Single-cell multiomics profiling in the study of colorectal cancer evolution

By

Silvia Uhunoma Ogbeide Igbinoba

Thesis submitted to the University of East Anglia for the Degree of *Doctor of Philosophy*

September 2024

“The Lord is your strength; the sky is your limit.”

— My mother

Declaration

I hereby declare that this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. Specific details of work done in collaboration are given at the start of relevant chapters and in the methods section. The contents of this thesis are not substantially the same as any that has been submitted or is being submitted for a degree. The total length of the main body of this thesis including figure legends is 74,201 and therefore does not exceed the limit of 100,000 words.

Silvia Ogbeide

September 2024

Acknowledgments

First, I would like to express my deepest gratitude to my primary supervisor, Dr Iain Macaulay, for giving me the opportunity to carry out this PhD project in his group and for his unwavering support throughout. When I began searching for a PhD, my sole focus was on finding a project I was passionate about. It never occurred to me how equally important it is to have a supervisor who consistently supports, encourages, and stays positive during challenging times.

I would also like to thank my secondary supervisor, Dr Wilfried Haerty, for providing a different perspective and sparking an unexpected interest in bioinformatics. My only regret is not having more meetings with him initially, as I was mistakenly intimidated. In reality, he is a wonderful person.

I am grateful to my team members: those who have left, those who have stayed, and the newcomers. Thank you for fostering a supportive, kind, and relaxed environment where I never felt embarrassed for making mistakes. I would also like to thank the Earlham Institute for creating a peaceful and welcoming atmosphere for all PhD students and staff. I once assumed that every PhD student experienced the same supportive environment I did, but I now realise how fortunate I am to have been surrounded by such a wonderful community. I would also like to acknowledge the scientists involved in the Accelerator Award project, both here in the UK and in Milan. I look forward to future collaborations with equally wonderful people.

I am also grateful to my housemates during the toughest months of the COVID-19 pandemic. Despite the challenges of living with six people, I did not have a single bad moment. A special thanks to my former housemate Dixie, who shared the PhD journey with me through countless memes and reels that humorously depicted our struggles. Thanks for introducing me to the best show in the world, “The Office,” which brought much-needed laughter. To my current housemate, Sanja, God bless you for arriving when I needed you most, especially during the final year of my PhD.

Lastly, I want to thank my family and friends. To all my siblings, especially my sister Carmen, for providing me with a place to think and reflect when I needed it most. To my nieces and nephews for their joyful chaos that helped me switch off and relax. Most importantly, I am deeply grateful to my parents for their countless sacrifices, which have allowed me to be where I am today. As two Nigerian parents who emigrated to Spain and then moved to the UK, they have faced many challenges, but their resilience has always been an inspiration to me. Their strength taught me to persevere and never give up, even in difficult times. A special thank you to my mom, who always ends our calls with “I love you,” “I’m proud of you,” “The sky is your limit,” and “The Lord is your strength.” I hope to be as sweet and supportive to my own children as she has always been to me.

Abstracts

Colorectal cancer (CRC) remains a significant global health concern, with metastatic CRC (mCRC) presenting particularly poor prognoses due to the high failure rate of existing treatments, including targeted therapies. The emergence of drug resistance severely undermines the efficacy of these therapies, highlighting the urgent need for new approaches to overcome or prevent resistance to improve patient outcomes.

This study aimed to investigate the characteristics of mCRC patient-derived organoids (PDOs) in response to two AKT inhibitors (AKTi): MK-2206 and AZD5363/capivasertib. The primary focus was on characterising the PDOs after they developed resistance to these inhibitors, with the goal of uncovering resistance mechanisms at both transcriptional and genomic levels. To achieve this, single-cell genome and transcriptome sequencing (G&T-seq) was applied to MK2206-resistant, AZD5363-resistant, and control mCRC organoids. This method allowed for extensive profiling of mCRC cells, and provided valuable insights into the relationship between genomic alterations and their effects on the transcriptome of resistant cells.

Using this single-cell multiomics approach, the research identified genes potentially associated with resistance to AKT inhibition, implicating various processes such as energy metabolism, extracellular matrix remodelling, and immune response regulation. A key finding was the expansion of a pre-existing resistant subclone, characterised by specific copy number alterations (CNAs) on chromosomes (chr) 2 and 5, in both AKTi-resistant organoids. This suggests a potential drug-agnostic resistance mechanism, with the same subclone being selected for under different selective pressures. Furthermore, a direct correlation between CNAs on chr2 and chr5 and gene expression was evident in MK-2206-resistant cells. However, this correlation was not consistently observed for chr5 in AZD5363-resistant cells, suggesting that compensatory mechanisms could have modulated gene dosage effects in this organoid.

By integrating genomic and transcriptomic datasets, researchers can identify altered genes at the DNA level and correlate these changes with gene expression patterns that drive malignant processes. This approach underscores the importance of considering both molecular layers to fully understand cancer evolution and inform the development of new treatments.

Table of contents

Abbreviations

Abbreviation	Description
5-FU	5-Fluorouracil
Akti	AKT inhibitor
CDPS	Cap-dependent protein synthesis
CF	Folinic acid
chr	Chromosome
CIN	Chromosomal instability
CMS	Consensus molecular subtypes
CNA	Copy number alteration
CNV	Copy number variation
CRC	Colorectal cancer
CSC	Cancer stem cell
DEG	Differentially expressed gene
DGE	Differential gene expression
EGFR	Epidermal growth factor receptor
FOrMAT	Feasibility of a Molecular Characterisation Approach to Treatment
G&T-seq	Genome and transcriptome sequencing
gDNA	Genomic DNA
GEMM	Genetically engineered mouse model
GI	Gastrointestinal
GOF	Gain of function
GP	Genomics pipeline
GSEA	Gene Set Enrichment Analysis
GTF	Gene transfer format
HGCA	Human Gut Cell Atlas
IBD	Inflammatory bowel disease
ICR	Institute of Cancer Research
IGV	Integrative Genomics Viewer
IOD	Index of dispersion
ISC	Intestinal stem cells
ITH	Intratumor heterogeneity
LOF	Loss of function
LOH	Loss of heterozygosity
mAbs	Monoclonal antibodies
MAD	Mean absolute deviation
MAPD	Median absolute pairwise difference
mCRC	Metastatic colorectal cancer
MDR	Multidrug resistance
MDSC	Myeloid-derived suppressor cell
miRNA	MicroRNA
MMR	Mismatch repair
mRNA	Messenger RNA
MSI	Microsatellite instability
MSigDB	Molecular Signatures Database
MSS	Microsatellite stable
NAC	Neoadjuvant chemotherapy

NCI	National Cancer Institute
NF-H_2O	Nuclease-free water
ORA	Over-representation analysis
PAP	Primary analysis pipeline
PCA	Principal component analysis
PDO	Patient-derived organoid
PDX	Patient-derived xenograft
PE	paired-end
PH	Pleckstrin homology
PKA	cAMP-dependent protein kinase
QC	Quality control
RNA-seq	RNA sequencing
RPCA	Robust Principal Component Analysis
RT	Reverse transcription
scRNA-seq	Single-cell RNA sequencing
sc-seq	Single-cell sequencing
scWGS	Single cell whole genome sequencing
SNV	Single-nucleotide variant
TA cells	Transit-amplifying cells
TAM	Tumour-associated macrophage
TME	Tumour microenvironment
Treg	T regulatory cell
TSG	Tumour suppressor gene
uBAM	Unaligned BAM
UMAP	Uniform Manifold Approximation and Projectio
WGD	Whole-genome doubling
WGS	Whole genome sequencing
WTA	Whole transcriptome amplification

List of figures

List of tables

Chapter 1

Introduction

1.1 Molecular characteristics of colorectal cancer

Colorectal cancer (CRC) is the third most commonly diagnosed cancer type worldwide for both men and women, and it is the second leading cause of cancer-related deaths (1). Between 2016 and 2018, the UK reported approximately 42,900 new cases of CRC annually, with 23,900 cases in males and 19,000 in females, which translates to nearly 120 cases diagnosed daily (2). The incidence of CRC in the UK is expected to rise to 47,000 new cases per year by 2040 (2), highlighting the critical need for a deeper understanding of the molecular characteristics of the disease, early detection, and the development of more effective treatment strategies.

1.1.1 Molecular classification

Nearly all cases of CRC originate from the malignant transformation of colorectal polyps (3). A polyp is a benign mass of glandular epithelial cells that arises from the mucosal layer and protrudes into the lumen of the gastrointestinal (GI) tract, predominantly in the descending colon and rectum, but can also occur in the genitourinary or respiratory tracts (4). Polyps are usually asymptomatic and often detected during colorectal endoscopic (colonoscopy) screenings. By the age of 50, approximately 30% of the population will develop these lesions, which include non-neoplastic inflammatory polyps, hamartomatous polyps, sessile serrated lesions, and adenomatous polyps (also known as adenomas). Although polyps can undergo changes that may lead to cancer over time, not all polyps exhibit malignant potential. The unpredictable progression of polyps underscores the importance of regular monitoring and, when necessary, their removal to prevent the development of CRC (4).

The molecular classification of CRC reflects the underlying mechanisms of tumorigenesis and is instrumental for determining the clinical, pathological, and biological characteristics of the disease (5). Currently, CRC can be categorised into distinct molecular subtypes based on genetic, epigenetic, and transcriptomic characteristics.

The Cancer Genome Atlas (TCGA) project initially classified CRC using a combination of array-based and sequencing technologies (6). This classification employed data from exome sequencing, DNA copy number analyses, promoter methylation profiles, as well as messenger RNA (mRNA) and microRNA (miRNA) expression studies. In the analysis of 224 CRC and matched normal samples, 15% were identified as hypermutated tumours, with 77% of these exhibiting a high frequency of microsatellite instability (MSI). Microsatellites, also known as Simple Sequence Repeats (SSRs) or Short Tandem Repeats (STRs), are repetitive sequences of 1-6 nucleotides in length found throughout the genome in both coding and non-coding regions (7). Due to their repetitive nature, microsatellites are especially susceptible to replication

errors caused by defective DNA mismatch repair (MMR) pathways (6, 7). This characteristic makes microsatellite status a critical marker for assessing MMR deficiency in CRCs, particularly involving the MMR genes *MLH1* and *MSH2* (8, 9). The remaining hypermutated tumours were identified as harbouring somatic mutations in DNA MMR pathways and DNA Polymerase Epsilon (*POLE*) (6).

The next group of CRCs identified was characterised by the absence of MSI. Instead, these microsatellite stable (MSS) CRCs exhibited extensive chromosomal rearrangements, including a high rate of gains and losses of chromosomal segments, known as copy number alterations (CNAs), leading to aneuploid genomes (6, 9). Additionally, this group demonstrated a significant loss of heterozygosity (LOH) events, resulting in the loss of function (LOF) of critical gene products (6, 9). These MSS CRCs, marked by chromosomal instability (CIN) (9), are typically recognised by an accumulation of mutations that alter or neutralise the function of genes that regulate cell proliferation, differentiation, senescence, and apoptosis, collectively known as tumour-suppressor genes (TSG) and proto-oncogenes (10, 11). Examples of affected TSGs in CRCs include *APC*, *TP53* and *SMAD4*, as well proto-oncogenes such as *KRAS* and *PIK3CA* (12).

Mutations in driver genes disrupt pathways critical for CRC initiation and progression, including Wingless (Wnt) signalling, RAS/MAPK, PI3K/AKT pathways, DNA MMR pathways and those involving *TGF-β* and *TP53* (Figure 1.1) (6, 9). Moreover, these mutations confer advantageous phenotypes, known as hallmarks, to neoplastic cells. Initially proposed by Hanahan and Weinberg in 2000 and later updated in 2011 and 2022, the hallmarks of cancer encapsulate essential processes that enable neoplastic cells to achieve malignant potential (13, 14). These hallmarks include the ability to sustain proliferative signalling, evade growth suppressors, reprogram cellular metabolism, avoid immune destruction, exhibit phenotypic plasticity, undergo disrupted cell differentiation, thrive in a permissive tumour microenvironment, and undergo epigenetic reprogramming.

