAMME

Patients with advanced cancer occasionally ask us whether they can donate their bodies to research after death

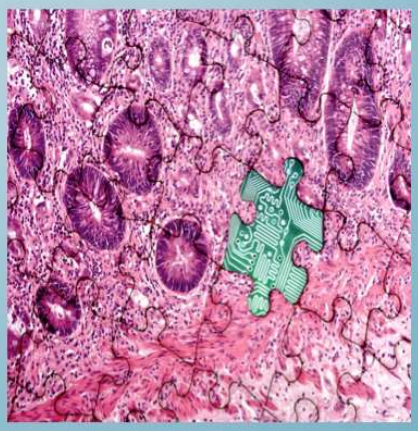
We have a research project where this is possible for patients with ***colorectal*** *or* ***liver*** *cancer*

D.Beirne - Oncology Research, Bexley Wing

GIFT RESEARCH PROGRAMME

This can be of real value in adding to our knowledge about cancer

If this is something you are willing to consider please discuss this with your consultant



D.Beirne - Oncology Research, Bexley Wing

11 Appendix 2: Primers for pyrosequencing

Mutation	Primer name	Primer sequence	Primer length	Amplicon length (bp)	Amplicon sequences. Primers in bold and italics
KRAS CODONS 12+13	KRAS_12+13_MDF_v2_F	GGCCTGCTGAAAATGACTGA	20	80	***GGCCTGCTGAAAATGACTGA***ATATAAACTTGTGGTAGTTGGAGCTGGTGGCGTAGGCA***AGAGTGCCTTGACGATACAGCT***
	KRAS_12+13_MDF_v2_R_B	AGCTGTATCGTCAAGGCACTCT	22		
	KRAS_12+13_MDF_v2_S	AAACTTGTGGTAGTTGGA	18		
KRAS CODON 61	MDF_KRAS_61_F	AATTGATGGAGAAACCTGTCTCTT	24	86	***AATTGATGGAGAAACCTGTCTCTT***GGATATTCTCGACACAGCAGGTCAAGAGGAGTACAGTGC***AATGAGGGACCAGTACATGAGGA***
	MDF_KRAS_61_R_B	TCCTCATGTACTGGTCCCTCATT	23		
	MDF_KRAS_61_S	GGATATTCTCGACACAGC	18		
KRAS CODON 146	KRAS_146_F_B	TCAGGACTTAGCAAGAAGTTATGG	24	100	***TCAGGACTTAGCAAGAAGTTATGG***AATTCCTTTTATTGAAACATCANNAAAGACAAGACAGGTAAGTAACACTGAA***ATAAATACAGATCTGTTTTCTGCA***
	KRAS_146_R	TGCAGAAAACAGATCTGTATTTAT	24		
	KRAS_146_S	GTGTTACTTACCTGTCTTGT	20		
BRAF CODON 600	BRAF_V600E_v2_F	TGAAGACCTCACAGTAAAAATAGG	24	91	***TGAAGACCTCACAGTAAAAATAGG***TGATTTTGGTCTAGCTACAGTGAAATCTCGATGGAGTGGGTCCC***ATCAGTTTGAACAGTTGTCTGGA***
	BRAF_V600E_v2_R_B	TCCAGACAACTGTTCAAACTGAT	23		
	BRAF_V600E_v2_S	TGATTTTGGTCTAGCTACA	19		
NRAS CODONS 12+13	NRAS_12+13_F	CTTGCTGGTGTGAAATGACTGAG	23	79	***CTTGCTGGTGTGAAATGACTGAG***TACAAACTGGTGGTGGTTGGAGCAGGTGGTGTTGG***GAAAAGCGCACTGACAATCCA***
	NRAS_12+13_R_B	TGGATTGTCAGTGCGCTTTTC	21		
	NRAS_12+13_S	CTGGTGGTGGTTGGA	15		
NRAS CODON 61	NRAS_61_F_B	GAAACCTGTTTGTTGGACATACTG	24	83	***GAAACCTGTTTGTTGGACATACTG***GATACAGCTGGACAAGAAGAGTACAGTGCCATGAGAGAC***CAATACATGAGGACAGGCGA***
	NRAS_61_R	TCGCCTGTCCTCATGTATTG	20		
	NRAS_61_S	CTCTCATGGCACTGTACT	18		
PIK3CA CODON 542	PIK1624_F_B	AAAGCAATTTCTACACGAGATCC	23	79	***AAAGCAATTTCTACACGAGATCC***TCTCTCTGAAATCACTGAGCAGGAGAAAGATTTT***TCTATGGAGTCACAGGTAAGTGC***
	PIK1624_R	GCACTTACCTGTGACTCCATAGA	23		
	PIK1624_S	TTCTCCTGCTCAGTGAT	17		
PIK3CA CODONS 545+546	PIK1633_F	ACAGCTCAAAGCAATTTCTACACG	24	95	***ACAGCTCAAAGCAATTTCTACACG***AGATCCTCTCTCTGAAATCACTGAGCAGGAGAAAGATTTTCTATGGA***GTCACAGGTAAGTGCTAAAATGGA***
	PIK1633_R_B2	TCCATTTTAGCACTTACCTGTGAC	24		
	PIK1633_S	GATCCTCTCTCTGAAATC	18		
PIK3CA CODON 1047	PIK3140_F_B	TGAGCAAGAGGCTTTGGAGTAT	22	102	***TGAGCAAGAGGCTTTGGAGTAT***TTCATGAAACAAATGAATGATGCACATCATGGTGGCTGGACAACAAAAATGGATTG***GATCTTCCACACAATTAAACAGCA***
	PIK3140_R	TGCTGTTTAATTGTGTGGAAGATC	24		
	PIK3140_S	GTTGTCCAGCCACCA	15		

12 Appendix 3: Pyrosequencing results by sample

The tables presented below show the mutations present in each sample taken from each individual; where a mutation is present the variant allele frequency is also stated.

Table 34: Table showing pyrosequencing results from GD1

Sample no	*BRAF*	Variant allele frequency	*KRAS*	*NRAS*	*PIK3CA*
1	c.1799T>A	56.4	WT	WT	WT
3	c.1799T>A	56.7	WT	WT	WT
5	c.1799T>A	37.1	WT	WT	WT
7	c.1799T>A	44.4	WT	WT	WT
9	c.1799T>A	47.1	WT	WT	WT
11	c.1799T>A	40.8	WT	WT	WT
13	c.1799T>A	43.4	WT	WT	WT
15	c.1799T>A	50.9	WT	WT	WT
17	c.1799T>A	43.7	WT	WT	WT
19	c.1799T>A	58.6	WT	WT	WT
21	c.1799T>A	46.9	WT	WT	WT
23	c.1799T>A	68.8	WT	WT	WT
25	c.1799T>A	57.5	WT	WT	WT
27	c.1799T>A	32.0	WT	WT	WT
29	c.1799T>A	36.1	WT	WT	WT
31	c.1799T>A	50.1	WT	WT	WT
33	c.1799T>A	53.4	WT	WT	WT
35	c.1799T>A	23.4	WT	WT	WT
37	c.1799T>A	42.4	WT	WT	WT
39	c.1799T>A	31.9	WT	WT	WT
41	c.1799T>A	44.3	WT	WT	WT
43	c.1799T>A	25.1	WT	WT	WT
45	c.1799T>A	48.5	WT	WT	WT
47	c.1799T>A	29.8	WT	WT	WT
49	c.1799T>A	48.6	WT	WT	WT

51	c.1799T>A	44.8	WT	WT	WT
53	c.1799T>A	48.2	WT	WT	WT
55	c.1799T>A	33.7	WT	WT	WT
57	c.1799T>A	28.3	WT	WT	WT
59	c.1799T>A	49.0	WT	WT	WT
61	c.1799T>A	48.6	WT	WT	WT
63	c.1799T>A	36.6	WT	WT	WT
65	c.1799T>A	54.6	WT	WT	WT
67	c.1799T>A	31.1	WT	WT	WT
69	c.1799T>A	29.2	WT	WT	WT
71	c.1799T>A	40.7	WT	WT	WT
74	c.1799T>A	41.1	WT	WT	WT

Table 35:Table showing pyrosequencing results from GD2

Sample no.	*BRAF*	*KRAS*	Variant allele frequency	*NRAS*	*PIK3CA*
1	WT	c.35G>A	34.0	WT	WT
2N	WT	c.35G>A	17.5	WT	WT
2EN	WT	c.35G>A	26.1	WT	WT
3N	WT	c.35G>A	33.8	WT	WT
3EN	WT	c.35G>A	32.8	WT	WT
4	WT	c.35G>A	39.0	WT	WT
5	WT	c.35G>A	46.7	WT	WT
6	WT	c.35G>A	34.4	WT	WT
7	WT	c.35G>A	42.9	WT	WT
8	WT	c.35G>A	18.7	WT	WT
9L	WT	c.35G>A	23.1	WT	WT
9NL	WT	c.35G>A	19.44	WT	WT
22	WT	c.35G>A	46.2	WT	WT
23	WT	c.35G>A	42.6	WT	WT
24	WT	c.35G>A	42.1	WT	WT
25	WT	c.35G>A	29.0	WT	WT
26	WT	c.35G>A	47.0	WT	WT
27	WT	c.35G>A	46.0	WT	WT

28	WT	c.35G>A	43.1	WT	WT
29	WT	c.35G>A	28.7	WT	WT
30	WT	c.35G>A	31.2	WT	WT
31	WT	c.35G>A	42.2	WT	WT
32	WT	c.35G>A	32.5	WT	WT
33	WT	c.35G>A	43.7	WT	WT
34	WT	c.35G>A	43.3	WT	WT
36	WT	c.35G>A	41.2	WT	WT
37	WT	c.35G>A	43.0	WT	WT
38	WT	c.35G>A	44.1	WT	WT
39	WT	c.35G>A	61.3	WT	WT
40	WT	c.35G>A	37.8	WT	WT
41	WT	c.35G>A	42.1	WT	WT
43	WT	c.35G>A	54.1	WT	WT
44	WT	c.35G>A	57.4	WT	WT
45	WT	c.35G>A	34.2	WT	WT
46	WT	c.35G>A	40.3	WT	WT
47	WT	c.35G>A	52.1	WT	WT
48	WT	c.35G>A	57.0	WT	WT
49	WT	c.35G>A	48.1	WT	WT
50	WT	c.35G>A	47.7	WT	WT
51	WT	c.35G>A	28.7	WT	WT
52	WT	c.35G>A	50.7	WT	WT
53	WT	c.35G>A	31.4	WT	WT
54	WT	c.35G>A	30.1	WT	WT
55	WT	c.35G>A	55.0	WT	WT
56	WT	c.35G>A	46.6	WT	WT
57	WT	c.35G>A	46.6	WT	WT
58	WT	c.35G>A	24.8	WT	WT
59	WT	c.35G>A	37.1	WT	WT
60	WT	c.35G>A	36.8	WT	WT
61	WT	c.35G>A	58.8	WT	WT
62	WT	c.35G>A	44.0	WT	WT
64	WT	c.35G>A	46.3	WT	WT

65	WT	c.35G>A	42.7	WT	WT
66	WT	c.35G>A	45.0	WT	WT
67	WT	c.35G>A	43.1	WT	WT
68	WT	c.35G>A	49.9	WT	WT
69	WT	c.35G>A	47.4	WT	WT
70	WT	c.35G>A	48.9	WT	WT
71	WT	c.35G>A	55.6	WT	WT
73	WT	c.35G>A	31.8	WT	WT
74	WT	c.35G>A	23.8	WT	WT
75	WT	c.35G>A	32.0	WT	WT
76	WT	c.35G>A	38.5	WT	WT
77	WT	c.35G>A	20.9	WT	WT
78	WT	c.35G>A	28.9	WT	WT
79	WT	c.35G>A	27.3	WT	WT
80	WT	c.35G>A	26.7	WT	WT
81	WT	c.35G>A	24.1	WT	WT
82	WT	WT	WT	WT	WT
83	WT	c.35G>A	28.8	WT	WT

Table 36: Table showing pyrosequencing results from GD3

Sample no.	*BRAF*	Variant allele frequency	*KRAS*	*NRAS*	*PIK3CA*
6	c.1799T>A	24.6	WT	WT	WT
7	c.1799T>A	16.0	WT	WT	WT
8	c.1799T>A	24.3	WT	WT	WT
9	c.1799T>A	26.2	WT	WT	WT
10	c.1799T>A	22.8	WT	WT	WT
11	c.1799T>A	31.4	WT	WT	WT
12	c.1799T>A	29.3	WT	WT	WT
13	c.1799T>A	21.9	WT	WT	WT
14	c.1799T>A	30.5	WT	WT	WT
15	c.1799T>A	29.1	WT	WT	WT
16	c.1799T>A	28.8	WT	WT	WT
17	c.1799T>A	23.5	WT	WT	WT
18	c.1799T>A	33.6	WT	WT	WT

19	c.1799T>A	29.7	WT	WT	WT
20	c.1799T>A	23.9	WT	WT	WT
21	c.1799T>A	20.6	WT	WT	WT
22	c.1799T>A	28.2	WT	WT	WT
23	c.1799T>A	10.7	WT	WT	WT
24	c.1799T>A	10.2	WT	WT	WT
25	Did not amplify				
A17	c.1799T>A	24.8	WT	WT	WT
A21	c.1799T>A	25.5	WT	WT	WT
A24	c.1799T>A	28.0	WT	WT	WT
A25	c.1799T>A	26.5	WT	WT	WT
A26	c.1799T>A	26.2	WT	WT	WT
A29	c.1799T>A	26.7	WT	WT	WT
A30	c.1799T>A	16.3	WT	WT	WT
A31	c.1799T>A	25.0	WT	WT	WT
A32	c.1799T>A	30.3	WT	WT	WT

Table 37:Table showing pyrosequencing results from GD4

Sample no.	*BRAF*	*KRAS*	Variant allele frequency	*NRAS*	*PIK3CA*
1	WT	c.38G>A	8.0	WT	WT
4	WT	c.38G>A	19.8	WT	WT
6	WT	c.38G>A	26.4	WT	WT
8	WT	c.38G>A	23.9	WT	WT
10	WT	c.38G>A	26.4	WT	WT
12	WT	c.38G>A	12.4	WT	WT
14	WT	c.38G>A	26.4	WT	WT
16	WT	c.38G>A	20.2	WT	WT
18	WT	c.38G>A	3.9	WT	WT
26	WT	c.38G>A	35.4	WT	WT
30	WT	c.38G>A	27.0	WT	WT
32	WT	c.38G>A	23.8	WT	WT
34	WT	c.38G>A	13.5	WT	WT
36	WT	c.38G>A	33.3	WT	WT

38	WT	c.38G>A	21.2	WT	WT
52	WT	c.38G>A	14.0	WT	WT
54	WT	c.38G>A	28.5	WT	WT
56	WT	c.38G>A	33.6	WT	WT
62	WT	c.38G>A	30.8	WT	WT
64	WT	c.38G>A	26.8	WT	WT
66	WT	c.38G>A	24.0	WT	WT
68	WT	c.38G>A	24.9	WT	WT
70	WT	c.38G>A	27.1	WT	WT
72	WT	c.38G>A	25.3	WT	WT
74	WT	c.38G>A	32.6	WT	WT
76	WT	c.38G>A	32.5	WT	WT
78	WT	c.38G>A	27.1	WT	WT
80	WT	c.38G>A	28.0	WT	WT
82	WT	c.38G>A	26.2	WT	WT
84	WT	c.38G>A	29.1	WT	WT
86	WT	c.38G>A	22.5	WT	WT
88	WT	c.38G>A	35.8	WT	WT
90	WT	c.38G>A	18.8	WT	WT
92	WT	c.38G>A	20.2	WT	WT
94	WT	c.38G>A	15.7	WT	WT
96	WT	c.38G>A	29.1	WT	WT
98	WT	c.38G>A	25.4	WT	WT
A6	WT	WT	WT	WT	WT
A7	WT	c.38G>A	23.6	WT	WT
A8	WT	c.38G>A	21.1	WT	WT
A9	WT	c.38G>A	10.8	WT	WT
A10	WT	c.38G>A	15.1	WT	WT
A11	WT	c.38G>A	21.0	WT	WT
A12	WT	c.38G>A	24.6	WT	WT
A13	WT	c.38G>A	18.3	WT	WT
A14	WT	c.38G>A	18.4	WT	WT
A15	WT	c.38G>A	25.9	WT	WT
A17	WT	c.38G>A	19.3	WT	WT

A18	WT	c.38G>A	16.5	WT	WT
A19a	WT	c.38G>A	9.4	WT	WT
A19b	WT	c.38G>A	15.7	WT	WT

Table 38: Table showing pyrosequencing results from GD5

Sample no.	*BRAF*	*KRAS*	Variant allele frequency	*NRAS*	*PIK3CA*
2	WT	c.437C>T	28.1	WT	WT
4	WT	c.437C>T	20.1	WT	WT
6	WT	c.437C>T	30.0	WT	WT
8	WT	c.437C>T	34.9	WT	WT
10	WT	c.437C>T	31.5	WT	WT
12	WT	c.437C>T	32.9	WT	WT
14	WT	c.437C>T	36.0	WT	WT
16	WT	c.437C>T	32.9	WT	WT
18	WT	c.437C>T	17.2	WT	WT
20	WT	c.437C>T	30.3	WT	WT
22	WT	c.437C>T	42.8	WT	WT
24	WT	c.437C>T	23.1	WT	WT
26	WT	c.437C>T	35.6	WT	WT
28	WT	c.437C>T	22.4	WT	WT
30	WT	c.437C>T	19.6	WT	WT
32	WT	c.437C>T	23.8	WT	WT
34	WT	c.437C>T	22.7	WT	WT
36	WT	c.437C>T	22.9	WT	WT
38	WT	c.437C>T	20.3	WT	WT
40	WT	c.437C>T	23.1	WT	WT
42	WT	c.437C>T	30.7	WT	WT
44	WT	c.437C>T	19.0	WT	WT
46	WT	c.437C>T	26.2	WT	WT
47	WT	c.437C>T	12.9	WT	WT
50	WT	c.437C>T	26.3	WT	WT
52	WT	c.437C>T	17.2	WT	WT
54	WT	c.437C>T	19.2	WT	WT
56	WT	c.437C>T	26.9	WT	WT

58	WT	c.437C>T	25.6	WT	WT

Table 39: Table showing pyrosequencing results from GD6

Sample no.	*BRAF*	*KRAS*	*NRAS*	Variant allele frequency	*PIK3CA*
B1	WT	WT	c.182A>T	27.1	WT
B2	WT	WT	c.182A>T	32.9	WT
D	WT	WT	c.182A>T	37.7	WT
F	WT	WT	c.182A>T	37.4	WT
H	WT	WT	c.182A>T	25.2	WT
J	WT	WT	c.182A>T	25.4	WT
N	WT	WT	c.182A>T	29.6	WT
P	WT	WT	c.182A>T	35.3	WT
R1	WT	WT	c.182A>T	24.7	WT
R2	WT	WT	c.182A>T	28.6	WT
T1	WT	WT	c.182A>T	27.6	WT
V	WT	WT	c.182A>T	28.3	WT
X	WT	WT	c.182A>T	36.5	WT
Z	WT	WT	c.182A>T	39.9	WT
BB	WT	WT	c.182A>T	40.9	WT
DD	WT	WT	c.182A>T	46.0	WT
FF	WT	WT	c.182A>T	22.2	WT
HH	WT	WT	c.182A>T	39.3	WT
JJ	WT	WT	c.182A>T	41.3	WT
LL	WT	WT	c.182A>T	47.3	WT
NN	WT	WT	c.182A>T	49.9	WT
PP	WT	WT	c.182A>T	44.1	WT
SS	WT	WT	c.182A>T	38.8	WT
UU	WT	WT	c.182A>T	50.5	WT
WW	WT	WT	c.182A>T	43.9	WT
YY	WT	WT	c.182A>T	43.6	WT
AB	WT	WT	c.182A>T	43.1	WT
AD	WT	WT	c.182A>T	44.5	WT
AF	WT	WT	c.182A>T	40.5	WT
AH	WT	WT	c.182A>T	40.9	WT

AJ	WT	WT	c.182A>T	49.9	WT
AL	WT	WT	c.182A>T	48.7	WT

Table 40: Table showing pyrosequencing results from GD7

Sample no.	*BRAF*	*KRAS*	Variant allele frequency	*NRAS*	*PIK3CA*
1	WT	c.38G>A	19.5	WT	WT
3	WT	c.38G>A	27.4	WT	WT
5	WT	c.38G>A	30.1	WT	WT
7	WT	c.38G>A	34.3	WT	WT
9	WT	c.38G>A	27.1	WT	WT
11	WT	c.38G>A	26.9	WT	WT
13	WT	c.38G>A	23.4	WT	WT
15	WT	c.38G>A	20.4	WT	WT
17	WT	c.38G>A	28.5	WT	WT
19	WT	c.38G>A	34.2	WT	WT
21	WT	c.38G>A	29.7	WT	WT
23	WT	c.38G>A	23.9	WT	WT
25	WT	c.38G>A	24.9	WT	WT
29	WT	c.38G>A	21.9	WT	WT
31	WT	c.38G>A	30.9	WT	WT
33	WT	c.38G>A	21.5	WT	WT
35	WT	c.38G>A	23.8	WT	WT
37	WT	c.38G>A	20.6	WT	WT
39	WT	c.38G>A	33.5	WT	WT
41	WT	c.38G>A	26.9	WT	WT
53	WT	c.38G>A	24.2	WT	WT
55	WT	c.38G>A	22.1	WT	WT
57	WT	c.38G>A	16.0	WT	WT
59	WT	c.38G>A	20.5	WT	WT
61	WT	c.38G>A	33.9	WT	WT
63	WT	c.38G>A	44.5	WT	WT
65	WT	c.38G>A	36.8	WT	WT
67	WT	c.38G>A	38.6	WT	WT
69	WT	c.38G>A	36.6	WT	WT

71	WT	c.38G>A	35.0	WT	WT
73	WT	c.38G>A	32.5	WT	WT
75	WT	c.38G>A	35.3	WT	WT
77	WT	c.38G>A	38.8	WT	WT
79	WT	c.38G>A	38.4	WT	WT
81	WT	c.38G>A	33.7	WT	WT
83i	WT	c.38G>A	40.7	WT	WT
83ii	WT	c.38G>A	29.4	WT	WT
88i	WT	c.38G>A	39.8	WT	WT
88ii	WT	c.38G>A	38.0	WT	WT
85	WT	c.38G>A	34.6	WT	WT
89	WT	c.38G>A	35.5	WT	WT
91	WT	c.38G>A	29.4	WT	WT
93	WT	c.38G>A	28.0	WT	WT
95	WT	c.38G>A	30.9	WT	WT
97	WT	c.38G>A	36.3	WT	WT
99	WT	c.38G>A	35.4	WT	WT
101	WT	c.38G>A	36.9	WT	WT
103	WT	c.38G>A	33.1	WT	WT
105	WT	c.38G>A	57.6	WT	WT
107	WT	c.38G>A	42.8	WT	WT

Table 41: Table showing pyrosequencing results from GD8

Sample no.	*BRAF*	*KRAS*	Variant allele frequency	*NRAS*	*PIK3CA*
1	WT	c.35G>A	14.8	WT	WT
3	WT	c.35G>A	71.1	WT	WT
5	WT	c.35G>A	64.0	WT	WT
7	WT	c.35G>A	47.5	WT	WT
9	WT	c.35G>A	43.7	WT	WT
11	WT	c.35G>A	32.5	WT	WT
13	WT	c.35G>A	49.6	WT	WT
15	WT	c.35G>A	34.4	WT	WT
17	WT	c.35G>A	48.6	WT	WT