Another subset of CRCs showed hypermethylation of CpG islands at promoter regions, leading to the transcriptional silencing of critical genes, including tumour suppressors. CRCs with the CpG island methylator phenotype (CIMP) often exhibit high microsatellite instability (MSI-high) as a result of their impact on DNA MMR genes (6, 9).

A more recent classification system divides CRCs into four consensus molecular subtypes (CMS), each characterised by distinct gene expression patterns (15). CMS1 (MSI immune) are MSI and CIMP high tumours featuring *MLH1* silencing, strong immune infiltration, and activation (15). Despite this, CMS1 CRCs are associated with a poor prognosis following tumour

relapse. CMS2 (canonical) CRCs are characterised by high expression of epithelial markers and exhibit more CIN than other subtypes. Additionally, CMS2 CRCs show widespread activation of Wnt and *MYC* signalling pathways, which are critical for regulating epithelial cell proliferation and survival. This subtype typically represents more differentiated tumours with features that closely resemble the normal epithelium. Although CMS3 (metabolic) CRCs also exhibit epithelial signatures, they are distinguished from CMS2 by the dysregulation of genes involved in metabolic processes, the presence of *KRAS* mutations, and low levels of CIMP. CMS2 and CMS3 CRCs have higher survival rates after relapse than the other CMS. Lastly, CMS4 (mesenchymal) CRCs frequently show activation of the TGF-β signalling and prominent expression of genes involved in inflammation, extracellular matrix remodelling, stromal invasion, and angiogenesis. Among all the consensus molecular subtypes, CMS4 CRCs are associated with the poorest prognosis (15).

CRC is a multifaceted disease, with a pathogenesis shaped by risk factors extending beyond these molecular classifications. CRCs are commonly categorised into sporadic, familial, or hereditary types (9). Approximately 70% of CRCs are sporadic, with no familial history or apparent genetic predisposition, and are primarily linked to environmental and lifestyle factors such as advanced age, sedentary lifestyle, obesity, an unhealthy diet, smoking status, and heavy alcohol consumption (15). CIN is a hallmark of 85% of sporadic cases, while the remainder predominantly exhibit MSI phenotypes (9). Sporadic MSI CRCs frequently feature loss of DNA MMR activity, often due to *MLH1* silencing (15).

Familial CRC involves individuals with a family history of the disease, indicating a potential genetic predisposition (15). On the other hand, inherited or genetic cases account for 5-10% of all cases and are categorised based on the presence or absence of precursor lesions known as colonic polyps, with specific conditions linked to each category. Diseases with polyposis following an autosomal dominant pattern of inheritance include Familial Adenomatous Polyposis (FAP) and hamartomatous polyposis syndromes (HPS), which encompass Peutz-Jeghers syndrome, Juvenile polyposis syndrome, and PTEN hamartoma tumour syndrome (15, 16). Notably, FAP is characterised by germline mutations in *APC,* with 95% of known mutations being frameshift or nonsense mutations, leading to a truncated protein (17). Conversely, hereditary polyposis CRC with autosomal recessive inheritance include MUTYH Associated Polyposis (MAP) (15, 17).

On the other hand, hereditary nonpolyposis CRC (HNPCC), also known as Lynch syndrome, is characterised by a lack of polyps. About 95% of HNPCC cases show a high frequency of MSI due to germline mutations in MMR genes. Specifically, mutations in *MLH1* or *MSH2* are associated

with a CRC risk of 70-80%, while mutations in *MSH6* or *PMS2* are associated with a comparatively lower risk, ranging from 25-60% (15, 18).

Other non-genetic risk factors associated with CRC include chronic inflammatory bowel diseases (IBDs) like Crohn's disease or ulcerative colitis (15).

1.1. Molecular characteristics of colorectal cancer

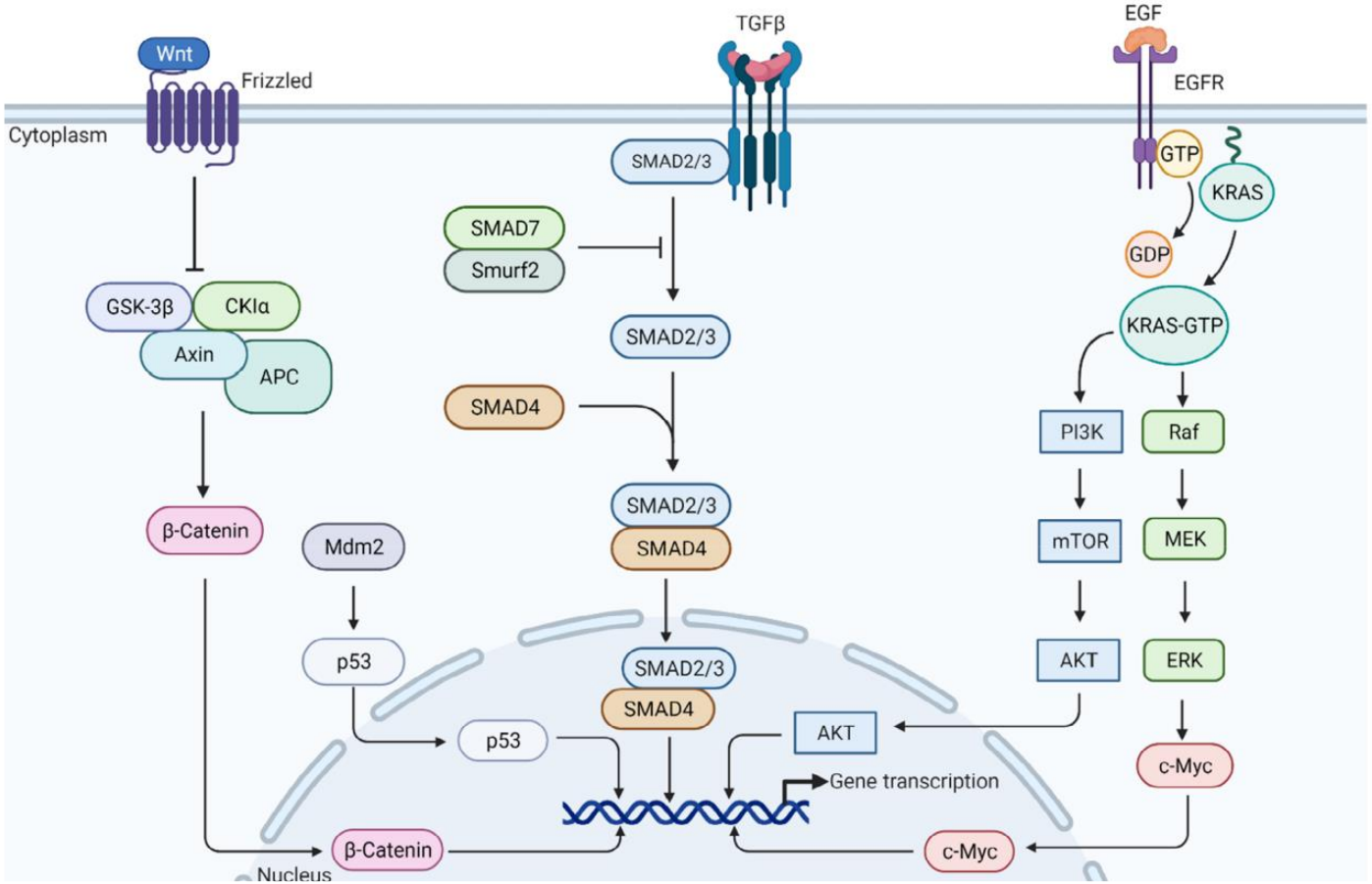


Figure 1.1. Major signalling pathways involved in CRC carcinogenesis.

Pathways involved in the development and progression of colorectal cancer: Wnt/β-catenin, TGF-β, MAPK and PI3K/AKT signaling. Figure reproduced from (19).

1.1. Molecular characteristics of colorectal cancer

1.1.2 Wnt/β-catenin signalling pathway in the development of CRC

Among the major signalling pathways implicated in CRC carcinogenesis (Figure 1.1), Wnt signalling is one of the most important as it regulates the homeostatic equilibrium between stemness, proliferation, and differentiation of normal ISCs at the intestinal crypt base (20). ISCs are responsible for the continuous renewal of the intestinal wall (21). The gradient of Wnt signalling, which is the highest at the crypt base, is essential for sustaining the undifferentiated state of ISCs and promoting their proliferation (22). As the progeny produced by ISCs begin to migrate upwards along the crypt-villus axis, they encounter a decreasing gradient of Wnt signalling, which promotes the differentiation of these cells into various specialised intestinal cell types, such as absorptive enterocytes, goblet cells, Paneth cells, and enteroendocrine cells (23).

The constitutive activation of canonical Wnt signalling pathway—also known as Wnt/β-catenin due to the critical role of β-catenin within the pathway—by a mutated *APC* or through other mechanisms is the initiating and rate-limiting step in the pathogenesis of CRC progression (Figure 1.2) (24). In the absence of Wnt ligands such as Wnt3a, β-catenin is kept at low levels in the cytoplasm through continuous degradation (25). This degradation is mediated by a "multi-protein" destruction complex that includes two anchoring proteins, APC and Axin1/2, and two kinases, GSK-3β and CK1α (25, 26). Upon binding to the destruction complex, β-catenin is sequentially phosphorylated, first by CK1α (at Ser45) and then by GSK-3β (at Ser33/37/Thr41), then ubiquitinated by the β-TrCP ubiquitin E3 ligase complex, marking it for ubiquitination and subsequent proteasomal degradation (25, 27-29). In the absence of Wnt signalling, T-cell factor/lymphoid enhancer factor (TCF/LEF) transcription factors, which are the nuclear mediators of Wnt signalling, are bound to Groucho/TLE co-repressors (27). This interaction keeps target genes repressed. Groucho/TLE proteins do not directly bind to DNA but are recruited to target genes through their interaction with DNA-bound TCF/LEF (20, 27, 29). The Groucho–TCF/TLE complex recruits other components of the transcriptional repression machinery, including histone deacetylases (HDACs), leading to a closed chromatin conformation and suppression of Wnt target gene expression (27). Therefore, as Wnt target genes are repressed, *APC* inhibits the transition of stem cells from G0/G1 to the S phase in unstimulated stem cells (15, 29).

Secretion of Wnt ligands by Paneth cells and subepithelial myofibroblasts in the stem cell compartment activates canonical Wnt signalling (30). The binding of Wnt ligands to the Frizzled (FZD) and LRP5 or LRP6 co-receptors brings these receptors in close proximity. This event leads to the phosphorylation of the intracellular domain of LRP5/6 by GSK-3β and CK1 α (26). Phosphorylated LRP5/6 then serves as a docking site for the recruitment of other

signalling proteins. For one, Axin1/2 translocation to the membrane is mediated by the phosphorylated LRP5/6 and by the protein Dishevelled (DVL), which is activated upon FZD receptor activation (20, 27, 29). The translocation of Axin to the membrane and its interaction with LRP5/6 disrupts the β-catenin destruction complex. As a result, β-catenin is no longer efficiently phosphorylated by GSK-3 and CK1α, which prevents its recognition and ubiquitination by β-TrCP. With the destruction complex inhibited, newly synthesised β-catenin accumulates in the cytoplasm and subsequently translocates to the nucleus. In the nucleus, β-catenin binds to TCF/LEF transcription factors bound to Wnt-responsive enhancers, displacing Groucho/TLE co-repressors from the complex. The β-catenin-TCF/LEF complex then recruits co-activators, including histone acetyltransferases (HATs), which modify chromatin to a more open configuration conducive to transcription. The displacement of Groucho/TLE by β-catenin converts TCF/LEF from transcriptional repressors into activators, leading to the transcription of Wnt target genes that drive cell proliferation, differentiation, and survival (20, 27, 29).