19	WT	c.35G>A	53.2	WT	WT
21	WT	c.35G>A	40.7	WT	WT
25	WT	c.35G>A	40.7	WT	WT
27	WT	c.35G>A	51.1	WT	WT
29	WT	c.35G>A	49.8	WT	WT
31	WT	c.35G>A	44.1	WT	WT
33	WT	c.35G>A	45.5	WT	WT
35	WT	c.35G>A	57.5	WT	WT
37	WT	c.35G>A	43.9	WT	WT
39	WT	c.35G>A	54.1	WT	WT
41	WT	c.35G>A	47.8	WT	WT
43	WT	c.35G>A	41.6	WT	WT
46	WT	c.35G>A	37.6	WT	WT
48	WT	c.35G>A	48.7	WT	WT
49	WT	c.35G>A	75.3	WT	WT
51	WT	c.35G>A	52.3	WT	WT
53	WT	c.35G>A	56.5	WT	WT
55	WT	c.35G>A	49.1	WT	WT
57	WT	c.35G>A	60.3	WT	WT
59	WT	c.35G>A	45.4	WT	WT
61	WT	c.35G>A	47.2	WT	WT
63	WT	c.35G>A	36.0	WT	WT
65	WT	c.35G>A	50.5	WT	WT
67	WT	c.35G>A	37.2	WT	WT
69	WT	c.35G>A	44.3	WT	WT
71	WT	c.35G>A	40.3	WT	WT
73	WT	c.35G>A	37.4	WT	WT
75	WT	c.35G>A	51.0	WT	WT
77	WT	c.35G>A	43.9	WT	WT
79	WT	c.35G>A	50.5	WT	WT
81	WT	c.35G>A	61.6	WT	WT
83	WT	c.35G>A	46.1	WT	WT
85	WT	c.35G>A	41.7	WT	WT
87	WT	c.35G>A	39.1	WT	WT

89	WT	c.35G>A	38.6	WT	WT
91	WT	c.35G>A	28.7	WT	WT
93	WT	c.35G>A	39.1	WT	WT
95	WT	c.35G>A	34.4	WT	WT
97	WT	c.35G>A	36.6	WT	WT
99	WT	c.35G>A	46.4	WT	WT
101	WT	c.35G>A	34.5	WT	WT
103	WT	c.35G>A	24.5	WT	WT
105	WT	c.35G>A	27.6	WT	WT
107	WT	c.35G>A	29.7	WT	WT
109	WT	c.35G>A	37.1	WT	WT
111	WT	c.35G>A	27.6	WT	WT
113	WT	c.35G>A	44.1	WT	WT
115	WT	c.35G>A	38.1	WT	WT
117	WT	c.35G>A	24.0	WT	WT
119	WT	c.35G>A	29.6	WT	WT
121	WT	c.35G>A	33.6	WT	WT
A6	WT	c.35G>A	27.1	WT	WT
A7	WT	c.35G>A	26.4	WT	WT
A8	WT	c.35G>A	26.0	WT	WT
A14i	WT	c.35G>A	26.5	WT	WT
A14ii	WT	c.35G>A	8.3	WT	WT
A14iii	WT	c.35G>A	<5	WT	WT
A15	WT	c.35G>A	10.0	WT	WT
A17	WT	c.35G>A	11.9	WT	WT
A18	WT	c.35G>A	10.8	WT	WT
A19	WT	c.35G>A	9.1	WT	WT
A20	WT	WT	WT	WT	WT

14 Appendix 4: Target capture gene panel

ABL1	ABL2	ACVR2A	ADAM29	AKAP9	AKT1	AKT2	AKT3	ALK
AMER1	APC	AR	ARAF	ARFRP1	ARID1A	ARID1B	ARID2	ARID5B
ASXL1	ATM	ATR	ATRX	AURKA	AURKB	AXIN1	AXIN2	AXL
BAG4	BAP1	BCL11A	BCL2	BCL2A1	BCL2L1	BCL2L2	BCL6	BCOR
BIRC2	BIRC7	BLM	BPTF	BRAF	BRCA1	BRCA2	BRIP1	BUB1B
C11orf30	CARD11	CASP8	CBFB	CBL	CCND1	CCND2	CCND3	CCNE1
CDC6	CDC73	CDH1	CDK12	CDK4	CDK6	CDK8	CDKN1A	CDKN1B
CDKN2A	CDKN2B	CDKN2C	CEBPA	CHD1	CHD1L	CHD8	CHEK1	CHEK2
CHUK	CIC	CKS1B	COL2A1	CREBBP	CRKL	CTCF	CTNNB1	CTNND1
DAXX	DCUN1D1	DDR2	DDX11	DDX3X	DDX5	DICER1	DNMT3A	DYRK1B
E2F3	EEF1A2	EGFR	EIF5A2	ELK3	EP300	EP400	EPHA3	EPHB1
ERBB2	ERBB3	ERBB4	ESR1	ETV1	EZH2	FADD	FANCA	FANCC
FANCD2	FANCE	FANCF	FANCG	FAS	FBXO11	FBXW7	FGFR1	FGFR2
FGFR3	FGFR4	FH	FLT1	FLT3	FLT4	FOXA1	FOXA2	FOXL2
FOXO1	FOXP1	GAB2	GATA1	GATA2	GATA3	GATA6	GNA11	GNAQ
GNAS	GPC5	GRB2	GRB7	GRIN2A	H3F3A	H3F3B	HIST1H3B	HMGA2
HNF1A	HRAS	HSPA8	ID1	IDH1	IDH2	IGF1R	IGF2R	IKBKE
IL7R	INHBA	IRS2	JAK1	JAK2	JAK3	JUN	KDM5A	KDM5C
KDM6A	KDR	KEAP1	KIT	KLF6	KRAS	MAFA	MAP2K1	MAP2K2
MAP2K4	MAP2K5	MAP2K6	MAP2K7	MAP3K1	MAP3K11	MAP3K12	MAP3K13	MAP3K14
MAP3K2	MAP3K3	MAP3K4	MAP3K5	MAP3K6	MAP3K7	MAP3K8	MAP3K9	MAP4K3
MAP4K4	MAPK1	MAPK10	MAPK7	MAPK8	MAPK9	MCL1	MDM2	MDM4
MED12	MED29	MEN1	MET	MITF	MLH1	MLL	MLL2	MLL3
MLLT3	MPL	MRAS	MRE11A	MSH2	MSH6	MSN	MTDH	MTOR
MUTYH	MYB	MYC	MYCL1	MYCN	MYD88	MYOC	NBN	NBPF10
NCOA2	NCOA3	NCOR1	NCOR2	NF1	NF2	NFE2L2	NGFR	NKX2-1
NOTCH1	NOTCH2	NOTCH3	NOTCH4	NOV	NPM1	NRAS	NTRK1	NTRK2
NTRK3	PAK1	PAK3	PALB2	PAX5	PAX9	PBRM1	PDGFRA	PDGFRB
PHGDH	PIK3C2A	PIK3C2B	PIK3C2G	PIK3C3	PIK3CA	PIK3CB	PIK3CG	PIK3R1
PIK3R2	PIM1	PLCG1	POU1F1	PPM1D	PPP2R1A	PRDM1	PREX2	PRKAR1A
PRKCG	PRKCI	PTCH1	PTEN	PTK6	PTP4A1	PTP4A3	PTPN11	PTPRB
RAB23	RAB25	RAC1	RAD21	RAD50	RAD51	RAF1	RARA	RB1
REG4	RET	RHEB	RHOA	RICTOR	RNF213	RNF43	ROBO1	ROBO2
ROS1	RPS6KB1	RPTOR	RRM2B	RSPO2	RSPO3	RUNX1	SETD2	SF3B1
SHC1	SKP2	SLIT2	SMAD2	SMAD3	SMAD4	SMARCA4	SMARCB1	SMC1A
SMC3	SMO	SMURF1	SOCS1	SOX10	SOX2	SOX9	SPOP	SRC
SRSF2	STAG2	STAT3	STAT4	STK11	SUFU	TBX22	TBX3	TCF7L2
TERT	TET2	TGFBR2	TNFAIP3	TOP1	TP53	TP73	TRAF2	TSC1
TSC2	TSHR	U2AF1	USP9X	VEGFA	VHL	WHSC1	WHSC1L1	WSB1
WT1	XPA	XPC	XPO1	YAP1	YWHAB	YWHAQ	YWHAZ	ZBTB10
ZFHX3	ZNF217	ZNF639	ZNF704	ZRSR2				

16 Appendix 5: mpileup ‘driver’ loci

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
1	115256529	NRAS	Q61L	2	25463169	DNMT3A	?
1	16202812	SPEN	R174*	2	25467150	DNMT3A	Q573_A575delQAA
1	16255256	SPEN	E841*	2	25470903	DNMT3A	?
1	16255882	SPEN	I1052fs*7	2	47643501	MSH2	Q337*
1	16256282	SPEN	E1183*	2	47690201	MSH2	S473*
1	16256320	SPEN	D1198fs*4	2	70315174	PCBP1	L100Q
1	16256666	SPEN	D1313_S1314ins*	2	86847539	RNF103	E94K
1	16256965	SPEN	E1412fs*5	2	148683685	ACVR2A	K437fs*5
1	16257359	SPEN	R1542*	2	148683731	ACVR2A	?
1	16257743	SPEN	E1670*	2	198265007	SF3B1	R957Q
1	16259452	SPEN	Y2239*	2	198266834	SF3B1	K700E
1	16259935	SPEN	L2402fs*13	2	198267484	SF3B1	R625C
1	16262679	SPEN	A3318fs*30	2	198270162	SF3B1	R425Q
1	27023430	ARID1A	G180fs*49	2	202137653	CASP8	E254*
1	27023615	ARID1A	S241fs*1	2	202149751	CASP8	Q398*
1	27024001	ARID1A	Q372fs*28	2	202150003	CASP8	Q482*
1	27057835	ARID1A	Q515*	2	202151270	CASP8	Q524*
1	27057894	ARID1A	Y534*	2	209113112	IDH1	R132H
1	27059206	ARID1A	S617fs*2	2	209113113	IDH1	R132C
1	27087360	ARID1A	S645*	2	219000488	CXCR2	R322C
1	27087892	ARID1A	P728fs*9	3	12645699	RAF1	S257L
1	27088658	ARID1A	Q758fs*59	3	30691806	TGFBR2	K130fs*19
1	27092809	ARID1A	Q944*	3	37090394	MLH1	?
1	27094320	ARID1A	K1010*	3	37090896	MLH1	?
1	27094454	ARID1A	Y1055fs*50	3	41265604	CTNNB1	?
1	27097621	ARID1A	K1072fs*21	3	41266027	CTNNB1	?
1	27099975	ARID1A	Y1285fs*2	3	41266113	CTNNB1	S37F
1	27100181	ARID1A	Q1334delQ	3	41266124	CTNNB1	T41A
1	27105738	ARID1A	N1784fs*13	3	47061252	SETD2	E2477*
1	27105930	ARID1A	D1850fs*4	3	47084094	SETD2	R2399*
1	27106354	ARID1A	R1989*	3	47125824	SETD2	E1816*
1	27106504	ARID1A	Q2039*	3	47129689	SETD2	S1730fs*1
1	27106617	ARID1A	Y2076*	3	47143008	SETD2	T1652N
1	27106921	ARID1A	D2178fs*22	3	47147552	SETD2	R1592*
1	27107082	ARID1A	R2232fs*33	3	47155365	SETD2	S1572fs*1
1	45797951	MUTYH	R274W	3	47158225	SETD2	R1492*
1	115256530	NRAS	Q61K	3	47161957	SETD2	S1390*
1	115258747	NRAS	G12D	3	47162506	SETD2	S1206fs*29
1	115258748	NRAS	G12C	3	47162636	SETD2	Q1164*
1	120458084	NOTCH2	E2420fs*3	3	47164118	SETD2	E670*
1	120458396	NOTCH2	Q2317*	3	47164577	SETD2	E517*
1	120458477	NOTCH2	E2290*	3	49412898	RHOA	Y42C
1	120458549	NOTCH2	E2266*	3	49412973	RHOA	G17E
1	120466609	NOTCH2	?	3	52439874	BAP1	Q280*
1	120547967	NOTCH2	Q134*	3	52685791	PBRM1	P227fs*2
1	201981291	ELF3	Q124*	3	71021785	FOXP1	R525*
1	226252155	H3F3A	G35W	3	71026113	FOXP1	R503fs*24
1	226252162	H3F3A	K37M	3	71064695	FOXP1	?
2	25458649	DNMT3A	Q842*	3	71102914	FOXP1	S98*

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
3	105439049	CBLB	G417fs*5	4	106180785	TET2	C1271W
3	142215210	ATR	W1964*	4	106190819	TET2	R1366H
3	142217556	ATR	R1814fs*10	4	153244092	FBXW7	R689W
3	142280229	ATR	M402fs*37	4	153245446	FBXW7	S582L
3	176752059	TBL1XR1	E393*	4	153247289	FBXW7	R505C
3	178916725	PIK3CA	R38C	4	153247367	FBXW7	R479*
3	178916854	PIK3CA	E81K	4	153249360	FBXW7	R473fs*25
3	178916876	PIK3CA	R88Q	4	153249384	FBXW7	R465H
3	178916891	PIK3CA	R93P	4	153249385	FBXW7	R465C
3	178916919	PIK3CA	E103_P104delEP	4	153249504	FBXW7	W425*
3	178916923	PIK3CA	P104_G106delinsR	4	153259052	FBXW7	E255*
3	178916924	PIK3CA	P104L	4	153268084	FBXW7	Q242*
3	178916928	PIK3CA	G106_R108delGNR	4	153332811	FBXW7	E49*
3	178916929	PIK3CA	G106R	4	153332919	FBXW7	R13*
3	178916930	PIK3CA	G106V	5	13753356	DNAH5	R3620*
3	178916943	PIK3CA	K111_I112delKI	5	33637760	ADAMTS12	R604W
3	178916945	PIK3CA	K111_L113delinsN	5	56111421	MAP3K1	R8fs*16
3	178916946	PIK3CA	K111N	5	56111785	MAP3K1	A130fs*49
3	178916949	PIK3CA	L113_N114delLN	5	56155714	MAP3K1	R270fs*30
3	178916950	PIK3CA	L113_N114delLN	5	56155720	MAP3K1	R273fs*27
3	178917478	PIK3CA	G118D	5	56160616	MAP3K1	P298fs*14
3	178921548	PIK3CA	V344M	5	56160636	MAP3K1	R307fs*8
3	178921553	PIK3CA	N345K	5	56160650	MAP3K1	N309fs*6
3	178922324	PIK3CA	E365K	5	56160663	MAP3K1	R313fs*13
3	178922328	PIK3CA	P366R	5	56160674	MAP3K1	L318fs*4
3	178927980	PIK3CA	C420R	5	56160695	MAP3K1	P324fs*2
3	178928070	PIK3CA	H450_D454delHGLED	5	56160748	MAP3K1	F341fs*44
3	178928073	PIK3CA	G451_L456delinsV	5	56161195	MAP3K1	H357fs*4
3	178928079	PIK3CA	E453K	5	56161198	MAP3K1	I356fs*38
3	178928080	PIK3CA	E453V	5	56161230	MAP3K1	Q367*
3	178936082	PIK3CA	E542K	5	56167737	MAP3K1	I435fs*4
3	178936091	PIK3CA	E545K	5	56167803	MAP3K1	T457fs*4
3	178936092	PIK3CA	E545A	5	56167838	MAP3K1	H469fs*12
3	178936093	PIK3CA	E545D	5	56168471	MAP3K1	C479fs*5
3	178936094	PIK3CA	Q546K	5	56170936	MAP3K1	S592fs*67
3	178936095	PIK3CA	Q546P	5	56171035	MAP3K1	S622fs*21
3	178937422	PIK3CA	C604R	5	56171039	MAP3K1	Q624fs*32
3	178938934	PIK3CA	E726K	5	56171066	MAP3K1	V632fs*19
3	178941923	PIK3CA	L748I	5	56171069	MAP3K1	E633*
3	178951957	PIK3CA	M1004I	5	56171091	MAP3K1	S640*
3	178952064	PIK3CA	M1040T	5	56171123	MAP3K1	Y651fs*1
3	178952072	PIK3CA	M1043V	5	56174872	MAP3K1	Q677fs*6
3	178952074	PIK3CA	M1043I	5	56174927	MAP3K1	S696fs*40
3	178952084	PIK3CA	H1047Y	5	56176975	MAP3K1	N749fs*14
3	178952085	PIK3CA	H1047R	5	56177017	MAP3K1	R763*
3	178952086	PIK3CA	H1047Q	5	56177053	MAP3K1	I776fs*4
3	178952090	PIK3CA	G1049R	5	56177563	MAP3K1	T847fs*10
3	185146495	MAP3K13	E44_L56delEDQQEKGMVRTEL	5	56177654	MAP3K1	L877fs*32
3	185191353	MAP3K13	P746fs*4	5	56177685	MAP3K1	A887fs*19

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
5	56177762	MAP3K1	G914fs*7	5	112170673	APC	S590fs*20
5	56177784	MAP3K1	L920fs*10	5	112170681	APC	W593fs*1
5	56177998	MAP3K1	S992fs*85	5	112170761	APC	Y622fs*7
5	56178257	MAP3K1	N1079fs*2	5	112173452	APC	G721*
5	56178433	MAP3K1	E1137fs*11	5	112173537	APC	L749*
5	56178474	MAP3K1	K1150fs*22	5	112173600	APC	S770*
5	56178528	MAP3K1	K1168fs*18	5	112173704	APC	R805*
5	56178672	MAP3K1	I1216fs*17	5	112173801	APC	S837*
5	56179379	MAP3K1	A1233fs*7	5	112173834	APC	D849fs*11
5	56179438	MAP3K1	G1252fs*21	5	112173836	APC	D849fs*2
5	56179454	MAP3K1	S1256fs*12	5	112173897	APC	P870fs*46
5	56179500	MAP3K1	Q1273fs*11	5	112173917	APC	R876*
5	56180550	MAP3K1	E1294fs*13	5	112174022	APC	E911fs*5
5	56180614	MAP3K1	E1315*	5	112174096	APC	Y935*
5	56180658	MAP3K1	?	5	112174151	APC	L954fs*6
5	56181757	MAP3K1	?	5	112174182	APC	L964*
5	56181844	MAP3K1	H1361fs*20	5	112174253	APC	E988*
5	56181891	MAP3K1	?	5	112174286	APC	Q999*
5	56183234	MAP3K1	Q1382*	5	112174398	APC	N1037fs*19
5	56183241	MAP3K1	R1385fs*37	5	112174574	APC	Q1095*
5	56183327	MAP3K1	I1413fs*7	5	112174631	APC	R1114*
5	56183330	MAP3K1	A1414fs*8	5	112174782	APC	K1165fs*17
5	56183346	MAP3K1	V1420fs*12	5	112174973	APC	Q1228*
5	56189415	MAP3K1	D1483fs*40	5	112174997	APC	Q1237fs*2
5	56189434	MAP3K1	L1491fs*3	5	112175021	APC	Q1244*
5	67522699	PIK3R1	D68fs*38	5	112175130	APC	L1280*
5	67522740	PIK3R1	I82fs*32	5	112175147	APC	E1286*
5	67588154	PIK3R1	Q329fs*15	5	112175167	APC	Q1294fs*6
5	67589168	PIK3R1	R386*	5	112175174	APC	A1296fs*9
5	67589494	PIK3R1	?	5	112175179	APC	D1297fs*8
5	67589568	PIK3R1	V445_H450delVGKKLH	5	112175189	APC	T1301fs*13
5	67589609	PIK3R1	K459delK	5	112175205	APC	E1306fs*2
5	67589622	PIK3R1	D464_Y467delDRLY	5	112175207	APC	E1306*
5	67589664	PIK3R1	?	5	112175211	APC	E1309fs*4
5	67591097	PIK3R1	N564D	5	112175213	APC	K1308*
5	67591106	PIK3R1	K567E	5	112175219	APC	K1310*
5	67591116	PIK3R1	I571_R574delIQLR	5	112175225	APC	G1312*
5	67591144	PIK3R1	?	5	112175239	APC	E1317fs*5
5	112116592	APC	R213*	5	112175255	APC	E1322*
5	112128143	APC	R216*	5	112175260	APC	A1325fs*7
5	112128191	APC	R232*	5	112175273	APC	Q1328*
5	112137081	APC	?	5	112175303	APC	Q1338*
5	112151204	APC	R283*	5	112175358	APC	S1356*
5	112154907	APC	S393*	5	112175376	APC	G1365fs*12
5	112155039	APC	?	5	112175390	APC	Q1367*
5	112162891	APC	R499*	5	112175418	APC	Y1376fs*1
5	112164586	APC	R554*	5	112175437	APC	F1384fs*7
5	112164603	APC	K561fs*9	5	112175452	APC	C1387*
5	112170646	APC	?	5	112175475	APC	E1397fs*1

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
5	112175480	APC	E1397*	6	168281125	MLLT4	F275fs*7
5	112175489	APC	S1400fs*1	6	168297599	MLLT4	Q421*
5	112175496	APC	S1403fs*5	6	168303032	MLLT4	D538fs*2
5	112175507	APC	Q1406*	6	168307939	MLLT4	P581fs*12
5	112175513	APC	E1408*	6	168315877	MLLT4	M769fs*32
5	112175521	APC	S1411fs*4	6	168315898	MLLT4	R776*
5	112175576	APC	T1430fs*44	6	168325728	MLLT4	K1013fs*20
5	112175582	APC	M1431fs*42	6	168347439	MLLT4	Y1131fs*7
5	112175596	APC	S1436fs*1	7	6035259	PMS2	S270*
5	112175599	APC	T1438fs*35	7	55233036	EGFR	P596A
5	112175625	APC	A1446fs*27	7	55241707	EGFR	p.G719S
5	112175639	APC	R1450*	7	55241722	EGFR	G724S
5	112175646	APC	P1453fs*20	7	55242466	EGFR	p.E746_A750del
5	112175668	APC	P1467fs*1	7	55249071	EGFR	T790M
5	112175675	APC	S1465fs*3	7	55259515	EGFR	L858R
5	112175678	APC	R1463*	7	101833132	CUX1	E353*
5	112175723	APC	V1479fs*8	7	101877410	CUX1	W1171*
5	112175745	APC	D1486fs*21	7	140453136	BRAF	V600K
5	112175748	APC	L1488fs*23	7	140453145	BRAF	L597R
5	112175749	APC	T1487fs*27	7	140453155	BRAF	D594N
5	112175751	APC	L1488fs*26	7	140481411	BRAF	G466E
5	112175755	APC	L1489fs*18	7	151845466	MLL3	K4509fs*2
5	112175760	APC	A1492fs*22	7	151845485	MLL3	T4500fs*4
5	112175765	APC	T1493fs*14	7	151845512	MLL3	M4494fs*17
5	112175769	APC	E1494fs*13	7	151845958	MLL3	K4351fs*5
5	112175776	APC	T1496fs*18	7	151845964	MLL3	W4352fs*17
5	112175812	APC	L1509fs*12	7	151846177	MLL3	Q4279*
5	112175876	APC	Q1529*	7	151856024	MLL3	S3865*
5	112175921	APC	E1544*	7	151859568	MLL3	A3700fs*26
5	112175951	APC	T1556fs*3	7	151860212	MLL3	Q3484*
5	112176063	APC	P1594fs*38	7	151860728	MLL3	Q3312*
5	112179128	APC	R2613*	7	151860863	MLL3	I3267fs*59
6	106555250	PRDM1	C789*	7	151864252	MLL3	T3238fs*6
6	131908954	MED23	A1324fs*2	7	151873725	MLL3	R2933fs*21
6	131917741	MED23	R899*	7	151873749	MLL3	S2930*
6	131919439	MED23	L865fs*1	7	151873888	MLL3	R2884*
6	131919845	MED23	N759fs*12	7	151874147	MLL3	K2797fs*26
6	131923421	MED23	A678fs*7	7	151875059	MLL3	P2493fs*22
6	131927766	MED23	?	7	151876918	MLL3	?
6	131931387	MED23	?	7	151878287	MLL3	Q2220*
6	139167714	ECT2L	S268*	7	151878863	MLL3	R2028*
6	152419923	ESR1	Y537C	7	151879253	MLL3	D1896fs*40
6	152419926	ESR1	D538G	7	151879451	MLL3	K1832*
6	157405953	ARID1B	S665fs*27	7	151879521	MLL3	P1809fs*18
6	157502130	ARID1B	E1037*	7	151880090	MLL3	S1745*
6	157505463	ARID1B	Y1130*	7	151884502	MLL3	S1618*
6	157527738	ARID1B	G1804_H1815delGGGDTTEHIQTH	7	151891329	MLL3	L1509fs*3
6	157527960	ARID1B	Q1879fs*77	7	151932901	MLL3	?
6	157527961	ARID1B	Q1879fs*77	7	151932992	MLL3	K893fs*20

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
7	151945226	MLL3	E765*	10	8115814	GATA3	K388fs*112
7	151945631	MLL3	N621fs*6	10	8115850	GATA3	I407fs*106
7	151946974	MLL3	D599fs*12	10	8115851	GATA3	M401fs*107
7	151949719	MLL3	Q461*	10	8115852	GATA3	M401fs*107
7	151949738	MLL3	I455fs*3	10	8115853	GATA3	S402fs*106
7	151949758	MLL3	Q448*	10	8115857	GATA3	S403fs*105
7	151960177	MLL3	T408fs*3	10	8115873	GATA3	P409fs*99
7	151960206	MLL3	G398fs*7	10	8115874	GATA3	P409fs*99
7	152007138	MLL3	R254fs*7	10	8115875	GATA3	P409fs*99
7	152008968	MLL3	D218fs*43	10	8115892	GATA3	H415fs*93
7	152012385	MLL3	S143fs*3	10	8115911	GATA3	T421fs*87
8	68939479	PREX2	R155Q	10	8115919	GATA3	H424fs*84
9	21971035	CDKN2A	D108G	10	8115928	GATA3	S427fs*82
9	21971036	CDKN2A	D108N	10	8115944	GATA3	P433fs*75
9	21971096	CDKN2A	E88*	10	8115946	GATA3	P433fs*75
9	21971111	CDKN2A	H83Y	10	8115952	GATA3	H435fs*73
9	21971120	CDKN2A	R80*	10	8115955	GATA3	S437fs*71
9	21971208	CDKN2A	?	10	8115962	GATA3	M439fs*39
9	98229397	PTCH1	?	10	8115976	GATA3	M443fs*34
9	98232175	PTCH1	L590fs*37	10	8115982	GATA3	*445fs*31
9	135797205	TSC1	?	10	27702350	PTCHD3	T277M
9	139390935	NOTCH1	G2420fs*2	10	89624265	PTEN	R14fs*8
9	139397631	NOTCH1	?	10	89624274	PTEN	Y16*
10	8097752	GATA3	E46fs*143	10	89624275	PTEN	Q17*
10	8106058	GATA3	M294K	10	89624294	PTEN	D24fs*20
10	8111432	GATA3	?	10	89624305	PTEN	Y27D
10	8111433	GATA3	?	10	89653862	PTEN	?
10	8111461	GATA3	C318fs*38	10	89685307	PTEN	Y68N
10	8111472	GATA3	C321fs*32	10	89685314	PTEN	?
10	8111484	GATA3	W329fs*27	10	89692754	PTEN	?
10	8111489	GATA3	T327fs*30	10	89692904	PTEN	R130*
10	8111493	GATA3	W329fs*24	10	89692905	PTEN	R130Q
10	8111496	GATA3	R330fs*23	10	89692911	PTEN	G132D
10	8111497	GATA3	N332fs*26	10	89692920	PTEN	C136fs*12
10	8111499	GATA3	R330fs*22	10	89692980	PTEN	Y155C
10	8111500	GATA3	R331fs*22	10	89692984	PTEN	E157fs*7
10	8111510	GATA3	N334fs*19	10	89711873	PTEN	?
10	8111512	GATA3	N334fs*19	10	89711891	PTEN	S170I
10	8111513	GATA3	D336fs*21	10	89711899	PTEN	R173C
10	8111517	GATA3	D336fs*17	10	89711952	PTEN	V191_S207delVALLFH KMMFETIPMFS
10	8111537	GATA3	L344fs*9	10	89717609	PTEN	?
10	8111549	GATA3	Y346*	10	89717672	PTEN	R233*
10	8111554	GATA3	H349fs*8	10	89717708	PTEN	Q245*
10	8115704	GATA3	N352fs*3	10	89717719	PTEN	C250fs*2
10	8115705	GATA3	N352fs*19	10	89717739	PTEN	E256fs*8
10	8115710	GATA3	L355fs*1	10	89717752	PTEN	Q261fs*37
10	8115741	GATA3	R365fs*87	10	89717777	PTEN	?
10	8115770	GATA3	K375fs*6	10	89720650	PTEN	?
10	8115779	GATA3	C376*	10	89720725	PTEN	S294fs*4

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
10	89720741	PTEN	Q298*	12	25398284	KRAS	G12A
10	89720744	PTEN	E299*	12	25398285	KRAS	G12C
10	89720761	PTEN	C304*	12	40687400	LRRK2	E915*
10	89720778	PTEN	N311fs*7	12	46254668	ARID2	Q1620*
10	89720798	PTEN	T319fs*1	12	49415846	MLL2	R5501*
10	89720799	PTEN	T319fs*0	12	49422631	MLL2	T4787fs*29
10	89720804	PTEN	T319fs*24	12	49422656	MLL2	G4777fs*18
10	89720832	PTEN	N329fs*14	12	49426221	MLL2	L4089fs*17
10	89720849	PTEN	?	12	49432463	MLL2	G2892fs*18
10	89720852	PTEN	R335*	12	49433979	MLL2	G2523fs*129
10	89725036	PTEN	?	12	49436523	MLL2	?
10	104375125	SUFU	Q375*	12	49445040	MLL2	Q809fs*121
10	114710614	TCF7L2	N34fs*2	12	49447018	MLL2	E309fs*34
10	114710675	TCF7L2	N55fs*47	12	50484315	SMARCD1	Q359*
10	114911505	TCF7L2	P325fs*7	12	52380708	ACVR1B	R415*
10	114912149	TCF7L2	R384*	12	52387754	ACVR1B	?
10	114925316	TCF7L2	C463fs*8	12	52387835	ACVR1B	L488fs*16
10	114925728	TCF7L2	*597fs*19	12	56478854	ERBB3	V104L
10	123258034	FGFR2	N550K	12	70932000	PTPRB	R1961*
11	533874	HRAS	Q61L	12	70949004	PTPRB	C1693*
11	534285	HRAS	G13V	12	70954553	PTPRB	E1444*
11	64573730	MEN1	Q344fs*25	12	70983756	PTPRB	Q680*
11	108119660	ATM	F357fs*33	12	70988418	PTPRB	R449*
11	108159804	ATM	L1405fs*9	12	115111972	TBX3	Q590*
11	108164189	ATM	S1589fs*7	12	115112247	TBX3	L498fs*135
11	108170464	ATM	E1677*	12	115112482	TBX3	E420*
11	108175549	ATM	R1882*	12	115112489	TBX3	A418fs*3
11	108178641	ATM	R1898*	12	115114115	TBX3	?
11	108216545	ATM	R2832C	12	115114148	TBX3	A357fs*18
11	118772517	BCL9L	P645fs*26	12	115114166	TBX3	Q351*
11	118772774	BCL9L	P561fs*66	12	115114215	TBX3	T335fs*4
11	118772790	BCL9L	G554fs*2	12	115114232	TBX3	H329fs*10
11	118772999	BCL9L	E485*	12	115115404	TBX3	N308fs*16
11	118773209	BCL9L	R415*	12	115117307	TBX3	?
11	118773302	BCL9L	A384fs*79	12	115118704	TBX3	N212delN
11	118773402	BCL9L	T351fs*4	12	115118710	TBX3	T210delT
11	118778187	BCL9L	?	12	115118733	TBX3	T203fs*24
11	118778296	BCL9L	S143fs*39	12	115118853	TBX3	Y163fs*2
12	12870853	CDKN1B	C29fs*12	12	115118888	TBX3	Y149_L151delins*
12	12871092	CDKN1B	Q107*	13	32907420	BRCA2	I605fs*11
12	12871212	CDKN1B	Q147*	13	32911530	BRCA2	S1013*
12	12871797	CDKN1B	E172*	13	32913464	BRCA2	Q1658*
12	25378562	KRAS	A146T	13	32914239	BRCA2	H1918fs*20
12	25378647	KRAS	K117N	13	32914451	BRCA2	Q1987*
12	25378652	KRAS	N116H	13	32914562	BRCA2	Q2024*
12	25380243	KRAS	Y71_M72insSAMRDQY	13	32914574	BRCA2	E2028*
12	25380275	KRAS	Q61H	13	32929050	BRCA2	Q2354*
12	25398279	KRAS	V14I	13	32929161	BRCA2	M2393fs*19
12	25398281	KRAS	G13D	13	32929199	BRCA2	K2404fs*7

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
13	32937614	BRCA2	G2760fs*13	16	3823754	CREBBP	Q821*
13	32954022	BRCA2	T3033fs*11	16	14026058	ERCC4	R340*
13	32972663	BRCA2	S3338*	16	23625322	PALB2	?
13	48881488	RB1	R73fs*36	16	23641218	PALB2	R753*
13	48881513	RB1	E79*	16	23646654	PALB2	F404fs*7
13	48916759	RB1	E97*	16	67063312	CBFB	M1T
13	48934234	RB1	S230*	16	67063389	CBFB	?
13	48934237	RB1	P232fs*4	16	67063672	CBFB	Q41*
13	48934262	RB1	?	16	67063709	CBFB	S53*
13	48936786	RB1	?	16	67063717	CBFB	?
13	48941639	RB1	S318fs*1	16	67070577	CBFB	P70fs*13
13	48941720	RB1	D346fs*17	16	67070599	CBFB	Q77fs*13
13	48941721	RB1	D346fs*17	16	67070611	CBFB	T80fs*9
13	48942689	RB1	S360fs*2	16	67116115	CBFB	?
13	48947540	RB1	?	16	67645084	CTCF	Q117*
13	48947596	RB1	Q395*	16	67645922	CTCF	H284Y
13	48951083	RB1	K417fs*23	16	67645924	CTCF	H284Q
13	48954320	RB1	F482fs*10	16	67650756	CTCF	S354F
13	48955539	RB1	R552Q	16	67654643	CTCF	R377H
13	48955550	RB1	R556*	16	67670755	CTCF	?
13	49030490	RB1	?	16	68771311	CDH1	?
13	49033844	RB1	R661W	16	68771344	CDH1	S9*
13	49033955	RB1	R698W	16	68772218	CDH1	Q23*
13	49037866	RB1	?	16	68772225	CDH1	E26fs*8
13	49039505	RB1	?	16	68772229	CDH1	C28fs*6
13	49050003	RB1	?	16	68772236	CDH1	H29fs*27
14	38061240	FOXA1	S250F	16	68772259	CDH1	Y37fs*21
14	38061313	FOXA1	D226N	16	68772314	CDH1	?
14	38061461	FOXA1	I176M	16	68835613	CDH1	Y68*
14	69256797	ZFP36L1	L157fs*24	16	68835747	CDH1	V114fs*50
14	69257086	ZFP36L1	K61fs*19	16	68835751	CDH1	T115fs*53
14	105246445	AKT1	L52R	16	68842326	CDH1	?
14	105246532	AKT1	R23Q	16	68842348	CDH1	A137fs*78
14	105246551	AKT1	E17K	16	68842432	CDH1	N166fs*2
15	40477505	BUB1B	M300fs*31	16	68842471	CDH1	?
15	43748582	TP53BP1	E742*	16	68842663	CDH1	P201fs*15
15	43748660	TP53BP1	E716fs*12	16	68842714	CDH1	E218fs*4
15	67457303	SMAD3	R93*	16	68844175	CDH1	Q255*
15	67457637	SMAD3	L153fs*13	16	68844193	CDH1	E261*
15	67482799	SMAD3	P403fs*73	16	68844245	CDH1	?
15	90631837	IDH2	R172S	16	68845600	CDH1	E283fs*11
15	90631838	IDH2	R172K	16	68845658	CDH1	Y302fs*1
15	90631839	IDH2	R172G	16	68845723	CDH1	G324fs*30
16	348179	AXIN1	A443fs*26	16	68846053	CDH1	L343fs*13
16	396589	AXIN1	R146Q	16	68847276	CDH1	D400N
16	3779755	CREBBP	Q1765*	16	68847285	CDH1	N405fs*14
16	3807363	CREBBP	Q1209fs*25	16	68847293	CDH1	T406fs*6
16	3817720	CREBBP	I1084fs*15	16	68849439	CDH1	Q448*
16	3820624	CREBBP	Q943*	16	68849577	CDH1	E494*

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
16	68849628	CDH1	Q511*	17	7577118	TP53	V274F
16	68853186	CDH1	Y523*	17	7577120	TP53	R273P
16	68856049	CDH1	P620fs*11	17	7577121	TP53	R273C
16	68856077	CDH1	L630fs*33	17	7577124	TP53	V272M
16	68857460	CDH1	Q699*	17	7577127	TP53	E271Q
16	68857530	CDH1	?	17	7577130	TP53	F270L
16	68862187	CDH1	G759fs*11	17	7577136	TP53	N268fs*77
16	68862212	CDH1	?	17	7577139	TP53	R267W
16	68867205	CDH1	D819fs*6	17	7577141	TP53	G266E
16	68867284	CDH1	L845fs*15	17	7577151	TP53	G262delG
17	7572963	TP53	K382fs*40	17	7577156	TP53	?
17	7573996	TP53	L344R	17	7577498	TP53	?
17	7573997	TP53	E343fs*3	17	7577500	TP53	S261fs*84
17	7574002	TP53	R342fs*3	17	7577501	TP53	T253fs*3
17	7574003	TP53	R342*	17	7577505	TP53	D259V
17	7574006	TP53	G334_M340delGRERFEM	17	7577507	TP53	E258D
17	7574017	TP53	R335fs*8	17	7577514	TP53	I255delI
17	7574018	TP53	R337C	17	7577517	TP53	I255T
17	7574021	TP53	E336fs*9	17	7577522	TP53	M246fs*92
17	7574026	TP53	G334fs*12	17	7577526	TP53	L252P
17	7574034	TP53	?	17	7577535	TP53	R249T
17	7576827	TP53	?	17	7577538	TP53	R248Q
17	7576851	TP53	?	17	7577539	TP53	R248W
17	7576852	TP53	?	17	7577544	TP53	S241_M246delinsL
17	7576855	TP53	Q331*	17	7577545	TP53	C242_M246>L
17	7576862	TP53	F328fs*17	17	7577547	TP53	G245V
17	7576865	TP53	Y327*	17	7577548	TP53	G245S
17	7576873	TP53	G325*	17	7577550	TP53	G244D
17	7576885	TP53	K321*	17	7577551	TP53	G244S
17	7576891	TP53	K319*	17	7577556	TP53	C242Y
17	7576897	TP53	Q317*	17	7577557	TP53	C242fs*5
17	7577005	TP53	?	17	7577559	TP53	S241F
17	7577018	TP53	?	17	7577560	TP53	S241A
17	7577022	TP53	R306*	17	7577561	TP53	S240R
17	7577049	TP53	H297fs*10	17	7577565	TP53	N239T
17	7577058	TP53	E294*	17	7577567	TP53	C238*
17	7577060	TP53	E294fs*11	17	7577568	TP53	C238Y
17	7577082	TP53	E286K	17	7577570	TP53	M237I
17	7577085	TP53	E285K	17	7577572	TP53	Y236delY
17	7577094	TP53	R282W	17	7577574	TP53	Y236C
17	7577095	TP53	D281_R282delDR	17	7577579	TP53	Y234*
17	7577096	TP53	D281G	17	7577580	TP53	Y234C
17	7577097	TP53	D281H	17	7577581	TP53	Y234N
17	7577099	TP53	R280K	17	7577585	TP53	H233_Y234insH
17	7577105	TP53	P278L	17	7577593	TP53	T230P
17	7577111	TP53	C275_A276delinsS	17	7577609	TP53	?
17	7577113	TP53	C275_A276insFEVRVC	17	7578175	TP53	?
17	7577114	TP53	C275Y	17	7578176	TP53	?
17	7577117	TP53	V274D	17	7578189	TP53	Y220*

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
17	7578190	TP53	Y220C	17	7578454	TP53	A159V
17	7578191	TP53	Y220N	17	7578463	TP53	R156P
17	7578192	TP53	P219_Y220insP	17	7578470	TP53	T150fs*16
17	7578194	TP53	V218_P219insV	17	7578474	TP53	P153fs*28
17	7578198	TP53	H214_V217delinsQ	17	7578479	TP53	P151S
17	7578203	TP53	V216M	17	7578492	TP53	W146*
17	7578208	TP53	H214R	17	7578496	TP53	Q144fs*4
17	7578209	TP53	H214D	17	7578503	TP53	V143M
17	7578211	TP53	R213L	17	7578505	TP53	P142H
17	7578212	TP53	R213*	17	7578506	TP53	P142fs*7
17	7578218	TP53	T211fs*36	17	7578507	TP53	C141W
17	7578221	TP53	R209fs*6	17	7578508	TP53	C141Y
17	7578224	TP53	R209*	17	7578510	TP53	K139_T140delKT
17	7578236	TP53	Y205H	17	7578515	TP53	K139fs*31
17	7578239	TP53	E204*	17	7578516	TP53	K139fs*31
17	7578249	TP53	N200fs*9	17	7578526	TP53	C135Y
17	7578253	TP53	G199V	17	7578528	TP53	M133fs*36
17	7578257	TP53	P191_E198>Q	17	7578532	TP53	M133R
17	7578262	TP53	R196fs*51	17	7578534	TP53	K132N
17	7578263	TP53	R196fs*51	17	7578535	TP53	K132R
17	7578265	TP53	I195T	17	7578546	TP53	P128fs*42
17	7578268	TP53	L194R	17	7578550	TP53	S127F
17	7578271	TP53	H193R	17	7578555	TP53	?
17	7578275	TP53	Q192*	17	7578556	TP53	?
17	7578281	TP53	P190T	17	7579307	TP53	?
17	7578290	TP53	?	17	7579310	TP53	?
17	7578370	TP53	?	17	7579312	TP53	T125T
17	7578371	TP53	G187S	17	7579313	TP53	T125R
17	7578375	TP53	D184fs*62	17	7579320	TP53	V122fs*26
17	7578380	TP53	D184H	17	7579335	TP53	T118fs*5
17	7578382	TP53	S183*	17	7579344	TP53	L111fs*8
17	7578384	TP53	C182W	17	7579350	TP53	F113V
17	7578388	TP53	R181P	17	7579358	TP53	R110fs*13
17	7578389	TP53	R181C	17	7579359	TP53	R110fs*13
17	7578391	TP53	E180fs*68	17	7579373	TP53	G105V
17	7578394	TP53	H179R	17	7579377	TP53	Q104*
17	7578395	TP53	H179Y	17	7579382	TP53	T102fs*21
17	7578398	TP53	H178D	17	7579389	TP53	Q100*
17	7578402	TP53	C176W	17	7579391	TP53	T81fs*49
17	7578406	TP53	R175H	17	7579395	TP53	V97fs*25
17	7578407	TP53	R175G	17	7579397	TP53	W91fs*52
17	7578413	TP53	V173M	17	7579405	TP53	S95fs*54
17	7578416	TP53	V172F	17	7579409	TP53	L93fs*30
17	7578420	TP53	T170fs*4	17	7579410	TP53	L93fs*30
17	7578434	TP53	Q165fs*5	17	7579415	TP53	W91*
17	7578437	TP53	Q165*	17	7579463	TP53	P75fs*48
17	7578440	TP53	K164*	17	7579470	TP53	V73fs*50
17	7578441	TP53	A159fs*8	17	7579486	TP53	E68fs*81
17	7578442	TP53	Y163C	17	7579494	TP53	R65*

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
17	7579525	TP53	F54fs*69	17	37880261	ERBB2	D769Y
17	7579528	TP53	W53*	17	37881000	ERBB2	V777L
17	7579529	TP53	W53*	17	37881002	ERBB2	P780_Y781insGSP
17	7579534	TP53	I50fs*72	17	37881010	ERBB2	P780_Y781insGSP
17	7579536	TP53	E51*	17	39183145	KRTAP1-5	I88T
17	11958295	MAP2K4	Q69*	17	41201134	BRCA1	?
17	11984670	MAP2K4	?	17	41203096	BRCA1	F1772fs*29
17	11984698	MAP2K4	E82*	17	41209130	BRCA1	D1739V
17	11984705	MAP2K4	S84*	17	41234451	BRCA1	R1443*
17	11998942	MAP2K4	L149fs*3	17	41246005	BRCA1	E515*
17	12011144	MAP2K4	S184L	17	41247864	BRCA1	A224fs*4
17	12013748	MAP2K4	?	17	70117593	SOX9	S23fs*38
17	12016583	MAP2K4	S240fs*36	17	70117614	SOX9	E28*
17	12032602	MAP2K4	L346_C347insNL	17	70117715	SOX9	E63fs*189
17	12032605	MAP2K4	?	17	70118940	SOX9	Y172fs*80
17	15942803	NCOR1	S2300*	17	70119023	SOX9	N201fs*51
17	15961268	NCOR1	Q2041*	17	70119068	SOX9	S215fs*4
17	15974791	NCOR1	Q1362*	17	70119716	SOX9	T243fs*10
17	15983280	NCOR1	Q1167*	17	70119758	SOX9	R254*
17	15983798	NCOR1	S1114fs*11	17	70119762	SOX9	R257fs*39
17	15983936	NCOR1	E1095*	17	70119779	SOX9	E261*
17	16004595	NCOR1	E887*	17	70120002	SOX9	W335*
17	16022772	NCOR1	R627fs*14	17	70120234	SOX9	S421fs*57
17	16024380	NCOR1	A599fs*23	17	70120246	SOX9	A419fs*53
17	16029427	NCOR1	E535*	17	70120260	SOX9	F423fs*48
17	16029445	NCOR1	E529fs*7	17	70120308	SOX9	Q439fs*32
17	16029452	NCOR1	E526fs*9	17	70120310	SOX9	D441fs*32
17	16029456	NCOR1	T525fs*11	17	73775146	H3F3B	K37M
17	16046958	NCOR1	E379*	18	45368223	SMAD2	S460*
17	16049835	NCOR1	Q313*	18	45375016	SMAD2	S276*
17	29486049	NF1	N78fs*7	18	45395659	SMAD2	E159*
17	29527439	NF1	?	18	45395732	SMAD2	W134*
17	29533298	NF1	V437fs*3	18	48575071	SMAD4	G89*
17	29546065	NF1	E524*	18	48575671	SMAD4	S144*
17	29550505	NF1	Q589*	18	48584607	SMAD4	Y260*
17	29557277	NF1	?	18	48584611	SMAD4	N263fs*74
17	29562998	NF1	S1312fs*0	18	48591918	SMAD4	R361C
17	29576134	NF1	Y1369*	18	48591919	SMAD4	R361H
17	29586089	NF1	E1458*	18	48603032	SMAD4	R445*
17	29586092	NF1	E1459*	18	48603038	SMAD4	Q450fs*17
17	29652922	NF1	I1641fs*9	18	48603116	SMAD4	G473*
17	29661916	NF1	C1960fs*4	18	48604733	SMAD4	A532fs*7
17	29677227	NF1	R2450*	18	48604749	SMAD4	W524S
17	29677284	NF1	E2469*	18	48604764	SMAD4	H530fs*46
17	29685497	NF1	?	19	1207079	STK11	E57fs*107
17	37868208	ERBB2	S310Y	19	1219401	STK11	C151*
17	37879658	ERBB2	R678Q	19	11118599	SMARCA4	Q675*
17	37880220	ERBB2	L755S	19	11169548	SMARCA4	?
17	37880257	ERBB2	I767M	19	40762959	AKT2	E17K