In addition, Wnt signalling is enhanced by R-spondins (RSPOs), which are secreted by cells in the intestinal crypt base (28). RSPOs bind to Leucine-rich repeat-containing G-protein coupled receptors such as LGR5 or LGR6, expressed in stem cells and tissues where Wnt signalling is crucial. The binding of RSPOs to LGR5/6 prevents the internalisation and subsequent degradation of the FZD and LRP5/6 co-receptors by the E3 ubiquitin ligases ZNRF3 and RNF43. Thus, binding of RSPOs to LGR5/6 inhibits the action of ZNRF3 and RNF43, leading to an increase in the number of Wnt receptors (Frizzled and LRP5/6) available on the cell surface, which ultimately permits the inhibition of the destruction complex and accumulation of β-catenin in the cytoplasm (28).

Contrarily, the canonical Wnt signalling pathway is negatively regulated by the Dickkopf (DKK) family of proteins. DKK1/3/4 inhibit canonical Wnt signalling by binding to the extracellular domains of LRP5/6, thus preventing it from forming a functional receptor complex with FZD (31). Additionally, the binding of DKK1/3/4 to LRP5/6 can lead to the formation of inactive tertiary complexes with Kremen 1 and 2 proteins. Kremen proteins function as high-affinity receptors for DKK1/2/3, and their interaction facilitates the removal of LRP5/6 from the cell surface through endocytosis, further inhibiting Wnt signalling. However, DKKs are transcriptional targets of Wnt/β-catenin signalling, suggesting that the increased expression of DKK proteins following Wnt pathway activation serves as a negative feedback mechanism (31).

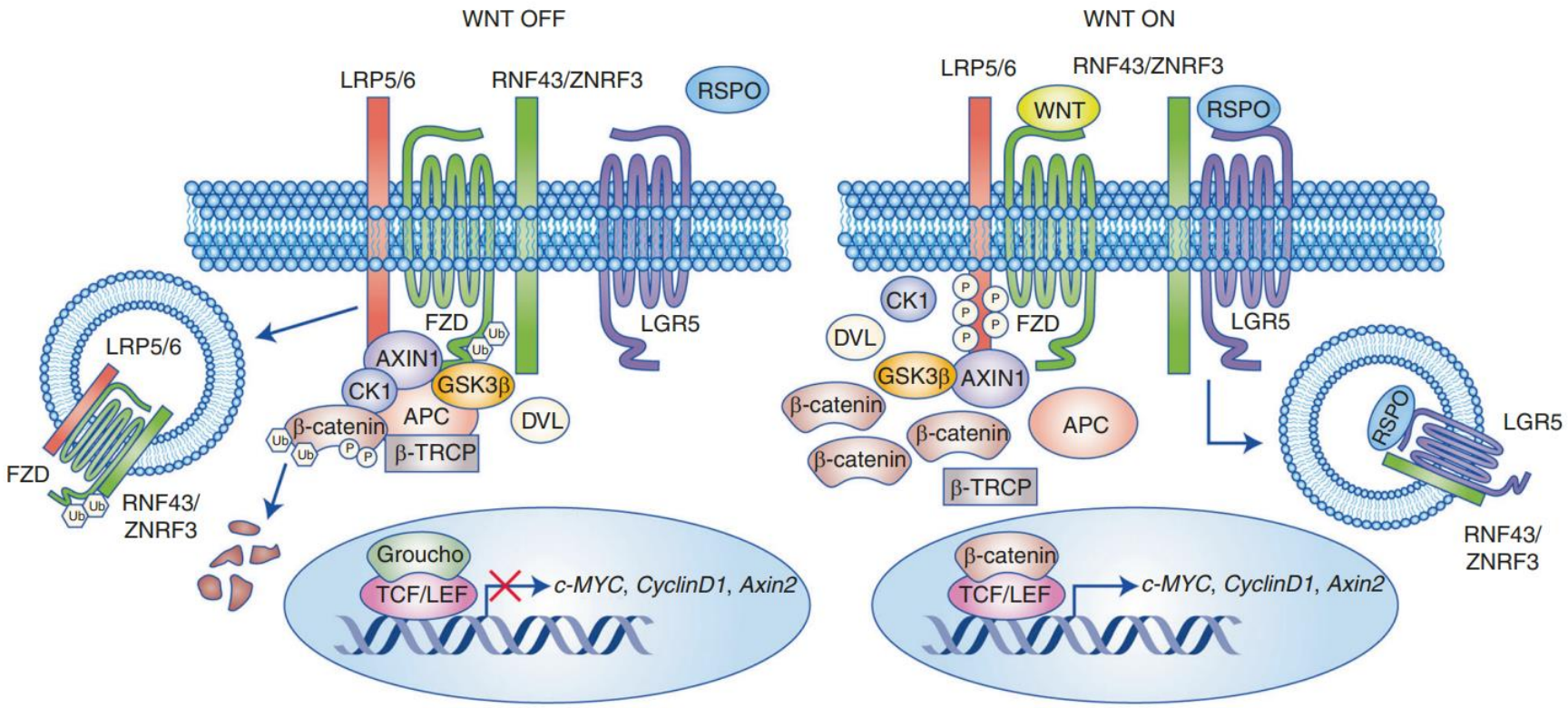


Figure 1.2. Canonical Wnt/β-catenin signalling.

In the absence of a Wnt ligand, the destruction complex phosphorylates and ubiquitinates β-catenin, leading to its subsequent proteasomal degradation (left). The binding of Wnt to the LRP5/6 and FZD co-receptors results in the disassembly of the destruction complex, which leads to the stabilisation and cytoplasmic accumulation of β-catenin. This is followed by its translocation into the nucleus, where it promotes the expression of genes that drive cell proliferation, differentiation, and survival (right). Figure reproduced from (28).

As scaffolding proteins within the destruction complex, APC (and Axin1/2) facilitate the proximity of β-catenin to GSK-3β and CK1α (17). Mutations in APC, which often result in truncated proteins unable to form a functional destruction complex, lead to a decreased affinity for β-catenin (17). This results in an impaired ability of the complex to promote β-catenin's phosphorylation (32). A dysfunctional destruction complex ultimately precludes β-catenin degradation, which then accumulates in the cytoplasm and translocates into the nucleus, ultimately leading to the ligand-independent constitutive activation of genes regulated by Wnt signalling that promote the proliferation, migration, invasion and metastasis of cancerous cells, including *MYC*, the cyclin D1 gene (*CCND1*), vascular endothelial growth factor (*VEGF*) genes, and the peroxisome proliferator-activated receptor delta (*PPAR-δ*) gene (32).

Oncogenic mutations affecting the β-catenin gene (*CTNNB1*) and *AXIN1/2* mutations can also activate Wnt signalling, though to a lesser extent, even in the absence of *APC* mutations (32). Similarly, somatic mutations within the Ras/Raf/MEK/ERK and PI3K/AKT pathways can also affect Wnt signalling:

As previously stated, activated KRAS phosphorylates the p110α subunit of PI3K (33). Thus, by activating PI3K, KRAS directly activates the PI3K/AKT pathway. Once phosphorylated, PI3K converts PIP2 to PIP3. PIP3 serves as a docking site for AKT, its upstream activator, phosphoinositide-dependent kinase-1 (PDK1), both of which have pleckstrin homology (PH) domains that bind PIP3, localising them to the cell membrane (34). PDK1 phosphorylates AKT at the threonine 308 (Thr308) residue. Full activation of AKT also requires phosphorylation at serine 473 (Ser473) by the mTORC2 complex or other kinases like DNA-PK. Activated AKT can negatively regulate GSK-3β kinase activity through post-translational modifications (26). Therefore, through AKT-mediated inhibition of GSK-3β, KRAS can indirectly promote β-catenin stability and activity, highlighting the impact of both KRAS and PI3K/AKT pathway mutations on tumorigenesis of CRC.

RAS/RAF/MEK/ERK/p90 signalling can also inactivate GSK-3β at S9 upon pathway activation by ligand binding of epidermal growth factor (EGF) and platelet-derived growth factor (PDGF) to their respective receptors. Additionally, ERK phosphorylation at GSK-3β T43 induces a conformational change that impacts its function (35). Furthermore, recent studies have revealed new mechanisms explaining the cross-talk between the Wnt/β-catenin and RAS-ERK signalling pathways (36). These findings suggest that the destruction complex can regulate the stability of RAS proteins in a manner akin to β-catenin regulation, i.e., through GSK3β-mediated phosphorylation of RAS proteins at certain threonine residues. This phosphorylation facilitates RAS recognition by the E3 ubiquitin ligase β-TrCP, leading to its degradation. This process keeps

RAS levels in check under resting conditions. However, upon Wnt pathway activation, β-catenin and RAS proteins accumulate because of the disruption of the destruction complex, leading to their accumulation in the cytoplasm. In the case of RAS, the accumulated RAS proteins would then activate the RAF/MEK/ERK and PI3K/AKT signalling pathways, promoting cell proliferation and transformation (36).

Understanding the interactions of the Wnt signalling pathway and its interplay with other pathways like RAF/MEK/ERK, PI3K/AKT, and TGF-β helps identify how specific signalling aberrations contribute to disease progression. Insights into signalling cross-talks can identify key nodes within these networks that, when modulated, can correct pathological signalling. This is particularly relevant for developing targeted therapies in cancer treatment, where the goal is to specifically inhibit oncogenic signals without disrupting normal cellular functions.

1.1.3 The adenoma-carcinoma sequence model of CRC progression

Nearly 96% of CRC cases originate from adenomas (4), which are subclassified based on their histological appearance. Tubular adenomas, characterised by a tube-like morphology, are the most commonly removed polyps, constituting 60-80% of cases. Villous adenomas feature long, finger-like epithelial projections and account for 5-10% of polyps. These adenomas often exhibit more severe atypia and dysplasia compared to tubular adenomas. Lastly, tubulo-villous adenomas, which exhibit tubular and villous adenomas features, constitute 10-25% of polyps (4).

The likelihood of CRC development escalates with increasing dysplasia (4). Tubular adenomas smaller than 1 cm in diameter have a 5% chance of becoming cancerous. In contrast, villous adenomas larger than 2 cm have a 50% chance of malignant transformation, while tubule-villous adenomas present an intermediate risk of 22% (4). This gradual progression provides a critical window for early detection and removal, as colonoscopic polypectomies significantly reduce the risk of developing CRC (17). However, the transition from an early adenoma to an established CRC may take no less than 10 years, while a polyp may take as long as 18 years to develop into an invasive cancer (37).

Initially proposed by Fearon and Vogelstein, the adenoma-carcinoma sequence model outlines how CRC progresses from normal colonic epithelium through benign adenoma to malignant carcinoma in a stepwise manner (38, 39), suggesting that interventions at earlier stages could potentially prevent the progression to cancer. This progression is mediated by the three pathways previously described in the text: chromosomal instability (CIN), microsatellite instability (MSI), and the CpG island methylator phenotype (CIMP) (38, 39). Thus, each of these pathways represents a distinct mechanism by which genetic and epigenetic changes can drive the transformation of normal colonic cells into cancerous cells.

Focusing on the CIN-specific model of CRC progression (Figure 1.3), carcinogenesis begins with a mutation that inactivates the adenomatous polyposis coli (*APC*) TSG located on chromosome 5q21-q22. In patients with FAP, loss of *APC* function results from autosomal germline mutations followed by a second somatic mutation on APC that completely inactivates the gene (15). In contrast, sporadic cases of CRC are typically caused by somatic mutations—often frameshift or nonsense mutations—or allelic deletions at 5q, with a few sporadic cases exhibiting *APC* inactivation through promoter hypermethylation (9, 15). Mutations in *APC* lead to the constitutive activation of Wnt signalling, a critical pathway that regulates stemness in epithelial cells located at the bottom of the intestinal crypts of Lieberkühn (24). Wnt signalling activity is frequently dysregulated in the majority of CRC cases. As a result, mutations in *APC* disrupt the

regulation of normal intestinal stem cell (ISC) growth, ultimately leading to the development of polyps (24).