Chromosome	Position	Gene	Variant	Chromosome	Position	Gene	Variant
19	42793085	CIC	S326*	X	76776395	ATRX	?
19	54646887	CNOT3	E20K	X	123164816	STAG2	G46fs*41
20	31022301	ASXL1	I597fs*106	X	123197716	STAG2	R614*
20	31022380	ASXL1	Q623fs*79	X	133512118	PHF6	R76fs*5
20	31022592	ASXL1	R693*	X	133559304	PHF6	G348*
20	31023702	ASXL1	Q1063*				
20	31024197	ASXL1	E1228*				
20	39795235	PLCG1	R707Q				
20	40770613	PTPRT	W904*				
20	57484420	GNAS	R844C				
20	57484421	GNAS	R844H				
21	36206899	RUNX1	?				
21	36252984	RUNX1	D126fs*12				
21	36252994	RUNX1	D123fs*11				
21	36253012	RUNX1	?				
22	29095873	CHEK2	E364*				
22	29191600	XBP1	P240fs*20				
22	29191658	XBP1	P221fs*43				
22	29192037	XBP1	Q199fs*187				
22	29192046	XBP1	L196fs*190				
22	30035202	NF2	?				
22	30067890	NF2	L360fs*1				
X	39916576	BCOR	?				
X	39933958	BCOR	S214*				
X	40994073	USP9X	W140fs*1				
X	41007699	USP9X	H505fs*3				
X	41043647	USP9X	?				
X	41057947	USP9X	Q1518fs*1				
X	41075178	USP9X	K1788fs*10				
X	41075215	USP9X	R1799*				
X	41075671	USP9X	Y1953fs*2				
X	41077624	USP9X	?				
X	44894228	KDM6A	?				
X	44918338	KDM6A	W321*				
X	44922769	KDM6A	L545fs*8				
X	44929206	KDM6A	S769*				
X	44929499	KDM6A	K867*				
X	44929570	KDM6A	Y890*				
X	44938487	KDM6A	W1012*				
X	44949176	KDM6A	?				
X	63411169	FAM123B	V663fs*53				
X	63411228	FAM123B	Q647fs*6				
X	63411366	FAM123B	R601*				
X	63411514	FAM123B	R549fs*11				
X	63411678	FAM123B	R497*				
X	63412095	FAM123B	R358*				
X	63412110	FAM123B	R353*				
X	63412511	FAM123B	E219fs*63				
X	76776358	ATRX	Q2370*				

17 Appendix 6: Mean sequencing depth for each sample

Table 42: Mean sequencing depth per sample in GD1

Sample name	Mean sequencing depth
GD1_4_met	63.97
GD1_14_met	63.16
GD1_28_met	59.84
GD1_30_met	62.67
GD1_34_met	63.74
GD1_38_tumour	58.29
GD1_40_tumour	59.91
GD1_42_tumour	63.72
GD1_44_met	64.14
GD1_46_met	66.36
GD1_52_met	64.23
GD1_54_met	63.14
GD1_56_met	63.81
GD1_60_met	65.12
GD1_62_met	61.87
GD1_64_met	58.31
GD1_66_met	65.50
GD1_68_met	71.81
GD1_70_met	66.60
GD1_72_met	66.65
GD1_76_normal	28.57

Table 43: Mean sequencing depth per sample in GD3

Sample	Mean sequencing depth
GD3_33_recurrence	68.04
GD3_39_met 1	65.78
GD3_40_met 1	67.45
GD3_43_normal	30.80

Table 44: Mean sequencing depth per sample in GD4

Sample	Mean sequencing depth
GD4_3	61.47
GD4_5	58.43
GD4_7	59.45
GD4_9	61.11
GD4_11	57.90
GD4_13	62.44
GD4_15	59.21
GD4_17	59.80
GD4_19 (normal)	30.86
GD4_27	63.82
GD4_29	52.74
GD4_31	62.43
GD4_33	58.95
GD4_35	58.35
GD4_37	61.25
GD4_39 (normal)	30.87
GD4_53	57.95
GD4_61	59.07
GD4_63	57.68
GD4_65	62.11
GD4_67	65.85
GD4_69	65.50
GD4_71	62.79
GD4_73	55.64
GD4_75	60.70
GD4_77	60.22
GD4_79	57.88
GD4_81	63.95
GD4_83	60.63
GD4_85	59.97
GD4_87	58.43
GD4_89	58.83
GD4_91	60.54
GD4_93	61.19
GD4_95	59.30

Table 45: Mean sequencing depth per sample in GD6

Sample	Mean sequencing depth
GD6_E_sigmoid colon_primary	59.33
GD6_G_sigmoid colon_primary	62.75
GD6_I_sigmoid colon_primary	65.26
GD6_K_sigmoid colon_primary	60.47
GD6_M_sigmoid colon_primary	67.61
GD6_O_sigmoid colon_primary	61.50
GD6_S_sigmoid colon_primary	66.76
GD6_U_sigmoid colon_primary	65.59
GD6_Y_liver_liver_met_2	59.90
GD6_AA_liver_liver_met_3	61.52
GD6_AC_liver_liver_met_18	59.69
GD6_AK_liver_liver_met_13	63.99
GD6_AO_spleen _normal_spleen	32.46
GD6_AQ_spleen _normal_spleen	32.03
GD6_AS_peritoneum_serosal_met	62.93
GD6_CC_liver_liver_met_4	58.12
GD6_EE_liver_liver_met_5	58.35
GD6_GG_liver_liver_met_6	59.64
GD6_II_liver_liver_met_7	78.80
GD6_KK_liver_liver_met_8	69.74
GD6_MM_liver_liver_met_7	60.96
GD6_QQ_liver_liver_met_9	62.91
GD6_RR_liver_liver_met_10	63.44
GD6_TT_liver_liver_met_11	61.34
GD6_VV_liver_liver_met_12	58.49
GD6_W_liver_liver_met_1	59.67
GD6_XX_liver_liver_met_13	58.10
GD6_ZZ_liver_liver_met_17	57.90

Table 46: Mean sequencing depth per sample in GD8

Sample	Mean sequencing depth
GD8_1	282.10
GD8_3	328.87
GD8_5	302.03
GD8_7	418.32
GD8_9	445.56
GD8_11	332.87
GD8_13	426.50
GD8_15	380.97
GD8_17	361.45
GD8_19	392.83
GD8_21	313.25
GD8_25	278.67
GD8_27	307.31
GD8_29	243.28
GD8_31	284.02
GD8_33	276.18
GD8_35	261.08
GD8_37	290.90
GD8_39	319.40
GD8_41	328.37
GD8_43	297.89
GD8_46	300.32
GD8_48	336.47
GD8_49	268.00
GD8_51	259.33
GD8_53	262.86
GD8_55	288.24
GD8_57	250.60
GD8_59	277.35
GD8_61	329.87
GD8_63	293.09
GD8_65	296.55
GD8_67	265.19
GD8_69	265.11
GD8_71	311.01
GD8_73	274.77

GD8_75	307.41
GD8_77	304.39
GD8_79	273.62
GD8_81	304.60
GD8_83	290.24
GD8_85	276.77
GD8_87	256.13
GD8_89	281.15
GD8_91	253.93
GD8_93	335.26
GD8_95	317.97
GD8_97	314.25
GD8_99	296.53
GD8_101	279.15
GD8_103	301.53
GD8_105	320.62
GD8_107	338.04
GD8_109	329.05
GD8_111	318.45
GD8_113	309.65
GD8_115	346.62
GD8_117	327.21
GD8_119	350.37
GD8_121	352.42
GD8_A6	532.53
GD8_A7	696.32
GD8_A8	818.98
GD8_A14i	745.52
GD8_A14ii	717.52
GD8_A14iii	391.09
GD8_A15	693.54
GD8_A17	640.29
GD8_A18	731.53
GD8_A19	679.39
GD8_A20	840.02
GD8_123_normal	249.90
GD8_125_normal	417.84
GD8_127_normal	414.73

GD8_129_normal	384.82
GD8_131_normal	424.50

18 Appendix 7: driver variants with VAFs greater than 5% with more than one supporting read

Table 47: Table showing driver variants from GD1 with VAFs greater than 5% with more than one supporting read

		GD1_4_met_liver	GD1_14_met_liver	GD1_28_met_liver	GD1_30_met_liver	GD1_34_met_liver	GD1_38_tumour_primary	GD1_40_tumour_primary	GD1_42_tumour_primary	GD1_44_met_diaphragm	GD1_46_met_diaphragm	GD1_52_met_diaphragm	GD1_54_met_pancreatic	GD1_56_met_para-aortic	GD1_60_met_localnode	GD1_62_met_local_vascular_embolus	GD1_64_met_localnode	GD1_66_met_localperitoneal	GD1_68_met_localnode	GD1_70_met_localnode	GD1_72_met_localperitoneal
Variant		Variant allele frequency																			
TP53	V272M	44.3	39.9	33.3	28.1	26.9	20.7	19.8	31.5	23.7	45.8	39.1	23.0	28.9	42.9	46.3	37.5	49.3	37.0	26.7	31.0
BRAF	V600K	14.3	30.8	33.3	44.4	14.3	0.0	21.1	26.3	28.6	48.6	10.0	41.2	22.2	33.3	43.1	31.8	39.6	13.3	23.7	20.0
BRCA2	I605fs*11	0.0	5.6	2.1	4.5	0.0	0.0	0.0	7.0	0.0	0.0	0.0	1.2	0.0	1.4	1.2	1.9	1.2	1.5	2.2	0.0
NRAS	Q61K	6.4	0.0	0.0	0.0	0.0	0.0	2.9	2.0	2.8	0.0	3.0	4.8	0.0	0.0	0.0	4.8	1.3	1.3	1.7	0.0
PIK3CA	P104L	0.0	5.1	2.8	0.0	0.0	3.0	0.0	0.0	0.0	2.4	0.0	0.0	2.9	2.0	0.0	0.0	3.4	0.0	1.4	0.0
SMAD2	S276*	0.0	10.3	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	4.3	0.0	0.0	2.6	3.0	2.0	0.0	0.0	14.3
NF1	?	7.7	0.0	4.0	0.0	0.0	8.3	0.0	0.0	0.0	6.3	0.0	0.0	0.0	0.0	0.0	8.5	2.5	0.0	4.3	0.0

ATM	F357fs*33	0.0	0.0	0.0	5.3	0.0	0.0	7.7	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0	1.7	3.7	0.0
PTEN	S294fs*4	0.0	0.0	7.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.7	0.0	6.9	5.3	0.0	0.0	0.0
PREX2	R155Q	6.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	3.6	0.0	6.1	0.0	0.0
APC	R1114*	5.3	0.0	0.0	8.7	0.0	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0
PIK3CA	L113_N114delLN	0.0	0.0	2.8	0.0	0.0	9.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
APC	R499*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0
USP9X	R1799*	6.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 48: Table showing driver variants from GD3 with VAFs greater than 5% with more than one supporting read

		GD3_33_recurrence	GD3_39_met_1
Variant		Variant Allele Frequency	
BRAF	V600K	24.4	26.8
TP53	R249T	0.0	23.6
PIK3CA	M1004I	23.4	0.0

Table 49: Table showing driver variants from GD4 with VAFs greater than 5% with more than one supporting read

		GD4_3_met_Liver_lesion_1	GD4_5_met_Liver_lesion_2	GD4_7_met_Liver_lesion_2	GD4_9_met_Liver_lesion_2	GD4_13_met_Liver_lesion_4	GD4_27_met_Diaphragm_lesion_2	GD4_29_met_Diaphragm_lesion_3	GD4_31_met_Diaphragm_lesion_4	GD4_33_met_Diaphragm_lesion_5	GD4_35_met_Diaphragm_lesion_6	GD4_37_met_Peritoneal_deposit_1_duodenum	GD4_53_met_Para_aortic_node_2	GD4_61_met_Para_aortic_node_6	GD4_63_met_Anterorectal_mass	GD4_65_met_Anterorectal_mass	GD4_67_met_Anterorectal_mass	GD4_69_met_Anterorectal_mass	GD4_71_met_Peritoneal_deposit_2_gall_bladder	GD4_73_met_Peritoneal_deposit_3_bladder_1	GD4_75_met_Peritoneal_deposit_4_bladder_2	GD4_77_met_Right_iliac_fossa_mass	GD4_79_met_Right_iliac_fossa_mass	GD4_81_met_Right_iliac_fossa_mass	GD4_83_met_Right_iliac_fossa_mass	GD4_85_met_Right_iliac_fossa_mass	GD4_87_met_Peritoneal_deposit_5_right_iliac_fossa	GD4_89_met_Mesenteric_mass_node_1	GD4_91_met_Mesenteric_mass_node_2	GD4_93_met_Splenic_flexure_mass	GD4_95_met_Splenic_flexure_mass
Identifier		Variant Allele Frequency																														
ERBB3	V104L	25.4	34.1	30.1	35.0	17.3	20.8	29.8	22.1	15.3	32.3	37.8	26.5	17.3	18.5	22.4	28.2	28.3	29.6	20.5	21.3	33.6	21.1	22.6	41.5	27.4	29.7	22.5	24.1	25.3	35.5
KRAS	G13D	7.5	12.5	20.5	23.6	18.2	26.2	16.7	26.3	6.3	23.6	28.0	33.3	11.9	18.4	12.2	18.2	30.3	26.7	20.3	20.8	17.2	26.0	16.7	21.3	20.5	30.1	21.7	21.6	30.6	18.4
CBFB	M1T	2.2	1.8	2.4	0.9	1.0	2.9	1.7	3.8	2.9	3.7	1.2	2.4	2.0	2.2	1.7	5.0	2.3	3.8	5.6	1.2	2.0	3.3	3.3	3.3	2.7	2.0	3.1	1.1	1.7	4.4
AKT1	L52R	0.9	0.8	0.7	0.0	0.0	0.0	1.4	1.9	2.9	1.7	0.0	0.0	0.0	1.8	0.8	1.4	0.0	0.7	0.0	1.1	0.7	1.0	0.8	0.8	1.9	0.8	4.7	1.0	1.4	0.5
NCOR1	E1095*	3.3	2.4	3.8	3.6	0.0	3.0	0.0	5.0	4.1	5.8	0.0	0.0	0.0	2.6	2.0	2.4	0.0	3.6	6.8	2.7	2.1	1.5	1.4	2.4	1.5	1.8	2.0	0.0	0.0	2.6
PIK3CA	H450_D454delHGLED	2.1	1.4	2.7	2.4	0.0	1.2	1.8	4.3	4.0	1.5	2.7	0.0	2.4	0.0	0.0	1.8	0.0	2.3	2.2	1.4	0.0	0.0	2.0	5.0	1.5	2.9	0.0	8.0	3.6	2.0
ARID1A	S617fs*2	1.9	6.7	0.0	0.0	0.0	0.0	0.0	3.3	0.0	1.6	1.6	3.2	0.0	0.0	0.0	2.4	0.0	0.0	0.0	1.4	0.0	1.5	1.3	2.6	1.6	1.4	0.0	0.0	0.0	0.0
PTEN	Q298*	4.5	2.0	0.0	0.0	5.0	6.7	0.0	0.0	9.4	2.1	3.4	0.0	0.0	0.0	2.3	3.7	0.0	4.5	4.5	0.0	2.7	4.0	3.4	6.9	0.0	2.1	3.8	6.3	2.0	4.3
PIK3CA	L113_N114delLN	2.0	0.0	0.0	0.0	0.0	0.0	2.7	0.0	2.1	0.0	1.8	0.0	0.0	3.1	3.1	2.3	0.0	7.7	4.9	0.0	0.0	2.6	1.8	0.0	3.4	2.2	3.4	0.0	2.1	0.0
CDH1	P620fs*11	1.6	1.3	1.0	1.4	0.0	1.5	0.0	0.9	1.0	1.2	1.3	1.2	2.0	1.3	0.0	1.1	0.9	0.0	1.3	2.7	0.0	1.5	1.0	0.0	1.3	5.1	0.0	0.0	0.0	1.2
USP9X	?	7.0	0.0	0.0	0.0	0.0	3.2	0.0	2.2	10.6	3.1	2.4	2.6	0.0	0.0	3.1	4.8	0.0	2.1	5.6	0.0	5.1	7.3	3.3	0.0	2.1	2.4	0.0	0.0	0.0	0.0
TP53	G187S	0.6	0.0	0.0	1.1	0.0	0.0	0.0	0.0	1.1	1.4	0.0	1.2	1.0	1.1	1.1	0.0	0.0	0.6	1.0	0.0	1.0	0.0	1.1	0.0	1.3	5.5	0.0	0.0	0.0	0.8
MAP3K1	K1168fs*18	4.8	2.3	0.0	1.4	0.0	0.0	2.2	0.0	0.0	0.0	0.0	1.4	0.0	0.0	1.3	0.0	1.4	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	2.2
STK11	E57fs*107	0.7	0.0	0.7	1.0	0.0	2.0	1.0	1.0	0.9	0.0	0.0	1.6	1.2	0.0	0.0	0.0	0.9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	1.6	0.8	0.0

PTEN	C304*	6.5	2.0	0.0	6.1	0.0	1.7	3.1	0.0	0.0	2.0	1.7	0.0	3.8	3.0	0.0	0.0	0.0	0.0	2.2	2.5	2.6	0.0	0.0	3.0	3.1	0.0	0.0	0.0	2.2	2.0
RB1	?	12.5	0.0	2.3	0.0	0.0	0.0	0.0	0.0	1.5	1.4	2.8	1.4	0.0	1.7	0.0	0.0	0.0	0.0	2.5	1.9	0.0	0.0	0.0	1.7	0.0	2.1	0.0	0.0	0.0	0.0
KRAS	K117N	8.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.3	0.0	0.0	0.0	6.3	0.0	3.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0	0.0	0.0	0.0	3.8
PIK3CA	R93P	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	3.6	0.0	0.0	0.0	6.3	2.6	2.0	0.0	2.7	0.0	0.0	0.0	0.0	0.0	0.0	4.7	2.0
MLL3	L1509fs*3	6.1	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	0.0	1.6	0.0	0.0	0.0	2.4	0.0	0.0	0.0	0.0
APC	E1544*	0.0	0.0	6.7	0.0	0.0	1.3	0.0	3.4	2.3	2.2	0.0	2.9	1.8	1.4	1.4	0.0	0.0	0.0	0.0	0.0	1.3	0.0	1.1	2.6	2.3	1.3	1.5	0.0	2.6	5.0
TGFBR2	K130fs*19	0.0	0.0	4.5	2.6	5.6	2.7	0.0	1.6	5.0	8.5	0.0	2.9	0.0	2.0	1.8	0.0	4.4	3.3	0.0	0.0	2.0	3.2	2.7	1.3	2.7	3.4	2.0	0.0	0.0	2.7
APC	S1400fs*1	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0	2.4	1.2	1.3	1.3	0.0	0.0	0.0	0.0	0.0	1.8	1.3	1.1	1.1	1.0	1.5	3.6	2.3	0.0	0.0	4.5
SETD2	E670*	0.0	0.0	0.0	0.0	0.0	2.8	0.0	1.7	2.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	1.7	1.8	1.8	1.5	0.0	0.0	6.8	2.0
STAG2	G46fs*41	0.0	2.0	2.9	0.0	0.0	1.9	0.0	3.2	0.0	0.0	0.0	0.0	0.0	0.0	2.8	0.0	0.0	2.9	2.0	0.0	4.2	0.0	3.0	3.0	0.0	3.0	0.0	0.0	5.9	0.0
PIK3R1	?	0.0	2.7	4.8	3.3	0.0	1.8	0.0	2.1	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	1.9	2.4	0.0	1.6	2.4	0.0	2.7	0.0
SF3B1	K700E	0.0	0.0	5.3	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7	1.7	0.0	2.3	0.0	13.3	0.0	0.0
KDM6A	L545fs*8	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0	1.9	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	2.6	5.1	0.0	0.0	2.2	1.6	0.0	0.0
SMAD2	W134*	0.0	0.0	12.5	4.0	0.0	4.8	0.0	0.0	0.0	3.8	8.3	3.4	0.0	3.6	4.2	3.1	0.0	12.5	3.4	6.3	5.3	3.2	5.7	0.0	3.0	6.1	4.2	5.9	2.0	0.0
SMAD2	E159*	0.0	2.9	0.0	0.0	0.0	6.4	5.3	0.0	11.5	2.6	2.1	0.0	0.0	7.7	0.0	0.0	4.3	0.0	3.3	6.3	26.1	2.4	2.7	0.0	2.7	4.7	0.0	0.0	2.4	10.0
MAP2K4	L149fs*3	0.0	2.5	0.0	2.1	0.0	1.6	0.0	2.6	2.9	2.0	0.0	1.5	0.0	2.6	0.0	2.3	0.0	0.0	2.0	0.0	0.0	0.0	1.8	0.0	2.2	1.8	0.0	8.3	0.0	2.1
APC	A1325fs*7	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	1.5	3.1	0.0	9.1	2.6	1.4	0.0	0.0	1.0	0.0	1.8	1.2	0.0	0.0	0.0	4.2
APC	K1308*	0.0	0.0	0.0	1.2	5.3	1.0	0.0	0.0	0.0	1.2	0.0	2.8	0.0	1.1	0.0	0.0	0.0	0.0	2.1	0.0	0.0	1.1	1.0	0.0	1.2	1.1	1.1	0.0	0.0	0.0
MLL3	S1618*	0.0	0.0	0.0	2.4	4.8	0.0	0.0	2.4	0.0	0.0	1.9	1.6	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0	3.1	1.9	0.0	1.7	0.0	2.1	0.0	2.4	0.0
FBXW7	E49*	0.0	0.0	9.1	0.0	0.0	0.0	0.0	0.0	3.1	1.2	2.6	0.0	0.0	1.7	1.4	0.0	0.0	0.0	3.1	1.3	0.0	1.1	1.2	0.0	1.2	0.0	1.3	0.0	0.0	0.0
APC	Q1529*	0.0	0.0	7.1	2.2	0.0	2.7	0.0	2.2	2.1	4.5	0.0	0.0	1.7	1.7	1.4	0.0	0.0	0.0	0.0	1.4	1.3	0.0	1.1	0.0	2.7	0.0	5.6	0.0	0.0	0.0
RB1	S230*	0.0	0.0	0.0	0.0	0.0	4.4	0.0	4.3	3.0	3.8	0.0	4.5	0.0	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.7	3.8	0.0	2.9	0.0	0.0	8.3	0.0	0.0
USP9X	?	0.0	0.0	0.0	4.3	7.7	5.9	0.0	4.2	4.5	7.9	3.3	3.0	10.0	0.0	4.0	0.0	2.6	0.0	3.4	0.0	3.7	3.2	2.4	0.0	0.0	2.6	0.0	0.0	0.0	7.5
PIK3CA	G451_L456delinsV	0.0	0.0	0.0	4.9	0.0	1.2	3.4	4.2	0.0	1.5	1.4	2.2	2.4	0.0	0.0	3.7	0.0	2.3	4.7	1.4	0.0	4.3	2.1	0.0	0.0	1.5	0.0	0.0	1.8	0.0
APC	Y1376fs*1	0.0	0.0	0.0	4.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	1.4	0.0	2.5	1.2	0.0	3.7	0.0	2.9	0.0	0.0	0.0	1.1	0.0	0.0	0.0	1.1	0.0	3.8	0.0
PTEN	E299*	0.0	0.0	3.2	0.0	4.8	1.6	3.4	0.0	3.0	2.0	1.8	0.0	4.5	2.9	2.2	0.0	0.0	0.0	2.3	2.6	5.4	0.0	0.0	3.4	3.0	4.3	0.0	0.0	0.0	4.3
APC	P870fs*46	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	5.1	0.0	1.8	1.1	0.0	0.0	0.0	0.0	0.0	0.0	2.7	0.0	0.0	1.3	0.0	2.5	1.3	1.5	0.0	0.0	0.0	2.2
NF1	Q589*	0.0	0.0	0.0	0.0	5.6	0.0	0.0	0.0	0.0	0.0	2.7	0.0	0.0	5.9	11.8	0.0	0.0	4.3	0.0	1.7	0.0	4.7	0.0	2.5	2.2	1.8	0.0	0.0	0.0	3.7

PTPRB	C1693*	0.0	0.0	1.9	0.0	0.0	1.5	2.2	1.4	0.0	0.0	0.0	0.0	1.3	1.8	2.7	0.0	0.0	5.4	0.0	0.0	0.0	2.5	0.0	1.3	1.3	1.5	1.7	0.0	1.8	1.9
CTNNB1	?	0.0	0.0	12.5	0.0	0.0	0.0	6.3	1.7	2.3	0.0	0.0	0.0	1.8	1.6	1.7	0.0	0.0	0.0	0.0	1.5	0.0	1.0	0.0	1.3	1.5	1.5	1.3	0.0	0.0	11.1
APC	R1114*	0.0	0.0	0.0	0.0	0.0	0.0	3.6	1.7	0.0	1.7	0.0	0.0	0.0	0.0	2.3	0.0	3.6	4.0	2.5	0.0	0.0	1.5	0.0	4.2	1.9	0.0	1.7	0.0	2.9	0.0
NF1	R2450*	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	1.7	0.0	0.0	4.7	0.0	0.0	0.0	0.0	5.1	0.0	0.0	1.8	0.0	2.1	1.8	0.0	0.0	5.6	0.0	0.0
MAP3K1	Q677fs*6	0.0	4.8	0.0	0.0	0.0	1.4	0.0	4.4	0.0	1.4	0.0	0.0	0.0	1.9	0.0	0.0	0.0	0.0	1.5	1.7	0.0	2.1	0.0	1.9	1.4	0.0	3.0	0.0	0.0	0.0
CDH1	E218fs*4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	1.4	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	1.5	1.3	0.0	0.0	0.0	4.6	0.0
FBXW7	Q242*	0.0	7.1	0.0	6.3	0.0	0.0	0.0	1.3	7.5	1.3	1.4	1.2	3.8	1.5	0.0	0.0	2.6	2.1	1.5	0.0	1.4	2.4	0.0	1.3	1.6	0.0	0.0	0.0	0.0	0.0
APC	P1594fs*38	0.0	0.0	0.0	2.2	7.7	1.5	0.0	1.8	0.0	1.5	2.7	0.0	3.4	3.3	3.2	0.0	0.0	0.0	3.2	1.1	0.0	1.5	0.0	1.1	3.0	0.0	4.8	0.0	0.0	0.0
MAP3K1	R763*	0.0	0.0	0.0	0.0	0.0	1.4	11.1	0.0	0.0	0.0	1.9	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	1.5	1.8	0.0	0.0	3.6	0.0	1.5	1.8	16.7	0.0	3.3
SMAD4	W524S	0.0	0.0	0.0	1.9	0.0	0.0	2.6	0.0	3.0	1.2	1.8	0.0	1.4	0.0	0.0	5.9	0.0	0.0	0.0	1.3	1.0	1.4	0.0	1.0	0.0	4.0	1.1	0.0	0.0	0.0
PTEN	?	0.0	0.0	0.0	2.6	0.0	1.3	0.0	0.0	2.2	0.0	0.0	5.0	0.0	2.0	0.0	1.4	0.0	0.0	0.0	0.0	2.0	4.8	0.0	1.6	0.0	0.0	0.0	0.0	1.9	0.0
FBXW7	R465H	0.0	0.0	0.0	9.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	6.3	0.0	12.5	0.0	0.0	0.0	0.0	0.0	0.0
FBXW7	E255*	0.0	11.5	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	3.8	0.0	2.3	3.2	0.0	0.0	2.6	0.0	5.0	0.0	4.0	0.0	0.0	2.7	6.3	0.0	0.0
KDM6A	W321*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.2	7.1	4.8	5.3	4.8	9.1	0.0	7.1	0.0	0.0	0.0	0.0	0.0	3.8	3.7	0.0	0.0	0.0	7.1
PTEN	S294fs*4	0.0	0.0	5.9	0.0	0.0	1.6	0.0	2.6	0.0	0.0	3.6	3.7	0.0	3.2	0.0	0.0	0.0	3.6	0.0	0.0	5.7	3.8	0.0	0.0	3.0	0.0	0.0	0.0	0.0	2.2
KDM6A	?	0.0	0.0	2.4	0.0	0.0	0.0	0.0	3.1	0.0	0.0	0.0	0.0	0.0	0.0	5.4	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9	0.0	4.2	0.0	0.0	0.0
KDM6A	?	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.7	0.0	0.0	0.0	0.0	0.0	2.7	2.6	0.0	0.0	0.0	0.0
SF3B1	R957Q	0.0	0.0	0.0	0.0	0.0	0.0	4.3	0.0	0.0	0.0	2.5	2.3	0.0	5.0	3.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	2.1	0.0	0.0	6.1	0.0
PIK3CA	C604R	0.0	0.0	0.0	0.0	0.0	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0
ATM	R2832C	0.0	0.0	0.0	0.0	4.2	1.4	2.7	0.0	0.0	0.0	0.0	5.1	0.0	2.8	0.0	1.9	0.0	0.0	0.0	0.0	0.0	1.8	0.0	0.0	2.0	1.5	0.0	0.0	0.0	0.0
PIK3CA	E542K	0.0	0.0	0.0	0.0	0.0	2.8	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.8	0.0	0.0	7.1	0.0	0.0	2.0	2.6	0.0	0.0	3.6	0.0
BRCA2	Q2024*	0.0	0.0	6.1	3.3	0.0	0.0	0.0	0.0	1.9	1.6	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	0.0	0.0	0.0	1.9	1.4	2.4	0.0	3.6	0.0
PIK3CA	G1049R	0.0	0.0	0.0	0.0	0.0	5.6	4.2	0.0	2.7	0.0	0.0	3.0	0.0	2.6	3.0	0.0	2.8	0.0	2.1	0.0	0.0	2.7	0.0	0.0	1.9	0.0	2.5	10.0	2.0	0.0
PIK3CA	L748I	0.0	0.0	11.8	0.0	15.4	1.8	2.9	2.4	0.0	0.0	2.3	3.6	0.0	0.0	2.6	0.0	3.1	0.0	3.1	0.0	1.8	0.0	0.0	0.0	1.6	3.4	2.3	0.0	0.0	0.0
APC	R1450*	0.0	5.9	0.0	1.8	0.0	3.5	0.0	1.3	4.3	1.1	0.0	1.3	4.0	1.3	0.0	0.0	0.0	0.0	0.0	1.4	1.1	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	3.7
SETD2	Q1164*	0.0	0.0	0.0	0.0	5.9	1.4	0.0	2.0	0.0	0.0	0.0	1.4	0.0	0.0	0.0	0.0	5.7	0.0	0.0	0.0	1.6	1.8	0.0	0.0	1.4	1.9	0.0	0.0	8.3	0.0
MED23	N759fs*12	0.0	0.0	0.0	0.0	2.0	1.1	1.8	0.0	1.6	1.2	1.2	5.0	0.0	2.9	0.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0	0.0	0.0	1.4	1.2	0.0	0.0	0.0	1.4
RUNX1	D126fs*12	0.0	0.0	0.0	5.6	0.0	1.3	0.0	1.3	1.7	1.4	1.6	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	1.2	1.2	0.0	0.0	0.0	0.0

ATRX	?	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.1	0.0	0.0	8.3	0.0	0.0	0.0	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0	13.6	0.0	0.0	0.0	0.0
PTEN	N329fs*14	0.0	0.0	6.1	2.2	0.0	1.8	0.0	2.3	4.4	2.0	0.0	3.8	2.9	0.0	0.0	0.0	0.0	0.0	2.5	2.2	2.3	3.0	0.0	0.0	0.0	5.4	0.0	0.0	0.0	0.0
KRAS	A146T	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.0	2.7	0.0	4.8	0.0	3.8	4.8	0.0	0.0	5.6	4.8	0.0	0.0	5.6	0.0	0.0	0.0	3.1	5.6	11.1	3.4	0.0
PTEN	R335*	0.0	0.0	7.1	2.6	0.0	1.6	0.0	0.0	2.2	0.0	0.0	3.4	6.1	0.0	0.0	0.0	2.7	0.0	5.4	1.9	0.0	3.1	0.0	0.0	0.0	2.8	0.0	4.0	0.0	0.0
MLLT4	Q421*	0.0	0.0	0.0	0.0	2.1	0.0	1.8	0.0	1.6	1.4	0.0	0.0	2.0	0.0	1.6	0.0	0.0	0.0	1.9	0.0	0.0	1.5	0.0	0.0	0.0	2.6	0.0	5.0	1.6	0.0
PIK3CA	E545D	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.9	0.0	0.0	0.0	2.6	0.0	0.0	0.0	0.0
PIK3CA	H1047Q	0.0	3.2	0.0	3.6	6.7	5.8	0.0	0.0	0.0	1.9	2.0	0.0	3.1	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.0	2.6	0.0	0.0	0.0	2.1	0.0	0.0	0.0	3.8
PIK3CA	H1047R	0.0	3.2	0.0	0.0	6.7	1.9	0.0	0.0	0.0	0.0	2.0	0.0	3.1	5.0	0.0	0.0	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0
RB1	R556*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	6.9	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0
MLL3	R2884*	0.0	0.0	4.6	1.0	2.0	0.0	1.8	0.0	1.1	0.0	0.0	0.0	1.3	1.1	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0	2.3	0.0
APC	G1312*	0.0	0.0	0.0	0.0	6.7	1.1	0.0	2.8	1.9	1.2	0.0	1.0	1.1	0.0	1.5	2.3	3.0	4.8	2.4	0.0	1.2	0.0	0.0	0.0	0.0	1.2	1.3	0.0	0.0	0.0
MAP3K1	H469fs*12	0.0	0.0	0.0	0.0	3.7	0.0	0.0	0.0	2.4	0.0	0.0	1.7	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9
APC	T1496fs*18	0.0	0.0	0.0	2.1	0.0	0.0	0.0	1.7	2.6	0.0	0.0	1.4	0.0	1.7	0.0	4.0	0.0	6.7	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	3.3	0.0	0.0	0.0
APC	Q1367*	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0	0.0
MLL3	D218fs*43	0.0	0.0	0.0	0.0	0.0	0.0	7.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.8	0.0	0.0	0.0	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 50: Table showing driver variants from GD6 with VAFs greater than 5% with more than one supporting read

Variant		Variant Allele Frequency																								
		GD6_E_sigmoid colon_primary	GD6_G_sigmoid colon_primary	GD6_I_sigmoid colon_primary	GD6_K_sigmoid colon_primary	GD6_M_sigmoid colon_primary	GD6_O_sigmoid colon_primary	GD6_S_sigmoid colon_primary	GD6_U_sigmoid colon_primary	GD6_W_liver_liver_met_1	GD6_Y_liver_liver_met_2	GD6_AA_liver_liver_met_3	GD6_CC_liver_liver_met_4	GD6_EE_liver_liver_met_5	GD6_GG_liver_liver_met_6	GD6_II_liver_liver_met_7	GD6_MM_liver_liver_met_7	GD6_KK_liver_liver_met_8	GD6_QQ_liver_liver_met_9	GD6_RR_liver_liver_met_10	GD6_TT_liver_liver_met_11	GD6_VV_liver_liver_met_12	GD6_AK_liver_liver_met_13	GD6_XX_liver_liver_met_13	GD6_ZZ_liver_liver_met_17	GD6_AC_liver_liver_met_18	GD6_AS_peritoneum_serosa l_met
NRAS	Q61L	44.2	34.2	34.8	46.2	32.1	26.7	32.0	18.8	46.8	47.6	42.1	40.4	37.3	26.8	43.6	38.5	50.0	43.8	37.8	40.8	31.4	35.2	43.8	43.9	46.9	24.4
BRCA2	M2393fs*19	3.9	1.9	2.5	1.1	2.2	1.3	3.3	2.7	2.4	0.0	3.3	2.6	1.2	0.0	2.3	1.3	0.8	0.9	2.0	1.0	2.7	1.5	0.0	3.2	3.3	3.0
ARID1A	Q758fs*59	5.9	2.0	2.2	1.6	6.7	1.4	1.0	1.9	1.2	1.4	3.0	1.4	1.6	0.0	3.3	3.7	0.0	1.9	1.6	4.6	0.0	3.2	3.0	0.0	1.4	2.1
USP9X	?	0.0	3.7	2.7	0.0	2.7	2.0	1.4	0.0	1.6	1.7	1.4	5.9	3.1	0.0	3.8	1.7	3.8	4.8	2.6	5.8	1.8	0.0	1.5	1.5	5.1	1.6
BRCA2	I605fs*11	3.8	0.0	2.0	2.9	4.0	1.6	0.0	4.0	2.0	0.0	1.8	7.3	2.0	4.4	3.7	1.3	0.0	3.1	2.4	2.3	0.0	2.1	7.3	0.0	5.2	6.1
TGFBR2	K130fs*19	3.8	3.8	2.4	3.8	3.2	4.5	0.0	2.3	2.7	0.0	0.0	0.0	2.3	5.9	2.7	2.0	4.3	0.0	1.4	2.6	0.0	0.0	2.9	0.0	3.2	9.4
TP53	D259V	2.4	1.5	5.1	3.7	0.0	2.4	0.0	1.5	0.0	1.3	4.5	4.3	0.0	3.8	5.3	4.3	1.4	1.4	2.9	0.0	0.0	0.0	1.9	1.6	0.0	3.7
MLL3	P2493fs*22	4.5	5.0	6.1	0.0	0.0	0.0	2.4	0.0	0.0	0.0	3.9	5.4	2.4	2.0	3.9	3.3	3.6	0.0	2.9	4.5	2.0	1.9	3.6	5.9	0.0	3.0
PIK3CA	G451_L456delinsV	2.7	2.1	4.1	0.0	0.0	4.8	0.0	1.7	0.0	2.3	5.1	5.3	2.8	3.0	2.8	0.0	1.4	2.1	3.8	1.8	0.0	0.0	0.0	0.0	3.3	1.5
TP53	K382fs*40	5.0	1.2	2.1	0.0	1.6	2.1	1.7	1.0	2.9	5.1	1.3	1.8	1.8	0.0	1.7	0.0	2.1	1.8	0.0	0.0	1.5	0.0	0.0	0.0	1.6	0.0
PTEN	T319fs*1	7.7	0.0	4.2	5.3	5.3	3.2	0.0	3.1	0.0	5.6	6.1	4.5	0.0	0.0	5.6	3.2	5.3	0.0	0.0	3.2	0.0	6.3	0.0	10.5	9.5	0.0
SPEN	I1052fs*7	2.9	0.0	0.0	1.5	0.0	1.4	1.8	1.7	2.6	0.0	0.0	1.5	0.0	2.2	0.0	0.0	1.8	2.2	1.7	0.0	2.1	1.3	1.6	2.2	0.0	1.8
GATA3	W329fs*27	1.5	1.1	0.0	6.1	0.0	1.6	0.0	1.0	0.0	1.5	1.0	1.0	1.0	1.2	0.0	1.9	0.0	0.0	0.0	1.3	0.0	1.0	2.0	1.3	0.0	3.9
CDH1	N166fs*2	0.0	1.2	1.4	1.4	0.0	4.2	1.8	0.0	3.2	0.0	4.8	1.4	0.0	1.6	2.5	0.0	1.6	2.1	3.9	0.0	0.0	0.0	1.7	0.0	2.6	2.1
TP53	V274D	0.0	1.4	1.7	5.1	1.7	2.1	0.0	0.0	1.3	2.6	0.0	1.6	1.5	1.5	2.7	0.0	1.8	0.0	2.5	0.0	0.0	1.7	3.2	0.0	0.0	0.0
GATA3	?	1.6	1.3	1.4	4.7	0.0	1.5	1.4	0.0	0.0	0.0	1.4	0.0	1.1	1.3	0.0	1.7	0.0	0.0	2.4	0.0	0.0	0.0	2.1	1.5	1.3	0.0
KDM6A	?	0.0	1.7	4.6	3.6	0.0	1.4	0.0	0.0	1.6	1.2	4.0	0.0	0.0	1.6	0.9	0.0	4.7	2.7	0.0	1.5	0.0	0.0	1.6	0.0	0.0	2.0
RB1	S230*	0.0	3.7	0.0	0.0	0.0	0.0	0.0	3.1	0.0	2.3	0.0	4.3	2.4	2.4	2.7	0.0	5.1	3.3	2.2	2.9	0.0	8.7	0.0	0.0	2.4	0.0
ATM	F357fs*33	0.0	0.0	0.0	0.0	0.0	1.8	1.4	1.4	2.3	0.0	0.0	1.4	1.5	2.9	2.0	0.0	5.5	3.8	0.0	0.0	1.6	1.4	1.7	0.0	0.0	0.0
PTEN	E299*	0.0	0.0	14.3	5.3	3.3	0.0	4.8	0.0	5.6	12.5	3.4	0.0	5.9	0.0	3.0	3.4	4.5	0.0	0.0	9.1	0.0	0.0	0.0	0.0	5.6	0.0
MLL3	D1896fs*40	1.6	0.0	0.0	0.0	0.0	0.0	0.0	1.4	2.1	0.0	1.4	2.0	0.0	2.9	1.4	0.0	4.5	5.1	0.0	0.0	0.0	1.4	0.0	0.0	2.1	2.0
TBX3	T335fs*4	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.9	0.9	3.6	0.8	0.8	0.0	0.0	0.0	1.2	0.0	0.0	1.4	0.0	0.8	0.0	0.0	4.7	1.1	1.8
PTEN	R335*	0.0	0.0	10.7	0.0	0.0	0.0	3.7	0.0	0.0	5.9	3.4	4.3	0.0	0.0	2.7	0.0	2.4	3.2	2.4	0.0	8.3	0.0	0.0	0.0	3.7	6.7
SMAD2	S276*	6.3	0.0	3.7	0.0	0.0	3.4	0.0	0.0	9.7	5.3	0.0	5.0	13.8	3.0	4.0	0.0	0.0	5.9	0.0	5.4	7.1	0.0	0.0	0.0	0.0	0.0
PIK3CA	P104_G106delinsR	0.0	0.0	0.0	2.0	2.0	0.0	0.0	0.0	5.9	2.2	0.0	0.0	2.8	0.0	4.9	0.0	2.0	3.7	0.0	1.9	0.0	0.0	2.3	4.5	2.7	0.0
PIK3CA	P104L	4.5	2.6	4.9	2.0	0.0	0.0	0.0	0.0	0.0	2.2	2.4	2.5	0.0	5.7	2.4	4.3	2.0	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0

RB1	P232fs*4	0.0	0.0	0.0	0.0	0.0	2.8	0.0	3.2	0.0	4.9	0.0	0.0	0.0	4.8	2.6	0.0	3.2	3.3	1.1	5.9	0.0	4.3	0.0	0.0	4.9	0.0
PIK3CA	G106V	0.0	2.5	0.0	1.9	2.0	5.3	2.2	0.0	0.0	0.0	0.0	2.4	0.0	0.0	0.0	0.0	1.9	3.6	1.2	0.0	0.0	0.0	2.3	0.0	5.6	0.0
MLL3	Q448*	2.0	0.0	0.0	0.0	0.0	1.6	0.0	0.0	1.7	6.9	0.0	0.0	0.0	3.0	1.1	0.0	1.9	0.0	1.7	1.8	1.5	0.0	1.9	0.0	0.0	0.0
PIK3CA	R93P	0.0	5.7	0.0	1.9	0.0	0.0	0.0	0.0	0.0	2.3	0.0	0.0	0.0	0.0	5.1	0.0	2.2	3.2	0.0	0.0	3.7	0.0	2.4	3.6	2.4	1.8
SMAD4	G89*	0.0	0.0	1.7	0.0	0.0	1.6	0.0	1.4	5.9	3.1	2.0	0.0	2.0	0.0	0.0	0.0	2.0	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6
RB1	Q395*	0.0	0.0	0.0	1.8	0.0	0.0	1.4	0.0	0.0	2.3	4.8	1.8	4.0	1.8	0.0	3.4	2.0	0.0	0.0	0.0	2.6	0.0	0.0	0.0	0.0	0.0
RB1	R552Q	8.0	0.0	2.8	1.6	1.8	2.9	0.0	0.0	0.0	2.7	0.0	0.0	0.0	0.0	0.0	2.1	0.0	2.4	0.0	0.0	0.0	2.7	2.3	0.0	0.0	0.0
TBL1XR1	E393*	7.1	0.0	2.1	2.2	1.6	1.8	0.0	0.0	2.8	5.7	0.0	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	2.2	0.0
FBXW7	R689W	0.0	0.0	1.6	0.0	0.0	4.1	16.7	4.9	0.0	0.0	1.8	3.6	0.0	5.3	2.2	0.0	0.0	0.0	0.0	0.0	0.0	3.1	0.0	0.0	0.0	0.0
BRCA2	E2028*	0.0	0.0	0.0	1.4	1.1	0.0	0.0	0.0	1.6	2.1	0.0	6.5	0.0	0.0	1.2	0.0	1.1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6
RHOA	G17E	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	6.4	0.0	0.0	1.4	0.0	1.9	0.0	1.2	0.0	0.0	1.4	0.0	2.4	1.7	0.0	1.7	0.0
SPEN	D1313_S1314ins*	0.0	0.0	4.4	2.4	0.0	0.0	0.0	1.7	0.0	2.9	2.0	0.0	0.0	6.3	1.8	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SETD2	E670*	0.0	0.0	0.0	2.9	1.5	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	7.1	3.8	0.0	0.0	0.0	0.0	1.9	0.0	2.3	2.1
SMAD2	E159*	0.0	0.0	0.0	3.8	0.0	0.0	0.0	0.0	4.8	0.0	0.0	0.0	0.0	0.0	0.0	4.5	12.0	0.0	0.0	0.0	8.3	5.9	3.4	0.0	3.8	0.0
NF1	E2469*	0.0	0.0	3.1	2.3	0.0	2.6	2.0	0.0	0.0	0.0	0.0	0.0	0.0	2.7	2.2	0.0	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MAP3K1	F341fs*44	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	1.3	0.0	0.0	0.0	2.4	0.0	0.9	0.0	1.5	0.0	0.0	0.0	1.1	1.2	0.0	0.0	0.0	0.0
PIK3CA	H450_D454delHGLED	5.3	2.0	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	0.0	2.1	0.0	0.0	0.0	0.0	3.7	0.0	3.4	0.0
NF1	E1459*	0.0	0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9	2.2	8.3	0.0	0.0	2.4	0.0	0.0	4.8	0.0	0.0	2.6	0.0	0.0	0.0
APC	A1296fs*9	0.0	12.5	3.3	0.0	0.0	0.0	0.0	3.7	0.0	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.0	0.0	0.0	0.0	3.2	0.0
MLL3	S1745*	0.0	0.0	0.0	2.8	0.0	2.1	0.0	0.0	0.0	3.0	1.7	2.0	0.0	0.0	0.0	0.0	5.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CTNNB1	?	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	4.5	0.0	0.0	3.7	0.0	0.0	0.0	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.3
PTPRB	Q680*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	5.1	0.0	2.4	0.0	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4
ATRX	?	0.0	0.0	0.0	0.0	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.3	0.0	0.0	1.6	0.0	7.7	0.0	0.0	0.0	0.0	7.1
STAG2	R614*	0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	5.1	0.0	0.0	0.0	0.0	0.0	3.6	0.0	0.0	0.0	0.0	0.0	2.3	0.0	1.9	0.0	0.0	0.0
APC	D849fs*11	0.0	0.0	0.0	3.3	0.0	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
APC	Q1237fs*2	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SMAD4	N263fs*74	0.0	6.7	0.0	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ATR	W1964*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	0.0	0.0	0.0	0.0	12.5	0.0	0.0	0.0	5.6	0.0	0.0	0.0	0.0	0.0	0.0
APC	A1325fs*7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.8	0.0	0.0	12.5	0.0	0.0	0.0	0.0
NCOR1	Q1362*	0.0	0.0	0.0	0.0	4.3	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

19 References

1. International Agency for Research on Cancer (source: The Global Cancer Observatory). Colorectal cancer factsheet. (2019). Available at: https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf. (Accessed: 29th July 2019)

2. CRUK. Cancer Research UK Bowel Cancer: Key Facts. (2011).

3. Kune, G. A. *et al.* Survival in patients with large-bowel cancer. A population-based investigation from the Melbourne Colorectal Cancer Study. *Dis Colon Rectum* **33**, 938–946 (1990).