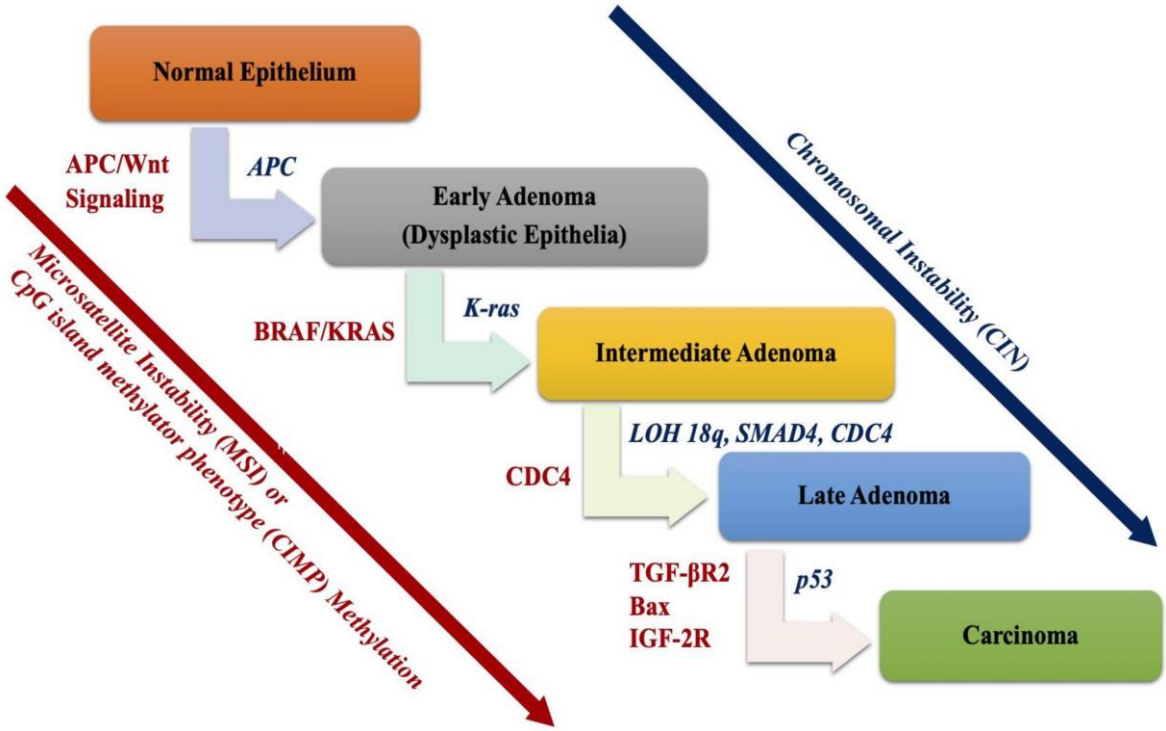


Figure 1.3. Adeno-carcinoma sequence model.

In this model of colorectal cancer progression, tumour development is regulated by chromosomal instability (CIN), microsatellite instability (MSI), and CpG island methylator phenotype (CIMP) hypermethylation. Figure reproduced from (15).

Several studies have highlighted the critical role of *APC* mutations in the initial stages of intestinal tumour formation. In patients with FAP, who are predisposed to developing hundreds to thousands of adenomatous polyps, loss of heterozygosity (LOH) of the second *APC* allele was observed in the majority of premalignant adenomas (40). While these findings did not conclusively prove that inactivation of the second allele occurred early in tumour formation, further research on *APC* heterozygous multiple intestinal neoplasia (Min) mice showed that 100% of the tumours had inactivated the remaining normal *Apc* allele, with this critical inactivation event detectable at the earliest recognisable stages of tumour development (40).

The association between *APC* silencing and the subsequent development of polyps was also demonstrated in murine models of sporadic GI cancers containing a mutant *APC* allele encoding a protein truncated at residue 716 ($Apc^{\Delta 716}$). These mice developed microadenomas throughout the GI tract three weeks after birth (41). It was later demonstrated that heterocyclic aromatic amines like PhIP (2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine), which are generated

when muscle meat (e.g., beef, pork) is cooked at high temperatures, stimulated the growth of intestinal polyps through the formation of PhIP-DNA adducts in *Apc*$^{\Delta 716}$ knockout mice (41). In contrast, omega-3 fatty acids such as docosahexaenoic acid (DHA) obtained from fish significantly reduced the number of polyps when fed to these mice (42).

Indeed, the biallelic inactivation of *APC* marks the initial step in CRC development in 75% of cases (24), conferring ISCs with these mutations a selective growth advantage that leads to their clonal expansion (43), eventually forming premalignant adenomas. However, *APC* mutations alone are insufficient to trigger malignant transformation. The subsequent step in the adeno-carcinoma sequence involves mutations in *KRAS*, which occur early in the adenomatous stage (15). Notably, *KRAS* mutations significantly increased the number of polyps in mice with combined Apc and Kras knockouts (44).

As a small GTPase, KRAS directly interacts with kinases such as RAF and PI3K (15). Consequently, oncogenic *KRAS* mutations lock the protein in its GTP-bound active state, leading to the constitutive activation of the MAPK/ERK (also known as RAS/RAF/MEK/ERK) and PI3K/AKT signalling pathways, which in turn stimulates cell proliferation, enhances cell survival, and facilitates tumour invasion and metastasis. Activating mutations of *KRAS* that are significant for cancer development predominantly affect codons 12 and 13 of exon 2, with less common mutations involving codons 61 and 146 (15, 33). Specifically, missense mutations at codons 12 and 13 result in the substitution of glycine with aspartate, referred to as p.G12D and p.G13D, respectively. Among these, the p.G13D *KRAS* mutation is more predominant in CRC adenocarcinomas (15, 33). Clinically, p.G12D *KRAS* mutations are associated with poorer overall survival in patients with advanced and recurrent CRCs (45, 46), while the predictive value of p.G13D *KRAS* remains inconclusive (47-49).

Continuing with the adeno-carcinoma sequence, the transition from the early to the later adenomatous stage is marked by the deletion of the long arm of chromosome 18 (18q) (50), which is detected in approximately 70% of primary CRCs (51). The high incidence of allelic loss at 18q suggests the presence of candidate TSGs in this region whose inactivation might be crucial for CRC progression, such as the “deleted in colorectal carcinoma” (*DCC*), *SMAD2* and *SMAD4* genes (9). While the evidence supporting *DCC’s* role as a TSG in sporadic CRCs is circumstantial (52), *SMAD2* and *SMAD4* are recognised TSGs that regulate tumour growth factor-β (TGF-β) signalling, playing pivotal roles in inhibiting tumour growth and invasion (53). However, the relatively low mutation frequency of *SMAD2* and *SMAD4* in CRCs, coupled with observations that smaller regions of loss may exclude these genes, suggests that they might not

be the primary targets of inactivation at 18q (9). Instead, other genes within this region could also be critical targets for inactivation (9).

The final genetic alteration in the transition to malignancy involves the inactivation of the *TP53* TSG on 17p (9). Commonly known as the "guardian of the genome", *TP53* encodes a transcription factor that regulates hundreds of genes involved in various biological processes, such as DNA repair, cell cycle arrest, senescence, apoptosis, and metabolism, in response to a wide range of stress signals. For example, in response to DNA damage, activated p53 binds to specific sequences in the promoter region of *CDKN1A* (*P21*), leading to increased transcription of the p21 cyclin-dependent kinase (CDK) inhibitor (54). Once activated, p21 inhibits the activity of cyclin-CDK complexes, including those formed by cyclin D/CDK4 and cyclin E/CDK2. This inhibition blocks the phosphorylation of proteins essential for initiating the S phase, effectively halting cell cycle progression from the G1 to the S phase. Given the pivotal role of p21 in cell cycle regulation, the loss of *TP53* function allows cells with damaged DNA to survive and propagate, thereby contributing to cancer progression (9).

Most *TP53* mutations in CRC are LOF missense mutations that impair normal p53 tumour-suppressing activities (15). These mutations are more frequently observed in colon carcinomas than in premalignant lesions, which highlights the role of *TP53* in the transition from an adenoma to a carcinoma. Beyond losing its tumour-suppressing function, mutant *TP53* can actively drive cancer progression through gain-of-function (GOF) mechanisms, as exemplified by the *p53*-R273H mutation (55). In compound mouse models harbouring both an $Apc^{\Delta 716}$ and a $Trp53^{R270H}$ mutation (the latter corresponding to the human *p53*-R273H mutation), the GOF mutant *p53*-R270H significantly increased the aggressiveness of intestinal tumours by promoting submucosal invasion (24). This demonstrates that the presence of this specific *p53* mutation not only facilitates tumour formation but also leads to a more aggressive tumour phenotype. The *p53*-R273H mutation in tumours is often linked with enhanced resistance to anoikis and overall aggressive cancer behaviour in breast and CRC cells (55).

Apart from the mutations discussed earlier, *BRAF* and *PIK3CA* mutations are also frequently identified in sporadic CRC. Like *KRAS*, *BRAF* is a critical kinase in the MAPK/ERK pathway, acting downstream of *KRAS* (56). The most common *BRAF* mutation involves a single nucleotide substitution that replaces valine with glutamic acid at codon 600 (V600E) (33). Present in 10 to 18% of CRC cases, *BRAF* mutations are predominantly observed in tumours with CIMP and *MLH1*-inactivated MSI, whereas they are less frequent in MSS tumours (33, 56). Activating *BRAF* mutations such as V600E result in the constitutive activation of the kinase independent of *KRAS*, leading to continuous activation of MAPK signalling (17). The co-

occurrence of *BRAF* and *KRAS* mutations in CRC is extremely rare, suggesting that oncogenic mutations in either gene can independently confer a selective proliferative advantage (33).

Conversely, *PIK3CA* mutations typically emerge late in tumorigenesis, as they are often found in very advanced adenomas (57). Missense mutations in *PIK3CA*, which are present in 14-25% of CRC cases irrespective of MSI or MSS status, may coexist with mutations in *KRAS* and *BRAF* (56). *PIK3CA* encodes the p110α catalytic subunit of PI3Kα, which is directly activated by KRAS, with hotspot mutations predominantly affecting the helical and kinase domains of p110α (33). Such mutations enhance the activity of PI3K/AKT/mTOR signalling independently of epidermal growth factor receptor (EGFR) activation, leading to the increased production of PIP3. This key second messenger recruits and activates downstream signalling proteins, including AKT, ultimately triggering events that promote cell survival, growth, proliferation, and metabolism (17).

Furthermore, array-based comparative genomic hybridisation (CGH) studies have identified frequent chromosomal alterations linked to CRC progression and metastasis (Figure 1.4) (58), impacting a diverse set of genes through these structural variants (Figure 1.5) (59). Early developments of primary carcinomas and liver metastases are marked by losses of 17p, 18, and 22q, alongside gains of 8q, 13q, and 20 (59) (Figure 1.4). Conversely, deletions at 4p and 8p and gains at 7p and 17q are associated with later changes in primary carcinomas but appear early in liver metastases. Furthermore, established liver metastases are characterised by losses at

14q and 17q and gains at 1q, 9p, 11, 12p, 19, and X. The genetic mutations that transform the normal colon into a carcinoma align with observed clinicopathological changes (50).

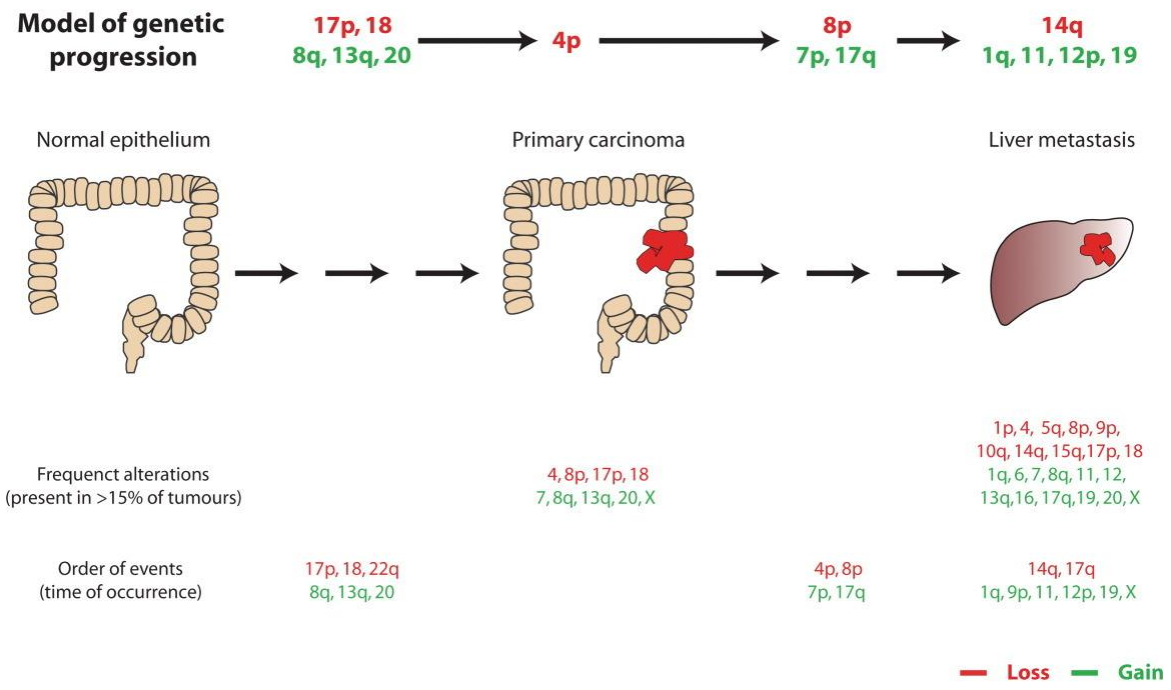


Figure 1.4. Model of genetic progression in colorectal cancer.

Sequence of DNA copy number changes in the progression of colorectal cancer, from a normal colonic epithelium through primary carcinoma development to liver metastasis. Figure reproduced from (59).

1.1. Molecular characteristics of colorectal cancer

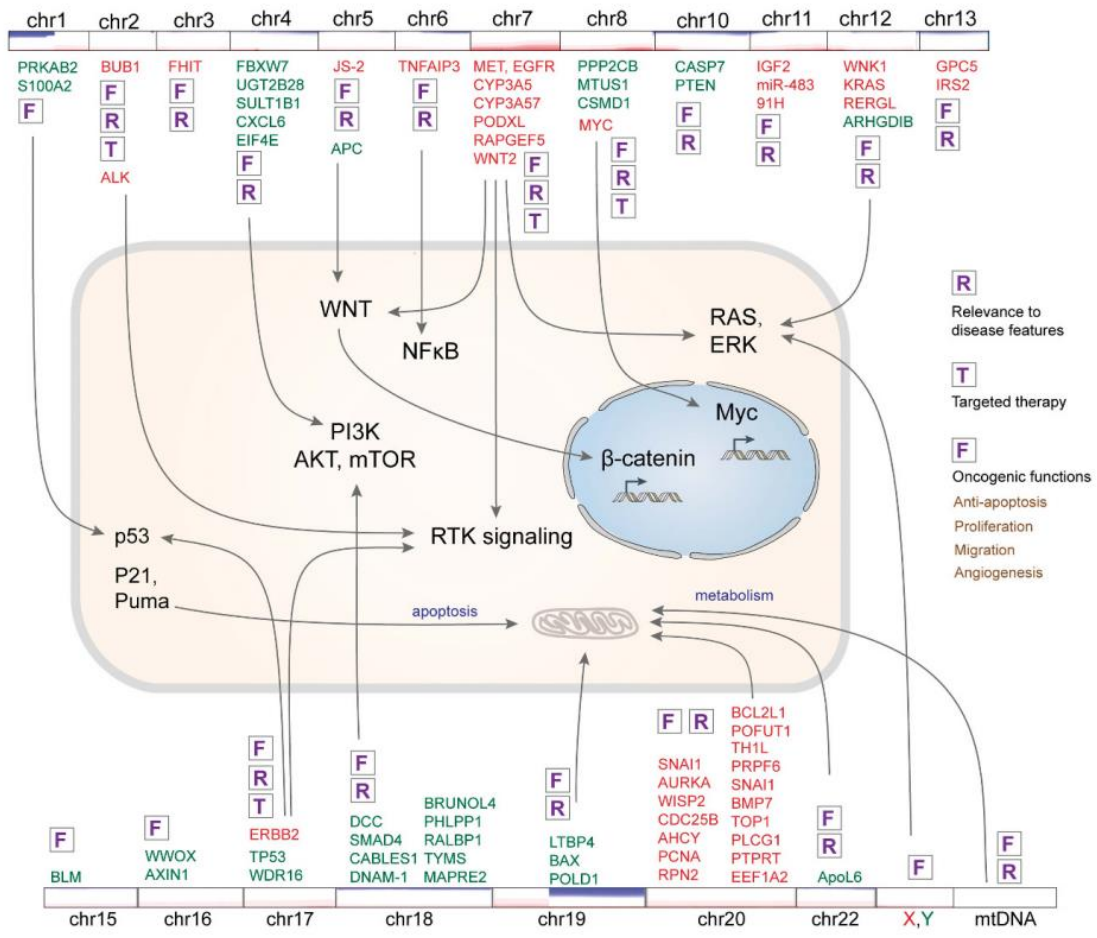


Figure 1.5. Genes impacted by somatic CNAs in CRC and pathways in which they operate.

Genes within CNAs that exhibit gains/amplifications are highlighted in red, while losses/deletions are highlighted in green. Arrows indicate the involvement of affected genes in signalling pathways. The relevance to disease features (R), targeted therapy (T) and oncogenic functions (F) labels are shown for each CNA region. Figure reproduced from (58).

1.2 Management and treatment approaches for CRC

1.2.1 Preventive strategies, screening and treatment

Preventative measures for CRC aim to reduce the risk of developing the disease through lifestyle modifications, regular screening, and in some cases, medical interventions (37). These preventive measures can be categorised into primary, secondary, and tertiary strategies, each targeting different stages in the prevention and management of the disease.

Primary prevention aims to avert the onset of CRC by addressing risk factors and promoting healthy behaviours in an otherwise healthy population (37). Epidemiological studies have identified several modifiable factors. For instance, the risk of CRC can be reduced by limiting alcohol intake, avoiding smoking, adopting a healthy diet rich in fruits and vegetables while reducing the consumption of red and processed meats, managing weight to prevent obesity, and promoting regular exercise. For individuals at high risk, such as those with irritable bowel diseases (IBDs), primary prevention may additionally include hormone replacement therapy and regular consumption of aspirin or other non-steroidal anti-inflammatory drugs (NSAIDs), both of which are associated with reduced CRC risk (37).

Secondary prevention focuses on the early detection and treatment of precancerous lesions or early-stage CRC to delay or halt its progression (37). Although early-stage cancer is usually asymptomatic, the progression of the disease can lead to observable clinical symptoms, e.g., changes in bowel habits, haematochezia—the passage of blood in stools—, rectal bleeding, iron-deficiency anaemia, weight loss, intestinal obstruction, abdominal pain, palpable abdominal mass and intestinal perforation (60, 61). These symptoms can prompt further diagnostic follow-up. For detailed examination, colonoscopy is considered one of the most effective screening tools, as it allows direct visualisation of the entire intestinal lining (mucosa), capturing detailed images that reveal the shape, size, and location of any abnormalities (60). During the procedure, a biopsy may be taken for subsequent pathological analyses to determine the nature of the lesion. In the UK, current colonoscopy screening recommendations suggest that average-risk men and women should begin screening at ages 60-74 (62). More frequent or earlier screening tests may be offered for monitoring high-risk populations, including individuals with IBDs, those with a family history of CRC, or those with known genetic predispositions or other risk factors (37).

Other examinations include: (i) sigmoidoscopies; (ii) digital rectal examinations for patients suspected of rectal cancer; (iii) stool-based tests such as the faecal occult blood test (FOBT) and the faecal immunochemical test (FIT) (60). The sensitivity and specificity of these tests are somewhat controversial, as haematochezia can be caused by non-cancerous conditions like haemorrhoids or may fail to detect cancers that do not bleed; (iv) detection of tumour markers in the blood. While there are no specific tumour markers for CRC, CEA (carcinoembryonic antigen) and CA19-9 (carbohydrate Antigen 19–9) are commonly used to monitor treatment effectiveness and check for CRC recurrence. However, their sensitivity and specificity for detecting CRC are moderate (40-70% and 73-90%, respectively), making them unreliable for initial screening and diagnosis. Additionally, CEA levels can vary greatly among healthy and asymptomatic individuals, which complicates its use as a diagnostic tool (60).

Treatment for colorectal polyps and CRC varies depending on the stage of the disease at detection (Figure 1.6). For Stage 0 and Stage I CRCs, removing the polyp during colonoscopy or sigmoidoscopy may be the only treatment required if the pathologist confirms clear margins that indicate no apparent risk of residual cancer (63). If cancer cells are found at the margins of the polyp, additional surgery is generally required. A partial colectomy, i.e., the removal of the section of the colon containing the cancer, may be necessary if the cancer is too large for local excision alone. For non-polyposis Stage I cancers, the standard treatment also involves a partial colectomy (63).

The primary treatment for Stage II and Stage III CRCs is typically a partial colectomy (63). However, if the tumour has significantly invaded neighbouring organs (T4b) or is initially inoperable, neoadjuvant therapy may be employed. This therapy usually includes chemotherapy and, in some cases, radiation therapy. Radiation is rarely used for Stage II patients unless it involves rectal cancer, but it is more likely recommended for Stage III patients to reduce the clinical stage of the tumour (60, 63). After neoadjuvant therapy has reduced the tumour size, surgery is performed to remove it. Following surgery, if high-risk features such as high-grade tumours or metastasis to nearby blood or lymph vessels are identified in the surgical specimen, adjuvant chemotherapy may be recommended for Stage II CRC patients. It is standard for all Stage III patients due to the increased risk of recurrence (63).

Despite the implementation of screening programs and subsequent treatment strategies, CRC might still recur. Tertiary prevention focuses on minimising the likelihood of cancer recurrence and addressing treatment-related complications in CRC patients (64). The predictive value of chemoprevention as a tertiary strategy is an active area of research, with observational studies

suggesting that regular, low-dose aspirin consumption after CRC diagnosis may improve survival rates by reducing the risk of adenoma recurrence (64).

Unfortunately, for very advanced cases like Stage IV, treatment typically focuses on managing symptoms and prolonging life rather than curing the cancer due to its extensive spread. Surgery might be considered if metastases are limited and operable, particularly in the liver or lungs. For inoperable liver metastases, techniques such as ablation or embolisation may be employed. Nevertheless, both neoadjuvant and adjuvant chemotherapy are central to treatment (63).