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

5. Bodmer, W. F. *et al.* Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature* **328**, 614–616 (1987).

6. Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).

7. Fernandes, M. S., Carneiro, F., Oliveira, C. & Seruca, R. Colorectal cancer and RASSF family--a special emphasis on RASSF1A. *Int J Cancer* **132**, 251–258 (2013).

8. Douillard, J. Y. *et al.* Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer. *N Engl J Med* **369**, 1023–1034 (2013).

9. Pretlow, T. P., Brasitus, T. A., Fulton, N. C., Cheyer, C. & Kaplan, E. L. K-ras mutations in putative preneoplastic lesions in human colon. *J Natl Cancer Inst* **85**, 2004–2007 (1993).

10. Vaughn, C. P., Zobell, S. D., Furtado, L. V, Baker, C. L. & Samowitz, W. S. Frequency of KRAS, BRAF, and NRAS mutations in colorectal cancer. *Genes Chromosom. Cancer* **50**, 307–312 (2011).

11. Richman, S. D. *et al.* KRAS and BRAF mutations in advanced colorectal cancer are associated with poor prognosis but do not preclude benefit from oxaliplatin or

irinotecan: results from the MRC FOCUS trial. *J Clin Oncol* **27**, 5931–5937 (2009).

12. Janku, F. Bringing target-matched PI3King from the bench to the clinic. *Cell Cycle* **12**, 1817–1818 (2013).

13. Polivka Jr., J. & Janku, F. Molecular targets for cancer therapy in the PI3K/AKT/mTOR pathway. *Pharmacol Ther* **142**, 164–175 (2014).

14. Samuels, Y. *et al.* High Frequency of Mutations of the PIK3CA Gene in Human Cancers. *Science (80-.).* **304**, 554 (2004).

15. Huang, M. Y. *et al.* Predictive value of ERCC1, ERCC2, and XRCC1 overexpression for stage III colorectal cancer patients receiving FOLFOX-4 adjuvant chemotherapy. *J Surg Oncol* **108**, 457–464 (2013).

16. Network, T. C. G. A. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

17. Frattini, M. *et al.* PTEN loss of expression predicts cetuximab efficacy in metastatic colorectal cancer patients. *Br J Cancer* **97**, 1139–1145 (2007).

18. Loupakis, F. *et al.* PTEN expression and KRAS mutations on primary tumors and metastases in the prediction of benefit from cetuximab plus irinotecan for patients with metastatic colorectal cancer. *J Clin Oncol* **27**, 2622–2629 (2009).

19. COSMIC. COSMIC (Catalogue of Somatic Mutations in Cancer) . (2014).

20. Carpten, J. D. *et al.* A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. *Nature* **448**, 439–444 (2007).

21. Li, X.-L., Zhou, J., Chen, Z.-R. & Chng, W.-J. P53 mutations in colorectal cancer - molecular pathogenesis and pharmacological reactivation. *World J. Gastroenterol.* **21**, 84–93 (2015).

22. Li, Q. & Lozano, G. Molecular pathways: targeting Mdm2 and Mdm4 in cancer therapy. *Clin. Cancer Res.* **19**, 34–41 (2013).

23. Ryan, K. M., Phillips, A. C. & Vousden, K. H. Regulation and function of the p53 tumor suppressor protein. *Curr. Opin. Cell Biol.* **13**, 332–337 (2001).

24. Smith, N. D., Rubenstein, J. N., Eggener, S. E. & Kozlowski, J. M. The p53 tumor

suppressor gene and nuclear protein: Basic science review and relevance in the management of bladder cancer. *J. Urol.* **169**, 1219–1228 (2003).

25. Dimri, G. P., Itahana, K., Acosta, M. & Campisi, J. *Regulation of a Senescence Checkpoint Response by the E2F1 Transcription Factor and p14 ARF Tumor Suppressor*. *MOLECULAR AND CELLULAR BIOLOGY* **20**, (2000).

26. Campisi, J. & d'Adda di Fagagna, F. Cellular senescence: when bad things happen to good cells. *Nat. Rev. Mol. Cell Biol.* **8**, 729–740 (2007).

27. Polager, S. & Ginsberg, D. p53 and E2f: partners in life and death. *Nat. Rev. Cancer* **9**, 738–748 (2009).

28. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).

29. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

30. Olivier, M., Hussain, S. P., Caron de Fromentel, C., Hainaut, P. & Harris, C. C. TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. *IARC Sci. Publ.* 247–70 (2004).

31. Russo, A. *et al.* The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment. *J. Clin. Oncol.* **23**, 7518–28 (2005).

32. Tsilimigras, D. I. *et al.* Clinical significance and prognostic relevance of KRAS, BRAF, PI3K and TP53 genetic mutation analysis for resectable and unresectable colorectal liver metastases: A systematic review of the current evidence. (2018). doi:10.1016/j.suronc.2018.05.012

33. Mouradov, D. *et al.* Survival in stage II/III colorectal cancer is independently predicted by chromosomal and microsatellite instability, but not by specific driver mutations. *Am. J. Gastroenterol.* **108**, 1785–93 (2013).

34. Sjo, O. H. *et al.* Peritoneal carcinomatosis of colon cancer origin: Highest incidence in women and in patients with right-sided tumors. *J. Surg. Oncol.* **104**, 792–797 (2011).

35. Tominaga, T. *et al.* Combination of *p53* codon 72 polymorphism and inactive *p53*

mutation predicts chemosensitivity to 5-fluorouracil in colorectal cancer. *Int. J. Cancer* **126**, NA-NA (2009).

36. Mazelin, L. *et al.* Netrin-1 controls colorectal tumorigenesis by regulating apoptosis. *Nature* **431**, 80–84 (2004).

37. Schutte, M. *et al.* DPC4 gene in various tumor types. *Cancer Res* **56**, 2527–2530 (1996).

38. Thiagalingam, S. *et al.* Evaluation of candidate tumour suppressor genes on chromosome 18 in colorectal cancers. *Nat Genet* **13**, 343–346 (1996).

39. Shi, Y. & Massagué, J. Mechanisms of TGF-β Signaling from Cell Membrane to the Nucleus. *Cell* **113**, 685–700 (2003).

40. Inman, G. J. Linking Smads and transcriptional activation. *Biochem. J.* **386**, e1–e3 (2005).

41. Jung, B., Staudacher, J. J. & Beauchamp, D. Transforming Growth Factor β Superfamily Signaling in Development of Colorectal Cancer. *Gastroenterology* **152**, 36–52 (2017).

42. Syed, V. TGF-β Signaling in Cancer. *J. Cell. Biochem.* **117**, 1279–1287 (2016).

43. Grady, W. M. *et al.* Mutational Inactivation of Transforming Growth Factor β Receptor Type II in Microsatellite Stable Colon Cancers. *Cancer Res.* **59**, (1999).

44. Markowitz, S. *et al.* Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science* **268**, 1336–8 (1995).

45. Kodach, L. L. *et al.* The Bone Morphogenetic Protein Pathway Is Inactivated in the Majority of Sporadic Colorectal Cancers. *Gastroenterology* **134**, 1332-1341.e3 (2008).

46. Calva-Cerqueira, D. *et al.* The rate of germline mutations and large deletions of SMAD4 and BMPR1A in juvenile polyposis. *Clin. Genet.* **75**, 79–85 (2009).

47. Aytac, E. *et al.* Genotype-defined cancer risk in juvenile polyposis syndrome. *Br. J. Surg.* **102**, 114–118 (2015).

48. Miyaki, M. *et al.* Higher frequency of Smad4 gene mutation in human colorectal

cancer with distant metastasis. *Oncogene* **18**, 3098–3103 (1999).

49. Ogino, S. *et al.* Prognostic significance and molecular associations of 18q loss of heterozygosity: a cohort study of microsatellite stable colorectal cancers. *J. Clin. Oncol.* **27**, 4591–8 (2009).

50. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* (2015). doi:10.1038/nm.3967

51. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).

52. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350–1356 (2015).

53. Hveem, T. S. *et al.* Prognostic impact of genomic instability in colorectal cancer. *Br J Cancer* **110**, 2159–2164 (2014).

54. Mouradov, D. *et al.* Survival in stage II/III colorectal cancer is independently predicted by chromosomal and microsatellite instability, but not by specific driver mutations. *Am. J. Gastroenterol.* **108**, 1785–93 (2013).

55. Birkbak, N. J. *et al.* Paradoxical relationship between chromosomal instability and survival outcome in cancer. *Cancer Res.* **71**, 3447–3452 (2011).

56. Smith, J. C. & Sheltzer, J. M. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. *Elife* **7**, (2018).

57. Swanton, C. *et al.* Chromosomal instability determines taxane response. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 8671–8676 (2009).

58. Burrell, R. A. & Swanton, C. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Mol. Oncol.* **8**, 1095–111 (2014).

59. Thompson, L. L., Jeusset, L. M.-P., Lepage, C. C. & McManus, K. J. Evolving therapeutic strategies to exploit chromosome instability in cancer. *Cancers* **9**, (2017).

60. Fanale, D. *et al.* Stabilizing versus destabilizing the microtubules: a double-edge sword for an effective cancer treatment option? *Anal. Cell. Pathol. (Amst).* **2015**, 690916 (2015).

61. Umar, A. *et al.* Revised Bethesda Guidelines for Hereditary Nonpolyposis Colorectal Cancer (Lynch Syndrome) and Microsatellite Instability. *JNCI J. Natl. Cancer Inst.* **96**, 261–268 (2004).

62. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98 (2008).

63. National Institute for Health and Care Excellence. Molecular testing strategies for Lynch syndrome in people with colorectal cancer | Guidance and guidelines | NICE. Available at: https://www.nice.org.uk/guidance/dg27. (Accessed: 30th December 2018)

64. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* **372**, 2509–2520 (2015).

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

66. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.* **45**, 136–44 (2013).

67. Gardiner-Garden, M. & Frommer, M. CpG Islands in Vertibrate Genomes. *J.Mol.Biol.* **196**, 261–282 (1987).

68. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* **22**, 271–282 (2012).

69. Lei, Z. *et al.* Identification of molecular subtypes of gastric cancer with different responses to pi3-kinase inhibitors and 5-fluorouracil. *Gastroenterology* **145**, 554–565 (2013).

70. Marisa, L. *et al.* Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Med.* **10**, e1001453 (2013).

71. Roepman, P. *et al.* Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int. J. Cancer* **134**, 552–562 (2014).

72. De Sousa E Melo, F. *et al.* Poor-prognosis colon cancer is defined by a

molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.* **19**, 614–618 (2013).

73. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* **19**, 619–625 (2013).

74. Fontana, E., Eason, K., Cervantes, A., Salazar, R. & Sadanandam, A. Context matters-consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **30**, 520–527 (2019).

75. Müller, M. F., Ibrahim, A. E. K. & Arends, M. J. Molecular pathological classification of colorectal cancer. *Virchows Arch.* **469**, 125–34 (2016).

76. Dietmaier, W. *et al.* Evaluation of KRAS and BRAF mutation analysis in primary metastatic colorectal cancers versus matched syn- and metachronous metastases. *Cancer Res. Conf. 101st Annu. Meet. Am. Assoc. Cancer Res. AACR 2010 Washington, DC United States.* **70**, (2010).

77. Etienne-Grimaldi, M. C. *et al.* K-Ras mutations and treatment outcome in colorectal cancer patients receiving exclusive fluoropyrimidine therapy. *Clin. Cancer Res.* **14**, 4830 (2008).

78. Knijn, N. *et al.* KRAS mutation analysis: a comparison between primary tumours and matched liver metastases in 305 colorectal cancer patients. *Br. J. Cancer* **104**, 1020–6 (2011).

79. Lee, S. Y. *et al.* Comparative genomic analysis of primary and synchronous metastatic colorectal cancers. *PLoS One* **9**, (2014).

80. Vermaat, J. S. *et al.* Primary colorectal cancers and their subsequent hepatic metastases are genetically different: Implications for selection of patients for targeted treatment. *Clin. Cancer Res.* **18**, 688–699 (2012).

81. Brannon, A. R. *et al.* Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol.* **15**, 454 (2014).

82. Goswami, R. S. *et al.* Hotspot mutation panel testing reveals clonal evolution in a study of 265 paired primary and metastatic tumors. *Clin. Cancer Res.* **21**, 2644–2651 (2015).

83. Mamlouk, S. *et al.* DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nat. Commun.* **8**, 14093 (2017).

84. Mogensen, M. B. *et al.* Genomic alterations accompanying tumour evolution in colorectal cancer: tracking the differences between primary tumours and synchronous liver metastases by whole-exome sequencing. *BMC Cancer* **18**, 752 (2018).

85. Naxerova, K. *et al.* Origins of lymphatic and distant metastases in human colorectal cancer. *Science (80-.).* **357**, 55–60 (2017).

86. Han, C. B., Li, F., Ma, J. T. & Zou, H. W. Concordant KRAS mutations in primary and metastatic colorectal cancer tissue specimens: A meta-analysis and systematic review. *Cancer Investigation* **30**, 741–747 (2012).

87. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).

88. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883–892 (2012).

89. Niida, A., Nagayama, S., Miyano, S. & Mimori, K. Understanding intratumor heterogeneity by combining genome analysis and mathematical modeling. *Cancer Sci.* **109**, 884–892 (2018).

90. Wood, H. M. *et al.* Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res* **38**, e151 (2010).

91. Diep, C. B. *et al.* The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. *Genes Chromosom. Cancer* **45**, 31–41 (2006).

92. Mekenkamp, L. J. *et al.* Chromosomal Copy Number Aberrations in Colorectal Metastases Resemble Their Primary Counterparts and Differences Are Typically Non-Recurrent. *PLoS One* **9**, (2014).

93. Stange, D. E. *et al.* Expression of an ASCL2 related stem cell signature and IGF2 in colorectal cancer liver metastases with 11p15.5 gain. *Gut* **59**, 1236–44 (2010).

94. Vakiani, E. *et al.* Comparative genomic analysis of primary versus metastatic

colorectal carcinomas. *J. Clin. Oncol.* **30**, 2956–2962 (2012).

95. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–9 (2015).

96. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).

97. Savas, P. *et al.* The Subclonal Architecture of Metastatic Breast Cancer: Results from a Prospective Community-Based Rapid Autopsy Program "CASCADE". *PLOS Med.* **13**, e1002204 (2016).

98. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).

99. Notta, F. *et al.* A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).

100. Diaz Jr., L. A. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537–540 (2012).

101. Nakamura, T. *et al.* Mechanisms of acquired resistance to afatinib clarified with liquid biopsy. *PLoS One* **13**, e0209384 (2018).

102. Cleary, A. S., Leonard, T. L., Gestl, S. A. & Gunther, E. J. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature* **508**, 113–117 (2014).

103. McFadden, D. G. *et al.* Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell* **156**, 1298–1311 (2014).

104. Brown, D. *et al.* Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nat. Commun.* **8**, 14944 (2017).

105. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* **46**, 225–233 (2014).

106. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).

107. Shi, J.-Y. *et al.* Inferring the progression of multifocal liver cancer from spatial and temporal genomic heterogeneity. *Oncotarget* **7**, 2867–2877 (2016).

108. Xue, R. *et al.* Variable Intra-Tumor Genomic Heterogeneity of Multiple Lesions in Patients With Hepatocellular Carcinoma. *Gastroenterology* **150**, 998–1008 (2016).

109. Bréchot, C. Pathogenesis of hepatitis B virus-related hepatocellular carcinoma: old and new paradigms. *Gastroenterology* **127**, S56-61 (2004).

110. Ferlay, J. *et al.* GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. (2013).

111. Eldredge, N. & Gould, J. G. Punctuated Equilibria: an alternative to phyletic gradualism. in *Models in Paleobiology* (ed. Thomas, J. M.) 82–115 (Freeman, Cooper and Company, 1972).

112. Kloosterman, W. P. *et al.* Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. *Genome Biol.* **12**, R103 (2011).

113. Murugaesu, N. *et al.* Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov.* **5**, 821–831 (2015).

114. Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *N Engl J Med* **319**, 525–532 (1988).

115. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–8 (1976).

116. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).

117. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

118. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).

119. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–11 (2013).

120. Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer evolution: Mathematical models and computational inference. *Syst. Biol.* **64**, e1–e25 (2015).

121. Miller, C. A. *et al.* SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput. Biol.* **10**, e1003665 (2014).

122. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).

123. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7**, 1740–1752 (2014).

124. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–93 (2014).

125. Letouze, E., Allory, Y., Bollet, M. A., Radvanyi, F. & Guyon, F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biol* **11**, R76 (2010).

126. Schwarz, R. F. *et al.* Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* **10**, e1003535 (2014).

127. Scheele, J., Stangl, R. & Altendorf-Hofmann, A. Hepatic metastases from colorectal carcinoma: impact of surgical resection on the natural history. *Br J Surg* **77**, 1241–1246 (1990).

128. Rees, M., Tekkis, P. P., Welsh, F. K., O'Rourke, T. & John, T. G. Evaluation of long-term survival after hepatic resection for metastatic colorectal cancer: a multifactorial model of 929 patients. *Ann Surg* **247**, 125–135 (2008).

129. Bismuth, H. *et al.* Resection of nonresectable liver metastases from colorectal cancer after neoadjuvant chemotherapy. *Ann Surg* **224**, 502–509 (1996).

130. NICE. Colorectal cancer: the diagnosis and management of colorectal cancer. NICE clinical guidance 131 [online]. Available at:

http://www.nice.org.uk/proxy/?sourceUrl=http%3a%2f%2fguidance.nice.org.uk%2fCG131%2fGuidance. (2011).

131. Adams, R. A. *et al.* Intermittent versus continuous oxaliplatin and fluoropyrimidine combination chemotherapy for first-line treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet Oncol* **12**, 642–653 (2011).

132. Seymour, M. T. *et al.* Different strategies of sequential and combination chemotherapy for patients with poor prognosis advanced colorectal cancer (MRC FOCUS): a randomised controlled trial. *Lancet* **370**, 143–152 (2007).

133. Koopman, M. *et al.* Sequential versus combination chemotherapy with capecitabine, irinotecan, and oxaliplatin in advanced colorectal cancer (CAIRO): a phase III randomised controlled trial. *Lancet* **370**, 135–142 (2007).

134. Saltz, L. B. *et al.* Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer: a randomized phase III study. *J Clin Oncol* **26**, 2013–2019 (2008).

135. Welch, S., Spithoff, K., Rumble, R. B., Maroun, J. & Gastrointestinal Cancer Disease Site, G. Bevacizumab combined with chemotherapy for patients with advanced colorectal cancer: a systematic review. *Ann Oncol* **21**, 1152–1162 (2010).

136. NICE. Bevacizumab in combination with oxaliplatin and either fluorouracil plus folinic acid or capecitabine for the treatment of metastatic colorectal cancer. NICE technology appraisal guidance 212 [online]. Available at: http://www.nice.org.uk/guidance/ta212. (2010).

137. Wilhelm, S. M. *et al.* Regorafenib (BAY 73-4506): A new oral multikinase inhibitor of angiogenic, stromal and oncogenic receptor tyrosine kinases with potent preclinical antitumor activity. *Int. J. Cancer* **129**, 245–255 (2011).

138. Tabernero, J. *et al.* Ramucirumab versus placebo in combination with second-line FOLFIRI in patients with metastatic colorectal carcinoma that progressed during or after first-line therapy with bevacizumab, oxaliplatin, and a fluoropyrimidine (RAISE): A randomised, double-blin. *Lancet Oncol.* **16**, 499–508 (2015).

139. Van Cutsem, E. *et al.* Addition of aflibercept to fluorouracil, leucovorin, and

irinotecan improves survival in a phase III randomized trial in patients with metastatic colorectal cancer previously treated with an oxaliplatin-based regimen. *J. Clin. Oncol.* **30**, 3499–3506 (2012).

140. Grothey, A. *et al.* Regorafenib monotherapy for previously treated metastatic colorectal cancer (CORRECT): An international, multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet* **381**, 303–312 (2013).

141. NICE. Ramucirumab for treating advanced gastric cancer or gastro – oesophageal junction adenocarcinoma pre previously viously treated with chemotherapy - Information for the public. *Natl. Inst. Heal. Care Excell. Guidel. T[a378]* 1–2 (2016).

142. NICE. Regorafenib for metastatic colorectal cancer after treatment for metastatic disease (terminated appraisal). (2015).

143. National Institute for Health and Care Excellence (NICE). Aflibercept in combination with irinotecan and fluorouracil-based therapy for treating metastatic colorectal cancer that has progressed following prior oxaliplatin-based chemotherapy. (2014).

144. Salomon, D., Brandt, R., Ciardiello, F. & Normano, N. Epidermal growth factor-related peptides and their receptors in human malignancies. *Crit. Rev. Oncol. Hematol.* **19**, 183–232 (1995).

145. Seymour, M. T. *et al.* Panitumumab and irinotecan versus irinotecan alone for patients with KRAS wild-type, fluorouracil-resistant advanced colorectal cancer (PICCOLO): a prospectively stratified randomised trial. *Lancet Oncol.* **14**, 749–759 (2013).

146. Amado, R. G. *et al.* Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol* **26**, 1626–1634 (2008).

147. Pietrantonio, F. *et al.* Predictive role of BRAF mutations in patients with advanced colorectal cancer receiving cetuximab and panitumumab: A meta-analysis. *Eur. J. Cancer* **51**, 587–594 (2015).

148. Osumi, H. *et al.* Clinical relevance of circulating tumor DNA assessed through deep sequencing in patients with metastatic colorectal cancer. *Cancer Med.* **8**, 408–417 (2019).

149. Knebel, F. H. *et al.* Circulating Tumor DNA Detection in the Management of Anti-EGFR Therapy for Advanced Colorectal Cancer. *Front. Oncol.* **9**, 170 (2019).

150. Nakamura, Y. & Yoshino, T. Clinical Utility of Analyzing Circulating Tumor DNA in Patients with Metastatic Colorectal Cancer. *Oncologist* **23**, 1310–1318 (2018).

151. Sugimachi, K. *et al.* Serial mutational tracking in surgically resected locally advanced colorectal cancer with neoadjuvant chemotherapy. *Br. J. Cancer* **119**, 419–423 (2018).

152. Pietrantonio, F. *et al.* Heterogeneity of acquired resistance to anti-EGFR monoclonal antibodies in patients with metastatic colorectal cancer. *Clin. Cancer Res.* **23**, 2414–2422 (2017).

153. Molinari, F. *et al.* Differing deregulation of EGFR and downstream proteins in primary colorectal cancer and related metastatic sites may be clinically relevant. *Br. J. Cancer* **100**, 1087–94 (2009).

154. Jia, J., Morse, M. A., Nagy, R. J., Lanman, R. B. & Strickler, J. H. Cell-Free DNA Profiling to Discover Mechanisms of Exceptional Response to Cabozantinib Plus Panitumumab in a Patient With Treatment Refractory Metastatic Colorectal Cancer. *Front. Oncol.* **8**, 305 (2018).

155. Mohan, S. *et al.* Changes in colorectal carcinoma genomes under anti-EGFR therapy identified by whole-genome plasma DNA sequencing. *PLoS Genet.* **10**, e1004271 (2014).

156. Stahler, A. *et al.* Influence of mRNA expression of epiregulin and amphiregulin on outcome of patients with metastatic colorectal cancer treated with 5-FU/LV plus irinotecan or irinotecan plus oxaliplatin as first-line treatment (FIRE 1-trial). *Int. J. Cancer* **138**, 739–746 (2016).

157. Khambata-Ford, S. *et al.* Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *J. Clin. Oncol.* **25**, 3230–3237 (2007).

158. Kopetz, S. *et al.* PLX4032 in metastatic colorectal cancer patients with mutant BRAF tumors. *J Clin Oncol* **28**, (2010).

159. Capalbo, C. *et al.* Vemurafenib and panitumumab combination tailored therapy in

BRAF-mutated metastatic colorectal cancer: A case report. *Cancer Biol Ther* **15**, 826–831 (2014).

160. Memorial Sloan Kettering Cancer Center. Vemurafenib and Panitumumab Combination Therapy in Patients With BRAF V600E Mutated Metastatic Colorectal Cancer. *June 09, 2016* Available at: https://clinicaltrials.gov/ct2/show/NCT01791309. (Accessed: 10th June 2016)

161. Mei, Z. B., Duan, C. Y., Li, C. B., Cui, L. & Ogino, S. Prognostic role of tumor PIK3CA mutation in colorectal cancer: a systematic review and meta-analysis. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **27**, 1836–48 (2016).

162. Karapetis, C. S. *et al.* PIK3CA, BRAF, and PTEN status and benefit from cetuximab in the treatment of advanced colorectal cancer--results from NCIC CTG/AGITG CO.17. *Clin Cancer Res* **20**, 744–753 (2014).

163. Domingo, E. *et al.* Evaluation of PIK3CA mutation as a predictor of benefit from nonsteroidal anti-inflammatory drug therapy in colorectal cancer. *J Clin Oncol* **31**, 4297–4305 (2013).

164. Liao, X. *et al.* Aspirin Use, Tumor *PIK3CA* Mutation, and Colorectal-Cancer Survival. *N. Engl. J. Med.* **367**, 1596–1606 (2012).

165. Vasudevan, K. M. *et al.* AKT-independent signaling downstream of oncogenic PIK3CA mutations in human cancer. *Cancer Cell* **16**, 21–32 (2009).

166. McCormack, F. X. *et al.* Efficacy and safety of sirolimus in lymphangioleiomyomatosis. *N Engl J Med* **364**, 1595–1606 (2011).

167. Eng, C. *et al.* A randomized, placebo-controlled, phase 1/2 study of tivantinib (ARQ 197) in combination with irinotecan and cetuximab in patients with metastatic colorectal cancer with wild-type KRAS who have received first-line systemic therapy. *Int. J. Cancer* **139**, 177–186 (2016).

168. Rodrik-Outmezguine, V. S. *et al.* mTOR kinase inhibition causes feedback-dependent biphasic regulation of AKT signaling. *Cancer Discov* **1**, 248–259 (2011).

169. Pike, K. G. *et al.* Optimization of potent and selective dual mTORC1 and mTORC2 inhibitors: the discovery of AZD8055 and AZD2014. *Bioorg Med Chem Lett* **23**,

1212–1216 (2013).

170. Blaser, B. *et al.* Antitumor activities of ATP-competitive inhibitors of mTOR in colon cancer cells. *BMC Cancer* **12**, 86 (2012).

171. Maughan, T. FOCUS4: Molecular selection of therapy in metastatic colorectal cancer: a molecularly stratified randonis. *8th NCRI Cancer Conference* (2012).

172. Lipson, E. J. *et al.* Durable cancer regression off-treatment and effective reinduction therapy with an anti-PD-1 antibody. *Clin. Cancer Res.* **19**, 462–468 (2013).

173. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

174. Study of Pembrolizumab (MK-3475) as Monotherapy in Participants With Previously-Treated Locally Advanced Unresectable or Metastatic Colorectal Cancer (MK-3475-164/KEYNOTE-164). *ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US).* (2015). Available at: https://clinicaltrials.gov/ct2/show/NCT02460198. (Accessed: 26th June 2016)

175. Herbst, R. S. *et al.* Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567 (2014).

176. Chung, K. Y. *et al.* Phase II study of the anti-cytotoxic T-lymphocyte-associated antigen 4 monoclonal antibody, tremelimumab, in patients with refractory metastatic colorectal cancer. *J. Clin. Oncol.* **28**, 3485–3490 (2010).

177. Chalabi, M. *et al.* LBA37_PRNeoadjuvant ipilimumab plus nivolumab in early stage colon cancer. *Ann. Oncol.* **29**, (2018).

178. Parkhurst, M. R. *et al.* T cells targeting carcinoembryonic antigen can mediate regression of metastatic colorectal cancer but induce severe transient colitis. *Mol. Ther.* **19**, 620–6 (2011).

179. Rao, B. *et al.* Clinical outcomes of active specific immunotherapy in advanced colorectal cancer and suspected minimal residual colorectal cancer: a meta-analysis and system review. *J. Transl. Med.* **9**, 17 (2011).

180. The Royal College of Pathologists. Types of post-mortems. *The Royal College of Pathologists* (2019). Available at: https://www.rcpath.org/discover-

pathology/what-is-pathology/information-about-post-mortems-for-friends-and-relatives-/types-of-post-mortems-.html. (Accessed: 5th August 2019)

181. Turnbull, A., Osborn, M. & Nicholas, N. Hospital autopsy: Endangered or extinct? *J. Clin. Pathol.* **68**, 601–4 (2015).

182. Davies, D. J. *et al.* The decline of the hospital autopsy: A safety and quality issue for healthcare in Australia. *Med. J. Aust.* **180**, 281–285 (2004).

183. Chariot, P. *et al.* Declining Autopsy Rate in a French Hospital. 739–745 (2009).

184. National Confidential Enquiry into Patient Outcome and Death. *The Coroner's Autopsy : Do we deserve better?* (2006).

185. Underwood, J. C. E. The impact on histopathology practice of new human tissue legislation in the UK. *Histopathology* **49**, 221–228 (2006).

186. Bailey, D. The future of the coronial autopsy service. *Royal College of Pathologists Bulletin Number 171* (2015). Available at: https://www.rcpath.org/profession/publications/college-bulletin/bulletin/the-future-of-the-coronial-autopsy-service.html. (Accessed: 24th August 2016)

187. Tsitsikas, D. a, Brothwell, M., Chin Aleong, J.-A. & Lister, A. T. The attitudes of relatives to autopsy: a misconception. *J. Clin. Pathol.* **64**, 412–4 (2011).

188. McPhee, S. J., Bottles, K., Lo, B., Saika, G. & Crommie, D. To Redeem Them from Death. *Am. J. Med.* **80**, 665–671 (1986).

189. Cohen, M. C., Blakey, S., Donn, T., McGovern, S. & Parry, L. An audit of parents'/guardians' wishes recorded after coronial autopsies in cases of sudden unexpected death in infancy: issues raised and future directions. *Med. Sci. Law* **49**, 179–184 (2009).

190. Boston University CTE centre. Boston University Chronic Traumatic Encephalopathy Centre Brain Bank. *Boston University Chronic Traumatic Encephalopathy Centre Website* (2014). Available at: http://www.bu.edu/cte/our-research/brain-bank/.

191. Toledo, J. B. *et al.* A platform for discovery: The University of Pennsylvania Integrated Neurodegenerative Disease Biobank. *Alzheimers. Dement.* **10**, 477-484.e1 (2014).

192. Cancer Research UK. A study looking at blood and tissue samples to learn more about advanced cancer (PEACE). *Cancer Research UK website* (2014).

193. Alix-Panabières, C. & Pantel, K. Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. *Cancer Discov.* **6**, 479–491 (2016).

194. Misale, S. *et al.* Blockade of EGFR and MEK Intercepts Heterogeneous Mechanisms of Acquired Resistance to Anti-EGFR Therapies in Colorectal Cancer. *Sci. Transl. Med.* **6**, 224ra26 LP-224ra26 (2014).

195. GIFT research autopsy service. GIFT research autopsy service website. *University of Leeds Website* (2016). Available at: http://www.gift.leeds.ac.uk/.

196. Loughrey, B., Quirke, P. & Shepherd, N. *The Royal College of Pathologists Dataset for histopathological reporting of colorectal cancer*. (Royal College of Pathologists, London, UK, 2017).

197. Cancer Research UK. Bowel cancer incidence statistics. *Cancer Research UK* (2010). Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence#heading-Four. (Accessed: 23rd February 2019)

198. Brenner, H. *et al.* Progress in colorectal cancer survival in Europe from the late 1980s to the early 21st century: The EUROCARE study. *Int. J. Cancer* **131**, 1649–1658 (2012).

199. Brouwer, N. P. M. *et al.* An overview of 25 years of incidence, treatment and outcome of colorectal cancer patients. *Int. J. Cancer* **143**, 2758–2766 (2018).

200. Fischer, J. *et al.* Outcome for stage II and III rectal and colon cancer equally good after treatment improvement over three decades. *Int. J. Colorectal Dis.* **30**, 797–806 (2015).

201. den Dulk, M. *et al.* Improved overall survival for patients with rectal cancer since 1990: The effects of TME surgery and pre-operative radiotherapy. *Eur. J. Cancer* **44**, 1710–1716 (2008).

202. Quirke, P. *et al.* Effect of the plane of surgery achieved on local recurrence in patients with operable rectal cancer: a prospective study using data from the MRC CR07 and NCIC-CTG CO16 randomised clinical trial. *Lancet* **373**, 821–828

(2009).

203. Petrelli, F. *et al.* Prognostic Survival Associated With Left-Sided vs Right-Sided Colon Cancer. *JAMA Oncol.* **3**, 211 (2017).

204. Cross, S. S., Bull, A. D. & Smith, J. H. Is there any justification for the routine examination of bowel resection margins in colorectal adenocarcinoma? *J. Clin. Pathol.* **42**, 1040–2 (1989).

205. Rocha, R. *et al.* Impact of bowel resection margins in node negative colon cancer. *Springerplus* **5**, 1959 (2016).

206. Rørvig, S. *et al.* Is the Longitudinal Margin of Carcinoma-Bearing Colon Resections a Neglected Parameter? *Clin. Colorectal Cancer* **13**, 68–72 (2014).

207. Amri, R., Bordeianou, L. G., Sylla, P. & Berger, D. L. Association of Radial Margin Positivity With Colon Cancer. *JAMA Surg.* **150**, 890 (2015).

208. Quirke, P., Durdey, P., Dixon, M. F. & Williams, N. S. Local recurrence of rectal adenocarcinoma due to inadequate surgical resection. Histopathological study of lateral tumour spread and surgical excision. *Lancet* **2**, 996–999 (1986).

209. Hohenberger, W., Weber, K., Matzel, K., Papadopoulos, T. & Merkel, S. Standardized surgery for colonic cancer: complete mesocolic excision and central ligation - technical notes and outcome. *Color. Dis.* **11**, 354–364 (2009).

210. Bertelsen, C. A. *et al.* Disease-free survival after complete mesocolic excision compared with conventional colon cancer surgery: a retrospective, population-based study. *Lancet. Oncol.* **16**, 161–8 (2015).

211. Brierley, J., Gospodarowicz, M. & Wittekind, C. *TNM classification of malignant tumours*. (Wiley-Blackwell, 2016).

212. Sobin, L. & Wittekind, C. *TNM Classification of Malignant Tumours*. (Wiley-Liss, 1997).

213. Compton, C., Fenoglio-Preiser, C. M., Pettigrew, N. & Fielding, L. P. American Joint Committee on Cancer prognostic factors consensus conference. *Cancer* **88**, 1739–1757 (2000).

214. Kornprat, P. *et al.* Value of Tumor Size as a Prognostic Variable in Colorectal

Cancer. *Am. J. Clin. Oncol.* **34**, 43–49 (2011).

215. Saha, S. *et al.* Tumor size predicts long-term survival in colon cancer: an analysis of the National Cancer Data Base. *Am. J. Surg.* **209**, 570–574 (2015).

216. Saha, S. *et al.* Tumor size as a prognostic factor for patients with colon cancer undergoing sentinel lymph node mapping and conventional surgery. *J. Clin. Oncol.* **31**, 546–546 (2013).

217. Bosman, F. T., Carneiro, F., Hruban, R. H. & Th eise, N. D. *WHO Classification of tumours of the colon and rectum*. (International Agency for Research on Cancer, 2010).

218. Kang, H., O'Connell, J. B., Maggard, M. A., Sack, J. & Ko, C. Y. A 10-Year Outcomes Evaluation of Mucinous and Signet-Ring Cell Carcinoma of the Colon and Rectum. *Dis. Colon Rectum* **48**, 1161–1168 (2005).

219. Verhulst, J., Ferdinande, L., Demetter, P. & Ceelen, W. Mucinous subtype as prognostic factor in colorectal cancer: a systematic review and meta-analysis. *J. Clin. Pathol.* **65**, 381–388 (2012).

220. Gonzalez, R. S., Cates, J. M. M. & Washington, K. Associations among histological characteristics and patient outcomes in colorectal carcinoma with a mucinous component. *Histopathology* **74**, 406–414 (2019).

221. Leopoldo, S. *et al.* Two Subtypes of Mucinous Adenocarcinoma of The Colorectum: Clinicopathological and Genetic Features. *Ann. Surg. Oncol.* **15**, 1429–1439 (2008).

222. Petersen, V. C., Baxter, K. J., Love, S. B. & Shepherd, N. a. Identification of objective pathological prognostic determinants and models of prognosis in Dukes' B colon cancer. *Gut* **51**, 65–69 (2002).

223. Halvorsen, T. B. & Seim, E. Degree of differentiation in colorectal adenocarcinomas: a multivariate analysis of the influence on survival. *J. Clin. Pathol.* **41**, 532–7 (1988).

224. Ueno, H., Murphy, J., Jass, J. R., Mochizuki, H. & Talbot, I. C. Tumour 'budding' as an index to estimate the potential of aggressiveness in rectal cancer. *Histopathology* **40**, 127–32 (2002).

225. Nakamura, T., Mitomi, H., Kanazawa, H., Ohkura, Y. & Watanabe, M. Tumor Budding as an Index to Identify High-Risk Patients with Stage II Colon Cancer. *Dis. Colon Rectum* **51**, 568–572 (2008).

226. Kanazawa, H. *et al.* Tumour budding at invasive margins and outcome in colorectal cancer. *Color. Dis.* **0**, 070621084454036-??? (2007).

227. Choi, H.-J. *et al.* Tumor budding as a prognostic marker in stage-III rectal carcinoma. *Int. J. Colorectal Dis.* **22**, 863–868 (2007).

228. Rogers, A. C. *et al.* Prognostic significance of tumor budding in rectal cancer biopsies before neoadjuvant therapy. *Mod. Pathol.* **27**, 156–162 (2014).

229. Quirke, P., Risio, M., Lambert, R., von Karsa, L. & Vieth, M. Quality assurance in pathology in colorectal cancer screening and diagnosis—European recommendations. *Virchows Arch.* **458**, 1–19 (2011).

230. Ogino, S. *et al.* Lymphocytic reaction to colorectal cancer is associated with longer survival, independent of lymph node count, microsatellite instability, and CpG island methylator phenotype. *Clin. Cancer Res.* **15**, 6412–20 (2009).

231. Roxburgh, C. S. D., Salmond, J. M., Horgan, P. G., Oien, K. A. & McMillan, D. C. Tumour inflammatory infiltrate predicts survival following curative resection for node-negative colorectal cancer. *Eur. J. Cancer* **45**, 2138–2145 (2009).

232. Komori, K. *et al.* Tumor necrosis in patients with TNM stage IV colorectal cancer without residual disease (R0 Status) is associated with a poor prognosis. *Anticancer Res.* **33**, 1099–105 (2013).

233. Richards, C. H. *et al.* The relationships between cellular components of the peritumoural inflammatory response, clinicopathological characteristics and survival in patients with primary operable colorectal cancer. *Br. J. Cancer* **106**, 2010–5 (2012).

234. Pollheimer, M. J. *et al.* Tumor necrosis is a new promising prognostic factor in colorectal cancer. *Hum. Pathol.* **41**, 1749–1757 (2010).

235. Nakamura, Y. *et al.* Importance of lymph vessels in gastric cancer: a prognostic indicator in general and a predictor for lymph node metastasis in early stage cancer. *J. Clin. Pathol.* **59**, 77–82 (2006).

236. Niakosari, F., Kahn, H. J., Marks, A. & From, L. Detection of Lymphatic Invasion in Primary Melanoma With Monoclonal Antibody D2-40. *Arch. Dermatol.* **141**, 440–4 (2005).

237. Arnaout-Alkarain, A., Kahn, H. J., Narod, S. A., Sun, P. A. & Marks, A. N. Significance of lymph vessel invasion identified by the endothelial lymphatic marker D2-40 in node negative breast cancer. *Mod. Pathol.* **20**, 183–191 (2007).

238. Gustafson, P. *et al.* Prognostic information in soft tissue sarcoma using tumour size, vascular invasion and microscopic tumour necrosis-the SIN-system. *Eur. J. Cancer* **39**, 1568–76 (2003).

239. Miles, E. W. The Spread of Cancer of the Rectum. *Lancet* **205**, 1218–1219 (1925).

240. Gibson, K. M., Chan, C., Chapuis, P. H., Dent, O. F. & Bokey, L. Mural and extramural venous invasion and prognosis in colorectal cancer. *Dis Colon Rectum* **57**, 916–926 (2014).

241. Minsky, B. D., Mies, C., Recht, A., Rich, T. A. & Chaffey, J. T. Resectable adenocarcinoma of the rectosigmoid and rectum: II The influence of blood vessel invasion. *Cancer* **61**, 1417–1424 (1988).

242. Kirsch, R. *et al.* Impact of an Elastin Stain on Detection and Interobserver Agreement Among Gastrointestinal and Nongastrointestinal Pathologists. *Am J Surg Pathol* **37**, 200–210 (2013).

243. Roxburgh, C. S. *et al.* Elastica staining for venous invasion results in superior prediction of cancer-specific survival in colorectal cancer. *Ann Surg* **252**, 989–997 (2010).

244. Vass, D. G., Ainsworth, R., Anderson, J. H., Murray, D. & Foulis, A. K. The value of an elastic tissue stain in detecting venous invasion in colorectal cancer. *J. Clin. Pathol.* **57**, 769–72 (2004).

245. Bosch, S. L., Teerenstra, S., de Wilt, J. H., Cunningham, C. & Nagtegaal, I. D. Predicting lymph node metastasis in pT1 colorectal cancer: a systematic review of risk factors providing rationale for therapy decisions. *Endoscopy* **45**, 827–834 (2013).

246. Knijn, N., Mogk, S. C., Teerenstra, S., Simmer, F. & Nagtegaal, I. D. Perineural

Invasion Is a Strong Prognostic Factor in Colorectal Cancer. *Am. J. Surg. Pathol.* **40**, 103–112 (2016).

247. Ueno, H. *et al.* Characterization of Perineural Invasion As a Component of Colorectal Cancer Staging. *Am. J. Surg. Pathol.* **37**, 1542–1549 (2013).

248. Maas, M. *et al.* Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *Lancet Oncol.* **11**, 835–844 (2010).

249. Capirci, C. *et al.* Prognostic Value of Pathologic Complete Response After Neoadjuvant Therapy in Locally Advanced Rectal Cancer: Long-Term Analysis of 566 ypCR Patients. *Int. J. Radiat. Oncol.* **72**, 99–107 (2008).

250. Chetty, R. *et al.* International study group on rectal cancer regression grading: interobserver variability with commonly used regression grading systems. *Hum. Pathol.* **43**, 1917–1923 (2012).

251. Amin, M. B. *et al. American Joint Cancer Committee Cancer Staging Manual*. (Springer International Publishing, 2017).

252. Ryan, R. *et al.* Pathological response following long-course neoadjuvant chemoradiotherapy for locally advanced rectal cancer. *Histopathology* **47**, 141–146 (2005).

253. McPhail, S., Johnson, S., Greenberg, D., Peake, M. & Rous, B. Stage at diagnosis and early mortality from cancer in England. *Br. J. Cancer* **112 Suppl**, S108-15 (2015).

254. NICE. *Capecitabine and oxaliplatin in the adjuvant treatment of stage III (dukes' c) colon cancer*. *National Institute for Health and Care Excellence (NICE) Guidleines ta100* (2006).