1.2. Management and treatment approaches for CRC

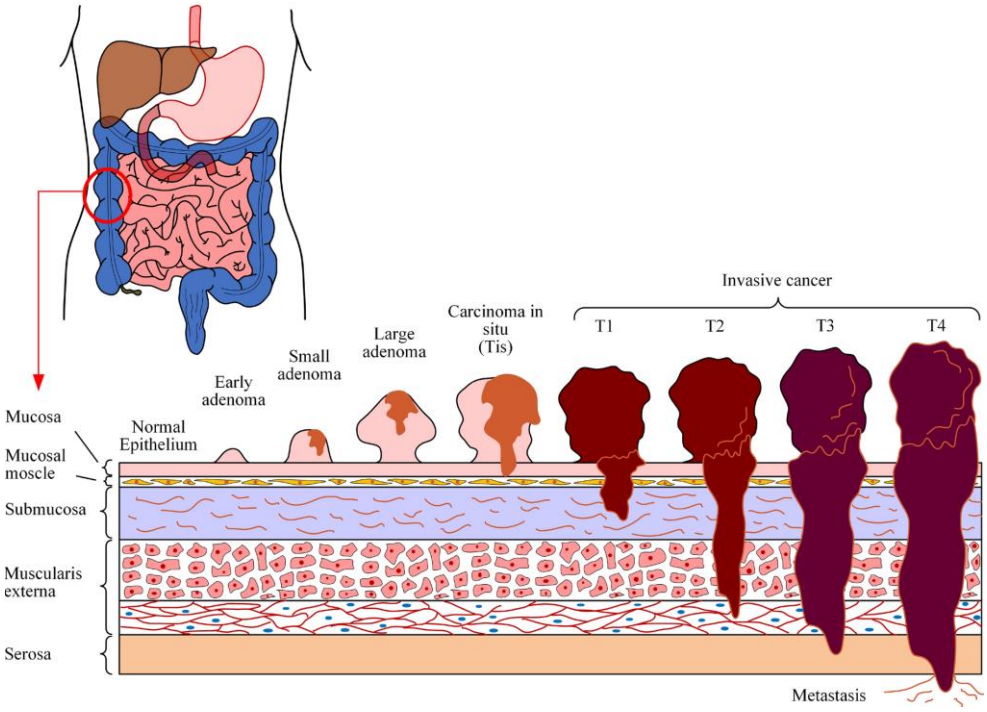


Figure 1.6. Colorectal cancer stages

***Stage 0**, also known as carcinoma in situ or intramucosal carcinoma (**Tis**), represents the earliest form of cancer where abnormal cells are present in the mucosa, the innermost layer of the colon or rectum. Although these cells are considered precancerous, they have not yet invaded deeper layers as they are generally contained within polyps (61).*

***Stage I**: The tumour has spread through the mucosa into the submucosa (**T1**) and possibly into the muscle layer or muscularis propria (**T2**). There is no spread to lymph nodes or distant sites.*

***Stage II**: Cancer has spread through the muscularis propria into the subserosa or into the visceral peritoneum without perforation (**Stage IIA, T3**), or through the visceral peritoneum **(Stage IIB, T4a**) or into nearby organs (**Stage IIC, T4b**) (65).*

***Stage III**: Cancer might have grown through all the layers of the colon into nearby tissues (**T3 or T4**) and results in more significant lymph node involvement (65).*

***Stage IV**: The tumour has spread or metastasised beyond the local region to distant organs via the lymphatic system or bloodstream, with the liver and the lungs being the most common target organs of CRC hematogenous metastasis (61). Figure adapted from (61).*

1.2. Management and treatment approaches for CRC

1.2.2 Treatment options and drug resistance mechanisms

CRC is among the most challenging diseases to treat, with treatment outcomes and survival rates closely linked to the point of intervention along the adeno-carcinoma pathway and disease stage at diagnosis (37, 66). Patients diagnosed at an early stage have a 5-year survival rate of 90%, which indicates that early detection is crucial for a favourable outcome (66). However, this rate decreases to 70% for patients with locally advanced tumours and falls to 15% for those with metastatic CRC (mCRC). The development of CRC is a slow process, often taking up to 18 years for pre-cancerous polyps, which are usually asymptomatic, to evolve into malignant tumours (37). Moreover, current screening methods can only detect about 40% of CRC cases in the early stages. Consequently, the majority of patients are diagnosed at an advanced stage, when treatment becomes more challenging, and the likelihood of long-term survival diminishes.

Figure 1.7 displays FDA-approved and candidate antineoplastic drugs for CRC treatment, highlighting their cellular targets. The most commonly employed first-line treatments include fluorouracil, often referred to as 5-FU; capecitabine (CAP), also known by the brand name Xeloda, which is a 5-FU prodrug; oxaliplatin (OX) and irinotecan (IR) (37). Treatment regimens typically consist of a combination of two or three of these agents, e.g., FOLFOX (5-FU + OX), FOLFIRI (5-FU + IRI + folinic acid), XELOX or CAPOX (CAP + OX), and CAPIRI (CAP + IRI) (37, 67). These cytotoxic chemotherapy drugs induce cell death by interfering with DNA and RNA synthesis. For example, 5-FU is a pyrimidine analogue that inhibits thymidylate synthase (TS), an enzyme crucial for synthesising thymidine monophosphate (dTMP), a nucleotide required for DNA synthesis. Additionally, 5-FU can be incorporated into RNA in place of uracil, and into DNA in place of thymidine, causing faulty RNA processing and DNA synthesis, leading to cell death (68). By inhibiting TS and disrupting RNA and DNA, 5-FU disrupts cell division and leads to apoptosis in rapidly dividing cancer cells.

1.2. Management and treatment approaches for CRC

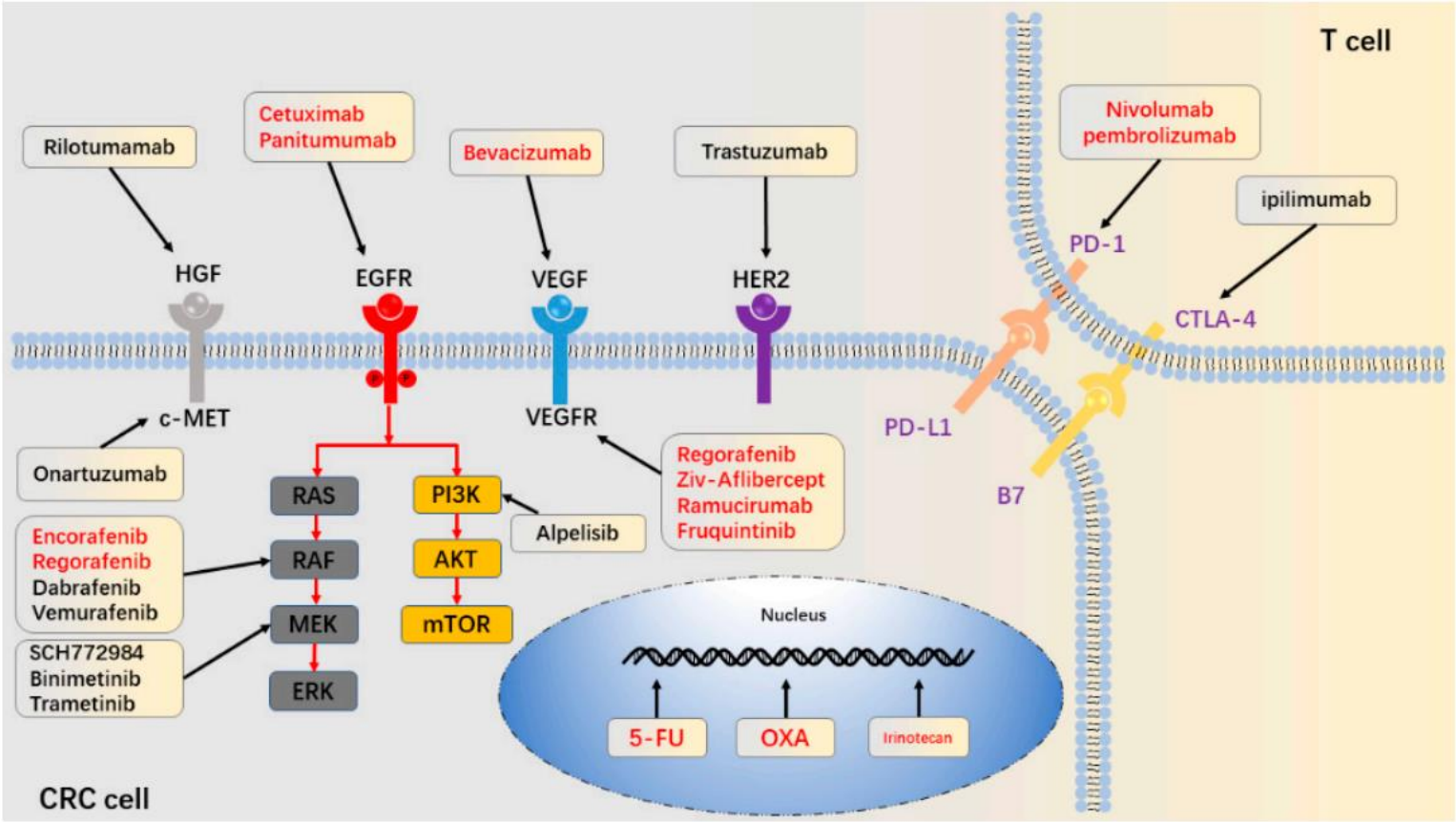


Figure 1.7. Anti-cancer drugs for the treatment of CRC and their cellular targets.

Anti-cancer drugs currently employed in the clinic or undergoing clinical trials for the treatment of CRC. Figure reproduced from (66).

On the other hand, OX is a platinum-based drugs like cisplatin that forms platinum-DNA adducts, causing DNA cross-linking and inhibiting DNA replication and transcription (69). The DNA damage induced by OX triggers cell death pathways. On the other hand, SN-38, the active metabolite of IRI, exerts a cytotoxic effect by inhibiting topoisomerase I (Top1), an enzyme that relieves torsional strain in DNA by inducing single-strand breaks, which allows DNA replication and transcription to proceed (70). SN-38 stabilises the complex between Top1 and DNA, forming the Top1-DNA-SN-38 complex and preventing the re-ligation of these single-strand breaks. The accumulation of single-strand breaks leads to replication fork collisions that result in double-strand breaks, ultimately causing cell death (70).

While first-line treatments are standardised for the general patient population, additional targeted therapies may be offered to CRC patients with more advanced stages (Stage III and Stage IV) or metastases, and those with a strong family history of CRC may also receive other forms of treatment Berner, 2022 #535}(63). These targeted treatments employ monoclonal antibodies (mAbs) and small molecule inhibitors as antagonists to exploit specific vulnerabilities of cancer cells (Figure 1.7) (71). These include mutations in surface receptors and other signalling molecules identified through immunohistochemistry or genetic testing. In the UK, all patients with CRC are eligible for testing for mismatch repair (MMR) deficiency through immunohistochemistry (IHC) to check for the presence or absence of MMR proteins (MSH2, MSH6 or PMS2) (72), which may be predictive of underlying genetic aberration (72, 73). In the case of abnormal IHC results, patients are also eligible for constitutional (germline) testing with whole genome sequencing (WGS) to determine the microsatellite instability (MSI) status of patients to help inform treatment options (74).

Targeted therapies aim to increase treatment efficacy and enhance survival rates in patients by inhibiting the function of critical proteins and disrupting downstream signalling pathways essential for tumour progression (71). For example, blocking EGFR signalling with the mAbs cetuximab or panitumumab inhibits EGFR signalling and can increase survival for 10–20% of mCRC patients (75). In addition, immunotherapies have improved cancer treatment by targeting immune checkpoint proteins such as CTLA-4 and PD-1 on T-cell receptors (TCRs), thereby preventing their inhibition and enhancing the immune system's ability to attack cancer cells (Figure 1.7) (76). For instance, MSI CRC patients whose tumours are deficient in DNA mismatch repair (dMMR) have mutations that produce a large number of neoantigens that can be recognised by the immune system (71). Meta-analysis of early studies revealed that while patients with dMMR tumours have a better prognosis, they are generally less responsive to 5-FU-based treatments (71, 77). This is likely because these tumours may repair the DNA damage caused by 5-FU less effectively or tolerate it without undergoing apoptosis due to their high

mutation burden and altered cell death pathways. Given that MSI-dMMR CRCs exhibit higher levels of lymphocyte infiltration, their enhanced response to cytotoxic T-cell activity is not surprising (71, 78). This response can be further amplified using immune checkpoint inhibitors (ICI), such as ipilimumab, leveraging the immune system's ability to target and eliminate cancer cells.

The emergence of primary (never-responders) and secondary (acquired) drug resistance significantly challenges CRC treatment (66). In addition, resistance to cancer therapies can arise through mechanisms that are intrinsic or extrinsic to tumours (76). Intrinsic resistance refers to the pre-existing characteristics of tumour cells that make them naturally resistant to a particular therapy even before exposure to the treatment, e.g., genetic mutations in cancer cells, molecular pathways that are inherently active or the presence of cancer stem cells which are less responsive to conventional treatments owing to their tumorigenic potential (20, 76). Extrinsic resistance involves changes that occur in tumour cells or their microenvironment after exposure to a therapy, leading to the development of resistance over time.