255. Augestad, K. M. *et al.* Metastatic spread pattern after curative colorectal cancer surgery. A retrospective, longitudinal analysis. *Cancer Epidemiol.* **39**, 734–744 (2015).

256. Mitry, E. *et al.* Epidemiology, management and prognosis of colorectal cancer with lung metastases: a 30-year population-based study. *Gut* **59**, 1383–8 (2010).

257. Tan, K. K., Lopes, G. D. L. & Sim, R. How uncommon are isolated lung metastases

in colorectal cancer? a review from database of 754 patients over 4 years. *J. Gastrointest. Surg.* **13**, 642–648 (2009).

258. Naxerova, K. *et al.* Origins of lymphatic and distant metastases in human colorectal cancer. *Science (80-.).* (2017). doi:10.1126/science.aai8515

259. Cerullo, G., Cassini, D. & Baldazzi, G. Application of Petersen Index Score for Dukes'B colorectal cancer in a population of 103 consecutive resected patients. *Updates Surg.* **64**, 95–99 (2012).

260. Yang, Y. *et al.* Prognostic Value of Perineural Invasion in Colorectal Cancer: A Meta-Analysis. *J. Gastrointest. Surg.* **19**, 1113–1122 (2015).

261. Betge, J. *et al.* Intramural and extramural vascular invasion in colorectal cancer: prognostic significance and quality of pathology reporting. *Cancer* **118**, 628–638 (2012).

262. Kaiser, a M., Kang, J.-C., Chan, L. S. & Beart, R. W. The prognostic impact of the time interval to recurrence for the mortality in recurrent colorectal cancer. *Colorectal Dis.* **8**, 696–703 (2006).

263. Huh, J. W. J., Kim, C. H., Lim, S. W. S., Kim, H. H. R. & Kim, Y. Y. J. Early recurrence in patients undergoing curative surgery for colorectal cancer: is it a predictor for poor overall survival? *Int. J. Colorectal Dis.* 1143–1149 (2013). doi:10.1007/s00384-013-1675-z

264. Samowitz, W. S., Curtin, K., Neuhausen, S., Schaffer, D. & Slattery, M. L. Prognostic implications of BAX and TGFBRII mutations in colon cancers with microsatellite instability. *Genes Chromosom. Cancer* **35**, 368–371 (2002).

265. Popat, S., Hubner, R. A. & Houlston, R. S. Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis. *J. Clin. Oncol.* **23**, 609–618 (2005).

266. Leach, F. S. *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75**, 1215–1225 (1993).

267. Boland, C. R., Koi, M., Chang, D. K. & Carethers, J. M. The biochemical basis of microsatellite instability and abnormal immunohistochemistry and clinical behavior in Lynch Syndrome: From bench to bedside. *Fam. Cancer* **7**, 41–52 (2008).

268. Van Cutsem, E. *et al.* ESMO consensus guidelines for the management of

patients with metastatic colorectal cancer. *Ann. Oncol.* mdw235 (2016). doi:10.1093/annonc/mdw235

269. Sanger, F., Nicklen, S. & AR, C. DNA sequencing with chain-terminating inhibitors. *Biochemistry* **74**, 5463–5467 (1977).

270. Tsiatis, A. C. *et al.* Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. *J Mol Diagn* **12**, 425–432 (2010).

271. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res* **15**, 1767–1776 (2005).

272. Check Hayden, E. Is the $1,000 genome for real? *Nature* (2014). doi:10.1038/nature.2014.14530

273. NICE. Evidence overview: KRAS mutation testing of tumours in adults with metastatic colorectal cancer. NICE diagnostic assessment programme. Available at: http://www.nice.org.uk/guidance/gid-dt14/resources/kras-mutation-testing-of-tumours-in-adults-with-metast. (2013).

274. Ronaghi M Pettersson B, Uhlén M, Nyrén P., K. S. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Anal Biochem* **242**, 84–89

275. Ogino, S. *et al.* Sensitive sequencing method for KRAS mutation detection by Pyrosequencing. *J Mol Diagn* **7**, 413–421 (2005).

276. Deschoolmeester, V. *et al.* KRAS mutation detection and prognostic potential in sporadic colorectal cancer using high-resolution melting analysis. *Br J Cancer* **103**, 1627–1636 (2010).

277. Whitehall, V. *et al.* A multicenter blinded study to evaluate KRAS mutation testing methodologies in the clinical setting. *J Mol Diagn* **11**, 543–552 (2009).

278. Taniguchi, K. *et al.* Quantitative detection of EGFR mutations in circulating tumor DNA derived from lung adenocarcinomas. *Clin Cancer Res* **17**, 7808–7815 (2011).

279. Johnson, K. A. The kinetic and chemical mechanism of high-fidelity DNA polymerases. *Biochim. Biophys. Acta* **1804**, 1041–8 (2010).

280. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nat. Protoc.* **7**, 1024–1041 (2012).

281. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

282. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).

283. Gollins, S. *et al.* Preoperative chemoradiation with capecitabine, irinotecan and cetuximab in rectal cancer: significance of pre-treatment and post-resection RAS mutations. *Br. J. Cancer* **117**, 1286–1294 (2017).

284. Hernan, I. *et al.* Detection of Genomic Variations in BRCA1 and BRCA2 Genes by Long-Range PCR and Next-Generation Sequencing. *J. Mol. Diagnostics* **14**, 286–293 (2012).

285. Ivády, G. *et al.* Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. *BMC Genomics* **19**, 158 (2018).

286. Templeton, A. R. *et al.* Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* **66**, 69–83 (2000).

287. Kosmidou, V. *et al.* Tumor Heterogeneity Revealed by KRAS, BRAF, and PIK3CA Pyrosequencing: KRAS and PIK3CA Intratumor Mutation Profile Differences and Their Therapeutic Implications. *Hum. Mutat.* **35**, 329–340 (2014).

288. Hardiman, K. M. *et al.* Intra-tumor genetic heterogeneity in rectal cancer. *Lab. Investig.* **96**, 4–15 (2015).

289. Maringe, C. *et al.* Stage at diagnosis and colorectal cancer survival in six high-income countries: A population-based study of patients diagnosed during 2000–2007. *Acta Oncol. (Madr).* **52**, 919–932 (2013).

290. West, N. P. *et al.* The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients. *Br. J. Cancer* **102**, 1519–23 (2010).

291. Laubert, T. *et al.* Stage-specific frequency and prognostic significance of aneuploidy in patients with sporadic colorectal cancer-a meta-analysis and current overview. *Int. J. Colorectal Dis.* 1015–1028 (2015). doi:10.1007/s00384-015-

2259-x

292. Bardi, G. *et al.* Cytogenetic analysis of 52 colorectal carcinomas - Non-random aberration pattern and correlation with pathologic parameters. *Int. J. Cancer* **55**, 422–428 (1993).

293. Ried, T. *et al.* Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes Chromosom. Cancer* **5**, 234–45 (1996).

294. Rooney, P. H. *et al.* Comparative genomic hybridization and chromosomal instability in solid tumours. *Br. J. Cancer* **80**, 862–73 (1999).

295. Douglas, E. J. *et al.* Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res* **64**, 4817–4825 (2004).

296. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (80-.).* **258**, 818–821 (1992).

297. Black, J. C. *et al.* H3K9/36me3 Demethylase KDM4A Promotes Site-Specific Copy Gain and Re-replication of Regions Amplified in Tumors. *Cell* **154**, (2013).

298. Kim, T. D., Shin, S., Berry, W. L., Oh, S. & Janknecht, R. The JMJD2A demethylase regulates apoptosis and proliferation in colon cancer cells. *J. Cell. Biochem.* **113**, 1368–1376 (2012).

299. Black, J. C. *et al.* Hypoxia drives transient site-specific copy gain and drug-resistant gene expression. *Genes Dev.* **29**, 1018–1031 (2015).

300. Mishra, S. & Whetstine, J. R. Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive. *Mol. Cell. Biol.* **36**, MCB.00652-15 (2016).

301. Rehen, S. K. *et al.* Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 13361–6 (2001).

302. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2008).

303. Diep, C. B. *et al.* Genome characteristics of primary carcinomas, local recurrences, carcinomatoses, and liver metastases from colorectal cancer

patients. *Mol Cancer* **3**, 6 (2004).

304. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).

305. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, 99–103 (2009).

306. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

307. Rechsteiner, M. *et al.* KRAS, BRAF, and TP53 deep sequencing for colorectal carcinoma patient diagnostics. *J Mol Diagn* **15**, 299–311 (2013).

308. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun* **3**, 811 (2012).

309. Sutton, K., Chambers, P., Taylor, G. & Quirke, P. Next generation sequencing mutation detection compared to pyrosequencing. *8th NCRI Cancer Conference* (2012).

310. Gusnanto, A., Wood, H. M., Pawitan, Y., Rabbitts, P. & Berri, S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28**, 40–47 (2012).

311. OncoScan® CNV FFPE Assay Kit. Affymetrix, part of Thermo Fisher Scientific. (2016). Available at: http://www.affymetrix.com/estore/catalog/prod960006/AFFY/OncoScan%26%23174%3B+CNV+FFPE+Assay+Kit#1_1. (Accessed: 22nd August 2016)

312. Lee, S. Y. *et al.* Comparative Genomic Analysis of Primary and Synchronous Metastatic Colorectal Cancers. *PLoS One* **9**, e90459 (2014).

313. Cina, S. J. Flow Cytometric Evaluation of DNA Degradation: A Predictor of Postmortem Interval? *Am. J. Forensic Med. Pathol.* **15**, 300–302 (1994).

314. Gilbert, M. T. P. *et al.* The Isolation of Nucleic Acids from Fixed, Paraffin-Embedded Tissues-Which Methods Are Useful When? *PLoS One* **2**, (2007).

315. Alanen, K. A., Joensuu, H. & Klemi, P. J. Autolysis Is a Potential Source of False

Aneuploid Peaks in Flow Cytometric DNA Histograms. *Cytometry* **10**, 417–425 (1989).

316. Yost, S. E. *et al.* Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **40**, 1–12 (2012).

317. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

318. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

319. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).

320. Newburger, D. E. *et al.* Genome evolution during progression to breast cancer. *Genome Res* **23**, 1097–1108 (2013).

321. Wood, H. M. *et al.* The clonal relationships between pre-cancer and cancer revealed by ultra-deep sequencing. *J Pathol* **237**, 296–306 (2015).

322. Schweiger, M. R. *et al.* Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. *PLoS One* **4**, 3–9 (2009).

323. Sottoriva, A. & Graham, T. A pan-cancer signature of neutral tumor evolution. *bioRxiv* 014894 (2015). doi:10.1101/014894

324. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression.

325. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).

326. Vermeulen, L. *et al.* Defining stem cell dynamics in models of intestinal tumor initiation. *Science* **342**, 995–8 (2013).

327. Marusyk, A. *et al.* Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54–58 (2014).

328. Chand, M. *et al.* Evidence for radiological and histopathological prognostic importance of detecting extramural venous invasion in rectal cancer: recommendations for radiology and histopathology reporting. *Color. Dis* **17**, 468–473 (2015).

329. Bertotti, A. *et al.* The genomic landscape of response to EGFR blockade in colorectal cancer. *Nature* **526**, 263–267 (2015).

330. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Rep* **6**, 514–527 (2014).

331. Genomics England. Genomics England: The 100,000 Genome Project. (2017). Available at: https://www.genomicsengland.co.uk/the-100000-genomes-project/. (Accessed: 19th August 2017)

332. Illumina. HiSeq X Instrument Performance Parameters. *Illumina, inc. webpage* (2016). Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf.

333. Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. & Seelow, D. Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. *PLoS One* **6**, e28240 (2011).

334. Samorodnitsky, E. *et al.* Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Hum. Mutat.* **36**, 903–914 (2015).

335. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–9 (2009).

336. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

337. Cancer Genome Project. CaVEMan: SNV expectation maximisation based mutation calling algorithm aimed at detecting somatic mutations in paired (tumour/normal) cancer samples. Supports both bam and cram format via htslib. (2017).

338. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions

from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

339. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

340. Ma, S. *et al.* SPARCoC: a new framework for molecular pattern discovery and cancer gene identification. *PLoS One* **10**, e0117135 (2015).

341. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14863–8 (1998).

342. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).

343. Behjati, S. *et al.* Recurrent PTPRB and PLCG1 mutations in angiosarcoma. *Nat Genet* **46**, 376–379 (2014).

344. Tarpey, P. S. *et al.* Frequent mutation of the major cartilage collagen gene COL2A1 in chondrosarcoma. *Nat Genet* **45**, 923–926 (2013).

345. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).

346. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

347. Broniscer, A. *et al.* Prospective collection of tissue samples at autopsy in children with diffuse intrinsic pontine glioma. *Cancer* **116**, 4632–7 (2010).

348. Sloothaak, D. A. M. *et al.* Intraperitoneal chemotherapy as adjuvant treatment to prevent peritoneal carcinomatosis of colorectal cancer origin: a systematic review. *Br. J. Cancer* **111**, 1112–1121 (2014).

349. Cavaliere, F. *et al.* Treatment of peritoneal carcinomatosis with intent to cure. *J. Surg. Oncol.* **74**, 41–4 (2000).

350. Franko, J. *et al.* Cytoreductive surgery and hyperthermic intraperitoneal chemoperfusion versus systemic chemotherapy alone for colorectal peritoneal carcinomatosis. *Cancer* **116**, 3756–3762 (2010).

351. Sloothaak, D. A. M. *et al.* Feasibility of adjuvant laparoscopic hyperthermic intraperitoneal chemotherapy in a short stay setting in patients with colorectal cancer at high risk of peritoneal carcinomatosis. *Eur. J. Surg. Oncol.* **40**, 1453–1458 (2014).

352. Boland, C. R., Shin, S. K. & Goel, A. Promoter methylation in the genesis of gastrointestinal cancer. *Yonsei Med. J.* **50**, 309–21 (2009).

353. Malkhosyan, S., Rampino, N., Yamamoto, H. & Perucho, M. Frameshift mutator mutations. *Nature* **382**, 499–500 (1996).

354. Higgins, G. S. *et al.* Overexpression of POLQ confers a poor prognosis in early breast cancer patients. *Oncotarget* **1**, 175–84 (2010).

355. Pillaire, M.-J. *et al.* A 'DNA replication' signature of progression and negative outcome in colorectal cancer. *Oncogene* **29**, 876–887 (2010).

356. Allera-Moreau, C. *et al.* DNA replication stress response involving PLK1, CDC6, POLQ, RAD51 and CLASPIN upregulation prognoses the outcome of early/mid-stage non-small cell lung cancer patients. *Oncogenesis* **1**, e30 (2012).

357. Wood, R. D. & Doublié, S. DNA polymerase θ (POLQ), double-strand break repair, and cancer. *DNA Repair (Amst).* **44**, 22–32 (2016).

358. Baldacci, G., Hoffmann, J.-S. & Cadoret, J.-C. Impact of the DNA polymerase Theta on the DNA replication program. *Genomics data* **3**, 90–3 (2015).

359. Yousefzadeh, M. J. & Wood, R. D. DNA polymerase POLQ and cellular defense against DNA damage. *DNA Repair* **12**, 1–9 (2013).

360. Hogg, M., Seki, M., Wood, R. D., Doublié, S. & Wallace, S. S. Lesion bypass activity of DNA polymerase θ (POLQ) is an intrinsic property of the pol domain and depends on unique sequence inserts. *J. Mol. Biol.* **405**, 642–652 (2011).

361. Lin, E. I. *et al.* Mutational profiling of colorectal cancers with microsatellite instability. *Oncotarget* **6**, 42334–44 (2015).

362. McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).

363. Therkildsen, C., Bergmann, T. K., Henrichsen-Schnack, T., Ladelund, S. & Nilbert,

M. The predictive value of KRAS, NRAS, BRAF, PIK3CA and PTEN for anti-EGFR treatment in metastatic colorectal cancer: A systematic review and meta-analysis. *Acta Oncol. (Madr).* **53**, 852–864 (2014).

364. Zhu, L. *et al.* Prognostic Role of BRAF Mutation in Stage II/III Colorectal Cancer Receiving Curative Resection and Adjuvant Chemotherapy: A Meta-Analysis Based on Randomized Clinical Trials. *PLoS One* **11**, e0154795 (2016).

365. Korphaisarn, K. *et al.* FBXW7 missense mutation: a novel negative prognostic factor in metastatic colorectal adenocarcinoma. *Oncotarget* **8**, 39268–39279 (2017).

366. Talbot, I. C. *et al.* Invasion of veins by carcinoma of rectum: method of detection, histological features and significance. *Histopathology* **5**, 141–163 (1981).

367. Dekker, J. W. T., Peeters, K. C., Putter, H., Vahrmeijer, A. L. & van de Velde, C. J. H. Metastatic lymph node ratio in stage III rectal cancer; prognostic significance in addition to the 7th edition of the TNM classification. *Eur. J. Surg. Oncol.* **36**, 1180–1186 (2010).

368. Martin, E. S. *et al.* Common and distinct genomic events in sporadic colorectal cancer and diverse cancer types. *Cancer Res.* **67**, 10736–10743 (2007).

369. Knösel, T. *et al.* Chromosomal alterations during lymphatic and liver metastasis formation of colorectal cancer. *Neoplasia* **6**, 23–8 (2004).

370. Pezza, J. A., Kucera, R. & Sun, L. Polymerase Fidelity: What is it and what does it mean for your PCR. *New England Biolabs Website* (2012). Available at: https://www.neb.com/tools-and-resources/feature-articles/polymerase-fidelity-what-is-it-and-what-does-it-mean-for-your-pcr. (Accessed: 25th September 2016)

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

372. Humphries, A. *et al.* Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. (2013). doi:10.1073/pnas.1220353110

373. Losi, L. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. **26**, (2005).

374. Bernier-Latmani, J. & Petrova, T. V. Intestinal lymphatic vasculature: structure, mechanisms and functions. *Nat. Rev. Gastroenterol. Hepatol.* **14**, 510–526 (2017).

375. Joung, J.-G. *et al.* Personalized Medicine and Imaging Tumor Heterogeneity Predicts Metastatic Potential in Colorectal Cancer. (2017). doi:10.1158/1078-0432.CCR-17-0306

376. Chan, D. L. H. *et al.* Epidermal growth factor receptor (EGFR) inhibitors for metastatic colorectal cancer. *Cochrane Database Syst. Rev.* **6**, CD007047 (2017).

377. Bekaii-Saab, T. S. *et al.* A phase I trial of paclitaxel and trastuzumab in combination with interleukin-12 in patients with HER2/neu-expressing malignancies. *Mol. Cancer Ther.* **8**, 2983–2991 (2009).

378. Adams, R. *et al.* Inhibition of EGFR, HER2, and HER3 signalling in patients with colorectal cancer wild-type for BRAF, PIK3CA, KRAS, and NRAS (FOCUS4-D): a phase 2-3 randomised trial. *lancet. Gastroenterol. Hepatol.* **3**, 162–171 (2018).

379. Sartore-Bianchi, A. *et al.* Dual-targeted therapy with trastuzumab and lapatinib in treatment-refractory, KRAS codon 12/13 wild-type, HER2-positive metastatic colorectal cancer (HERACLES): a proof-of-concept, multicentre, open-label, phase 2 trial. *Lancet Oncol.* (2016). doi:10.1016/S1470-2045(16)00150-9

380. Meric-Bernstam, F. *et al.* Pertuzumab plus trastuzumab for HER2-amplified metastatic colorectal cancer (MyPathway): an updated report from a multicentre, open-label, phase 2a, multiple basket study. *Lancet. Oncol.* **20**, 518–530 (2019).

381. Embuscado, E. E. *et al.* Immortalizing the complexity of cancer metastasis: genetic features of lethal metastatic pancreatic cancer obtained from rapid autopsy. *Cancer Biol. Ther.* **4**, 548–54 (2005).

382. DiSibio, G. & French, S. W. Metastatic patterns of cancers : Results from a large autopsy study. *Arch. Pathol. Lab. Med.* **132**, 931–939 (2008).

383. Kim, M.-J. *et al.* Different metastatic pattern according to the KRAS mutational status and site-specific discordance of KRAS status in patients with colorectal cancer. *BMC Cancer* **12**, 347 (2012).

384. Richman, S. D., Fairley, J., Butler, R. & Deans, Z. C. RAS screening in colorectal

cancer: a comprehensive analysis of the results from the UK NEQAS colorectal cancer external quality assurance schemes (2009-2016). *Virchows Arch.* **471**, 721–729 (2017).

385. Moorcraft, S. Y. *et al.* Molecular profiling of colorectal pulmonary metastases and primary tumours: implications for targeted treatment. *Oncotarget* **8**, 64999–65008 (2017).

386. Santini, D. *et al.* High Concordance of KRAS Status Between Primary Colorectal Tumors and Related Metastatic Sites: Implications for Clinical Practice. *Oncologist* **13**, 1270–1275 (2008).

387. Pereira, A. A. L. *et al.* Association between KRAS mutation and lung metastasis in advanced colorectal cancer. *Br. J. Cancer* **112**, 424–428 (2015).

388. Russo, M. *et al.* Tumor Heterogeneity and Lesion-Specific Response to Targeted Therapy in Colorectal Cancer. *Cancer Discov.* **6**, 147–153 (2016).

389. Aljehani, M. A. *et al.* Association of Primary Tumor Site With Mortality in Patients Receiving Bevacizumab and Cetuximab for Metastatic Colorectal Cancer. *JAMA Surg.* **153**, 60 (2018).

390. Boeckx, N. *et al.* Primary tumor sidedness has an impact on prognosis and treatment outcome in metastatic colorectal cancer: results from two randomized first-line panitumumab studies. *Ann. Oncol.* **28**, 1862–1868 (2017).

391. Seligmann, J. F. *et al.* Combined Epiregulin and Amphiregulin Expression Levels as a Predictive Biomarker for Panitumumab Therapy Benefit or Lack of Benefit in Patients With *RAS* Wild-Type Advanced Colorectal Cancer. *JAMA Oncol.* **2**, 633 (2016).

392. Cha, Y. *et al.* Association of CHFR Promoter Methylation with Treatment Outcomes of Irinotecan-Based Chemotherapy in Metastatic Colorectal Cancer. *Neoplasia* **21**, 146–155 (2019).

393. Ebert, M. P. A. *et al.* TFAP2E–DKK4 and Chemoresistance in Colorectal Cancer. *N. Engl. J. Med.* **366**, 44–53 (2012).

394. Korphaisarn, K. *et al.* Arginine methylation of EGFR: a new biomarker for predicting resistance to anti-EGFR treatment. *Am. J. Cancer Res.* **7**, 2587–2599

(2017).

395. Bormann, F. *et al.* Epigenetic regulation of Amphiregulin and Epiregulin in colorectal cancer. *Int. J. Cancer* **144**, 569–581 (2019).

396. Sunakawa, Y. *et al.* Combined assessment of EGFR-related molecules to predict outcome of 1st-line cetuximab-containing chemotherapy for metastatic colorectal cancer. *Cancer Biol. Ther.* **17**, 751–9 (2016).

397. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).

398. Kangaspeska, S. *et al.* Transient cyclical methylation of promoter DNA. *Nature* **452**, 112–115 (2008).

399. Hur, K. *et al.* MicroRNA-200c modulates epithelial-to-mesenchymal transition (EMT) in human colorectal cancer metastasis. *Gut* **62**, 1315–26 (2013).