Indeed, tumour cells exhibit various mechanisms to protect themselves against anti-cancer drugs (Table 1). For example, the metabolic enzyme thymidine phosphorylase (TP) is required to transform CAP into an active 5-FU form (79). Preclinical studies have shown that methylation of extracellular growth factor-1 (*ECGF-1*), the gene that encodes TP, can cause resistance to CAP, which was reversed using inhibitors of DNA methyltransferases (DNMTs) (79). In addition, aberrations in downstream signalling pathways can also lead to drug resistance (66). Assessing *KRAS* mutation status is a requirement for patients with mCRC, as constitutive activation of the kinase can impact both the MAPK/ERK and PI3K/AKT signalling pathways, rendering anti-EGFR therapies ineffective (33). It is also common to observe tumours resistant to bevacizumab (Figure 1.7), which aims to inhibit angiogenesis by blocking the vascular endothelial growth factor receptor (VEGFR) (80). This resistance occurs because angiogenic signalling can be activated by alternative ligands (e.g., Ang-1, EGF, FGF) and their respective receptors, thereby bypassing the blockage of VEGF signalling. Additional mechanisms of resistance include the upregulation of ATP-binding cassette (ABC) and solute carrier (SLC) membrane transporters, which regulate the transport of drugs into and out of cells, alterations in drug targets (previously exemplified in the text with the study on MSI-dMMR CRCs and their unresponsiveness to 5-FU), aberrations in cell death pathways, and changes within the tumour microenvironment (TME) (66).

Although advancements in chemotherapy, targeted therapy, and immunotherapy have improved survival rates, not all patients respond effectively to these treatments. The presence

of multiple mechanisms of drug resistance makes it challenging to predict patient responses to therapy accurately. Consequently, there is a pressing need to develop new treatment strategies that can predict the evolution of CRCs.

1.2. Management and treatment approaches for CRC

Table 1. Anti-cancer drugs employed for treating CRC and key molecular mechanisms underlying drug resistance.

Drug name	Type	Target/Mechanism of Action	Mechanism of drug resistance	References
5-Fluorouracil (5-FU)	Cytotoxic (nucleoside analogue)	Thymidylate synthase inhibitor. Inhibits DNA synthesis.	Increased expression of thymidylate synthase; alterations in drug uptake or metabolism.	(66)
Capecitabine/Xeloda (5-FU prodrug)			Conversion inefficiency due to thymine phosphorylase deficiency; increased thymidylate synthase activity.	(66)
Oxaliplatin	Cytotoxic (platinum-based drug)	Cross-links with DNA. Prevents DNA replication and transcription.	Overexpression of efflux pumps; increased autophagy.	(81) (82)
Irinotecan (SN-38)	Cytotoxic (topoisomerase I inhibitor)	Inhibits Top1-DNA complex after DNA cleavage. Prevents re-ligation of single strand breaks and induces replication arrest.	Overexpression of efflux pumps; alterations in topoisomerase I.	(70)
Cetuximab	Targeted (monoclonal antibody)	EGFR inhibitor. Blocks cell proliferation.	Mutations in *EGFR*; activation of alternative growth factor receptors; *RAS* mutations.	(66)
Bevacizumab	Targeted (monoclonal antibody)	VEGF inhibitor. Inhibits angiogenesis.	Upregulation of alternative angiogenic pathways (e.g., FGF); hypoxic tumour microenvironment adaptations.	(80)
Trastuzumab	Targeted (monoclonal antibody)	HER2 receptor inhibitor. Inhibiting HER2-driven cell proliferation.	Overexpression of *HER2*; activation of alternative growth factor pathways; truncated HER2 receptor or epitope masking.	(83)

Pembrolizumab	Targeted (monoclonal antibody)	PD-1 receptor inhibitor. Enhances the immune system's ability to detect and destroy cancer cells.	Upregulation of alternative immune checkpoints; immunosuppressive TME; cancer cell antigen loss; reduced antigen presentation.	(84)
Ipilimumab	Targeted (monoclonal antibody)	CTLA-4 receptor inhibitor. Enhances the immune system's ability to detect and destroy cancer cells.		(85)
Regorafenib	Targeted (tyrosine kinase inhibitor)	Multi-kinase inhibitor. Inhibits cell proliferation and angiogenesis.	Activation of compensatory signalling pathways; mutations in target kinases.	(86)
Vemurafenib	Targeted (serine/threonine kinase inhibitor)	Targets BRAF V600E mutations. Inhibits MAPK pathway and thus cell proliferation.	Upregulation of MAPK signalling through alternative pathways; mutations/copy number gains in BRAF.	(87)

1.3 Evolutionary biology shapes cancer research

Evolutionary adaptation by organisms that grow in ecological landscapes differs from the adaptation by somatic cells (88); however, insights from evolutionary biology can be extrapolated to cancer biology to understand the evolutionary dynamics in human cancers. This is possible because tumours present all the necessary conditions for evolutionary adaptation to occur, i.e., variation in fitness, selection, and inheritance (11).

1.3.1 Models of clonal evolution in cancer

Cancer evolutionary biology refers to the study of how tumours develop and change over time, focusing on understanding the genetic changes and cellular processes that contribute to the diversity of cells within a tumour (89). There are several models of clonal evolution in cancer:

The linear model of cancer evolution, originally proposed by Peter Nowell, was the first theory to describe cancer as an evolutionary process (43). According to this model, cancer evolves through a process where mutations are acquired linearly in a stepwise manner, with each new mutation providing a selective advantage that allows a single clone—a population of cells within a tumour that originated from a single ancestral cell—to dominate the tumour by outcompeting other clones (43, 90). In this model, tumour progression is characterised by sequential "selective sweeps", where each advantageous mutation leads to the rise of a new dominant clone that replaces the previous one (90). Consequently, tumours tend to be relatively homogeneous at any given time, with only a few remnants of earlier clones persisting (Figure 1.8A).

Although the linear model of cancer evolution aligns well with the adenoma-carcinoma sequence model of colorectal cancer progression, and further explains why tumours become increasingly aggressive over time (38, 39), tumours are not exclusively composed of genetically identical clones. Nowell further highlighted this complexity by observing that the mutation rate in neoplastic populations is higher than in healthy cells, concluding that genomic instability inevitably increases with cancer progression (43, 91).

Over time, variations in fitness within tumour clones lead to clonal diversification, resulting in the expansion and coexistence of multiple subclones simultaneously (90, 92). Each subclone descends from an earlier clone but acquires additional mutations unique to this subset of cells. This model of cancer evolution, known as branching

evolution (Figure 1.8B), illustrates how tumours evolve through a non-linear accumulation of genetic changes (90). Such diversity within the subclones that constitute a tumour is commonly known as intratumour heterogeneity (ITH), and is a critical factor in the progression and adaptability of many solid cancers, such as melanomas (93, 94), breast cancers (95, 96), lung cancers (93, 97), and gastrointestinal cancers (98, 99).

Neutral evolution is an extreme form of branching evolution, positing that in the absence of strong selective pressures, neutral mutations accumulate randomly in the cancer cell population (90) (Figure 1.8C). This leads to genetic drift, where allele frequencies change by chance rather than due to selective advantages or disadvantages, contributing to ITH. The effects of genetic drift are more pronounced in small populations, where random fluctuations can significantly shift allele frequencies, whereas in larger populations these impacts are diluted.

The Big Bang model was initially developed to describe the early and rapid expansion of CRCs (100), and has since been corroborated in other tumour types (101-103). In this model, CRCs grow predominantly through a single large expansion after initial transformation, generating numerous intermixed subclones early in their development (100). Most genetic alterations arise early and persist throughout the tumour's growth due to rapid expansion and spatial constraints that limit selective sweeps. Consequently, the timing of a mutation is the primary determinant of its frequency within the tumour rather than selection, with all major clones persisting during growth. Early mutations become widespread as the tumour expands, while late mutations are less frequent because they arise after significant tumour expansion, giving them less time to propagate. As a result, late mutations are often confined to smaller subpopulations and localised within the tumour mass, contributing less to the overall genetic diversity. Thus, high ITH is established early and remains uniformly high across the tumour. This aligns with effectively neutral evolution, where many mutations do not confer significant selective advantages (103). While this does not preclude selection, advantageous subclones may be rare or occur too late to expand to detectable frequencies. Nevertheless, the Big Bang model contrasts with the neutral evolution model by focusing specifically on the early establishment and persistence of genetic diversity and pervasive ITH in CRCs, providing a detailed framework within the broader concept of neutral evolution.

1.3. Evolutionary biology shapes cancer research

Contrary to the conventional gradualist view of cancer evolution described above, which posits that genetic changes accumulate slowly and steadily over time, the punctuated model suggests that tumour evolution can occur in bursts of rapid genetic changes due to cataclysmic genomic rearrangements and mutational bursts, followed by periods of relative genetic stability (103) (Figure 1.8D). During the earliest stages of tumour progression, this results in very high ITH. Over time, however, one or a few dominant clones expand to form the tumour mass. These bursts of genomic alterations can create new dominant clones and subclones, fitting into the broader framework of branching and parallel evolution (90).

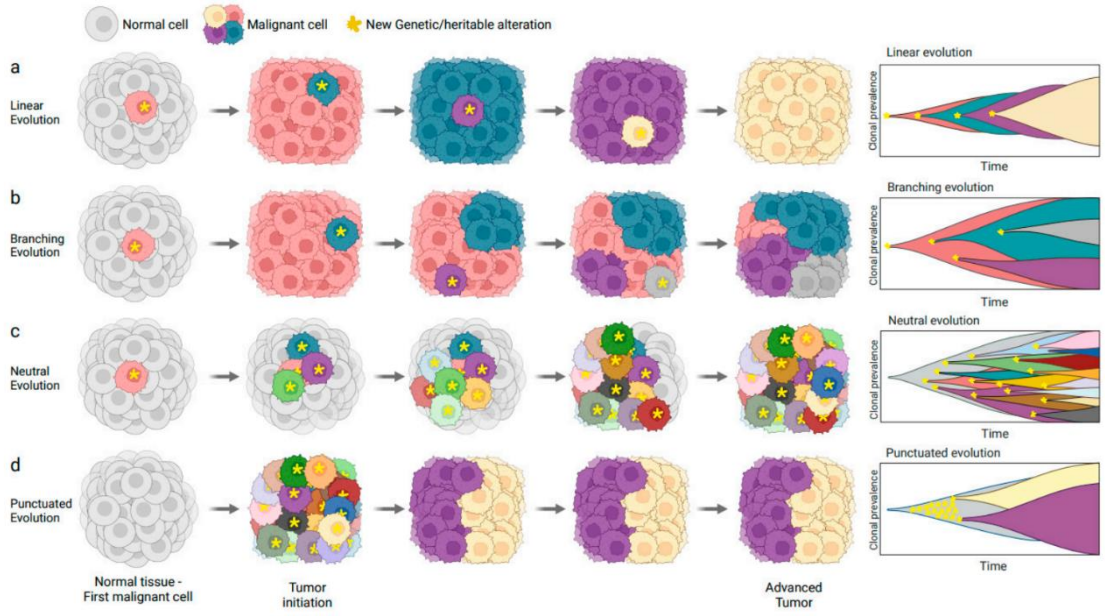


Figure 1.8. Models of clonal evolution in cancer.

The figure depicts the key models of tumour evolution, showing how cancer cells develop and proliferate over time (left), along with fish plots (right) that illustrate the prevalence of different clones over time for each evolutionary pattern: (a) Linear, (b) Branching, (c) Neutral, and (d) Punctuated Evolution. Figure reproduced from (104).

1.3. Evolutionary biology shapes cancer research

Besides the punctuated model of cancer evolution, the cancer stem cell (CSC) model offers a fundamentally different perspective on tumour growth and progression. The CSC model suggests that within a tumour, there exists a hierarchical organisation where a small population of stem cell-like cancer cells are responsible for initiating and sustaining tumour growth (105). These CSCs possess the ability to self-renew and differentiate into various cell types that constitute the bulk of the tumour. Unlike other models where genetic diversity and tumour progression are driven by mutations in all tumour cells, the CSC model emphasises the role of cellular hierarchy and the unique properties of CSCs in tumour dynamics. Compelling data support the CSC model in various human cancers, including malignant germ cell cancers, leukaemias, breast cancers, brain cancers, and colon cancers (105). In each of these cancers, only a small subpopulation of cells can transfer disease upon transplantation into immunocompromised mice, and specific markers distinguish these tumorigenic cells from the bulk of non-tumorigenic cells. This indicates that the tumorigenic cells are intrinsically different from the nontumorigenic cells, often without clear morphological distinctions (105).

These models of cancer evolution, which represent only a subset of the many models proposed in the field of cancer research, provide a comprehensive understanding of the genetic diversity within tumours and the dynamics of tumour progression. Traditionally, most studies have focused on a single model to explain cancer evolution. However, new evidence indicates that tumour evolution is more complex than previously thought, with different models potentially acting at different stages of cancer development (106). It is also possible that multiple models may coexist, especially for different types of mutations. Niida *et al*., proposed a mixed model of CRC evolution by integrating genomic analysis with mathematical modelling. This model highlights a shift from Darwinian selection in early-stage tumours to neutral evolution in late-stage tumours (106).

Acknowledging the complexity of tumour evolution and identifying the evolutionary processes that lead to drug resistance allows for the development of strategies to delay, prevent, or overcome it in the future.

1.4 Preclinical models to study intratumour heterogeneity

1.4.1 2D-3D cultures and animal models

Two-dimensional (2D) cultures of immortalised cell lines are essential in preclinical research as they offer a consistent and reproducible platform that is difficult to achieve with more complex preclinical models. A notable example is the NCI-60 panel, a collection of 60 different human cancer cell lines from various cancer types that facilitates high-throughput screening of thousands of compounds to identify those with potential anti-cancer activity (107). Despite their utility, ease of use, and cost-effectiveness, cancer cell line cultures lack the complex architecture and microenvironment of tissues from major organs, which limits their predictive accuracy for *in vivo* conditions. In addition, the shift towards personalised and precision medicine requires preclinical models that mirror the complexity and heterogeneity observed in human tumours more closely.

This gap has been partially bridged by the use of genetically engineered mouse models (GEMMs) and patient-derived xenografts (PDXs). These *in vivo* models enable the study of entire biological systems within a living organism, and are therefore suitable to study cancer biology, as they reflect the cellular and genetic diversity observed in human cancers (108). In particular, GEMMs and GEM-derived allografts are useful for studying the role of specific genes in cancer initiation and progression. However, they fail to encompass the diverse driver mutations and extensive genomic alterations observed in human cancers. On the other hand, chemical carcinogen-induced mouse models are traditional models for studying cancer aetiology and biology but produce unpredictable cancer landscapes that are difficult to replicate consistently, especially considering variations in dosing protocols and animal strains (108).

Indeed, variations between mouse species and differences between humans and mice often result in discrepancies in protein functions and oncogenic mechanisms (108). This can reduce the predictive value of animal models for studying drug efficacy and toxicity, ultimately limiting the applicability of results to human conditions. Additionally, the use of mouse models raises ethical, economical, and logistical concerns, alongside being more expensive and technically challenging than 2D cultures (109).

To better simulate healthy and diseased human tissues, many of the limitations associated with the preclinical models above have been addressed using 3D cultures, such as tissue explants, spheroids and more recently, organoids (108). Tissue explants are fragments of a tissue that are removed and cultured *in vitro* for research purposes (110). These *ex vivo* models retain the native architecture and cellular composition of the organ of origin for

a short period. Spheroids are 3D aggregates of cells created from various cellular sources (e.g., cell lines, multicellular mixtures, primary cells, tumour cells and tissues) that form spontaneously when cultured under non-adherent conditions (111). Spheroids are typically formed using techniques such as the hanging drop method, low-adhesion plates, or spinning bioreactors, which prevent floating cells from attaching to a surface, encouraging them to aggregate and form a spherical shape. Although tumour spheroids lack the organisation and cellular complexity found in other models, which makes them the simplest form of 3D cell culture models, they still exhibit cell-cell and cell-extracellular matrix interactions and are very similar to non-vascularised or poorly vascularised tumours. Additionally, the multilayered structure of spheroids, which consists of an outer layer of proliferating cells, a middle layer of quiescent cells, and an inner layer of hypoxic and necrotic cells, confers chemo- and radio-resistance similar to that seen in human cancers. For these reasons, spheroids are used to study drug efficacy, drug penetration into tumours, and vascularisation (111).

Lastly, organoids are generated from somatic cells, adult and pluripotent stem cells (112). Unlike spheroids, organoids grown into animal-derived extracellular matrices (e.g. Matrigel), form through self-organisation of cells into spatial patterns that resemble the morphology and functionality of real tissues (113). Organoids have been generated from several organs including retina, intestine, thyroid, liver, pituitary, inner ear, kidney, and brain. In particularly, patient-derived organoids (PDOs), which are typically established from primary tumours, mimic the biological characteristics of the parental specimens (109). PDOs represent a balance between physiological relevance and experimental feasibility, making them an ideal platform for studying cancer and assessing drug responses. This is especially important in co-clinical trials, where understanding a patient's susceptibility to candidate cancer treatments can not only inform and guide decision-making, but also help predict drug responses (109). Advances in microfluidics technology have made possible the establishment of multiple PDOs into microfluidics culture devices called organ-on-a-chip (OOAC). OOAC devices mimic the environment of a physiological organ and allow high-throughput drug screening (114). Similar to other 3D models, organoids are not without limitations. With a few exceptions, the vast majority of organoid models do not recapitulate the complete cellular diversity and interactions present in the tumour microenvironment (111).

1.4. Preclinical models to study intratumour heterogeneity

1.4.2 Predictive value of PDOs in drug response and resistance

Vlachogiannis and colleagues established a living biobank of PDOs from patients with metastatic, heavily pre-treated colorectal (mCRC), gastroesophageal (mGOC), and cholangiocarcinoma cancers, who were recruited in phase I or phase II clinical trials (109). These gastrointestinal PDOs derived from $LGR5^{+}$ stem cells, closely resembled the morphology, maintained the protein expression patterns, and mimicked the molecular landscape of their biopsies of origin across multiple passages. Next, to evaluate the feasibility of using PDOs for predicting drug responses based on known molecular targets, the team conducted 3D drug screening assays using 55 drugs employed in clinical practice or undergoing phase I to III clinical trials, over a two-week period. They discovered that within the PDO cohort, only one mCRC PDO, named F-016, which harboured an *AKT1* amplification and E17K mutation (E, glutamic acid; K, lysine), responded strongly to two AKT inhibitors in the drug library (MK-2206 and GSK690693). Similarly, the only mGOC PDO carrying an *ERBB2* amplification exhibited the strongest response to lapatinib, a dual ERBB2/EGFR inhibitor, whereas the same drug showed no effect on the viability of another PDO that exclusively carried an *EGFR* amplification (109).

After drug screening, Vlachogiannis *et al.* also investigated the predictive value of PDOs by comparing clinical responses observed in patients to xenografts created using PDOs from a mCRC patient before (BL) and after (PD) treatment with regorafenib, a multi-tyrosine kinase inhibitor used in the clinic for the treatment of mCRC, advanced GI cancers and hepatocellular carcinomas that blocks oncogenic and angiogenic signalling pathways (109, 115). Their findings showed that BL PDO-xenografts showed a 60% reduction in microvasculature after regorafenib treatment, mirroring clinical outcomes. However, this effect was not observed in PD PDO-xenografts, indicating resistance.

The research conducted by Vlachogiannis *et al.* demonstrates the value of PDOs in predicting patients' responses to chemotherapy options. This approach can identify the underlying mechanisms for treatment sensitivity or resistance, making PDOs powerful preclinical tools for modelling cancer evolution

.

1.5 Leveraging single-cell sequencing to explore cancer evolution

1.5.1 Introduction to single-cell sequencing

As fundamental biological units, cells within a multicellular organism exhibit remarkable diversity in form and function throughout development and disease (116). Single-cell sequencing (sc-seq) employs advanced next-generation sequencing technologies to analyse the genetic material from individual cells. By studying individual cell genomes and transcriptomes, sc-seq not only reveals cellular heterogeneity and temporal expression patterns, but also uncovers unique genomic signatures that are crucial for understanding both normal development and disease states.

Single-cell sequencing involves several key steps to isolate and analyse the genetic material from individual cells. The process begins with the enzymatic or mechanical dissociation of solid tissues into a single-cell suspension (117). This is followed by the isolation or sorting of cells using methods such as fluorescence-activated cell sorting (FACS), microfluidic or droplet-based approaches, and laser capture microdissection. Once the cells are isolated, they are lysed to release their nucleic acids, which are then extracted and purified for further processing. Given the low amounts of nucleic acids present in single cells, techniques like whole-transcriptome amplification (WTA) after reverse transcription of RNA into cDNA, and whole-genome amplification (WGA) are employed to ensure sufficient material for analysis.

The next stage involves library preparation, where DNA or cDNA is sheared into smaller fragments, either enzymatically or mechanically. Indices and adaptor sequences are then ligated to the ends of these fragments. This step is crucial as it allows for the sequencing of multiple cells and facilitates the sequencing process. The fragments are selected for a desired size range and purified to remove excess adaptors and other impurities. The quality and quantity of the single-cell libraries are then assessed using methods such as qPCR, Qubit, Bioanalyzer, or TapeStation before the libraries are loaded onto a sequencing platform, such as Illumina, PacBio, or Oxford Nanopore, at a desired concentration.

Technological advances in sc-seq are essential in innovative projects such as the Human Cell Atlas (HCA) (118). The HCA is an international collaboration aimed at sequencing all cell types in the human body to create a comprehensive reference map of healthy cells, which will be invaluable references for studies in health and disease.

1.5.2 Applications of single-cell sequencing in cancer research

Evolving knowledge of cancer biology has refined treatment strategies and interventions. Nevertheless, ITH represents a significant challenge and may be a major limitation for successfully treating cancers. Current knowledge of genomic abnormalities in cancer is mainly derived from array-based strategies and conventional high-throughput sequencing (HTS) analyses conducted on bulk tumour specimens (119). However, because bulk tumour samples typically contain a mixture of healthy, malignant, and other cells within the TME, bulk sequencing methods are not sensitive enough to detect molecular changes in minority subclones. Furthermore, RNA-sequencing from bulk samples generates average gene expression profiles, which may not reflect the transcriptional behaviour of all cells that comprised the bulk sample (120). Consequently, a significant limitation of bulk sequencing is the loss of information regarding individual cell behaviours and genetic diversity within tumour samples, which is critical for understanding the complexity of cancer. These challenges can be addressed with single-cell sequencing.

Single-cell sequencing remains one of the most powerful methods for studying the molecular substructure of tumours at a resolution high enough to identify the clones dominating the tumour mass but also rare subpopulations that may play a crucial role in cancer progression under therapy (121-123). In cancer research, sc-seq is employed to understand ITH, trace evolutionary dynamics, identify cellular subpopulations within tumours resistant to therapy, examine the impact of the TME on cancer progression, and unravel the complexity of T-cell diversity and function among other applications (116, 124, 125).

Particularly interesting is the application of sc-seq to identify resistant subpopulations. Primary resistance is inherent to tumours, with clonal populations harbouring resistant genetic alterations and phenotypes emerging from normal tumour progression even before the start of a treatment (126). On the other hand, cancer relapse in initial responders is often driven by cells that have acquired resistant traits following drug exposure. Sc-seq techniques allow researchers to understand how cancer treatment influences the evolutionary dynamics of cancer cells at an individual cell level. As an example, to investigate clonal evolution in response to neoadjuvant chemotherapy (NAC) with epirubicin, docetaxel and bevacizumab in triple-negative breast cancer (TNBC) patients, Kim *et al*. analysed longitudinal samples from 20 patients using bulk DNA sequencing (DNA-seq), single-cell DNA sequencing (scDNA-seq), and single-cell RNA sequencing (scRNA-seq) (127).

Initial bulk whole-exome sequencing (WES) of matched pre-treatment, mid-treatment, and post-treatment samples revealed that, in 50% of the patients, no detectable somatic mutations

were found in post-treatment samples compared to pre-treatment tumours (127). This suggests that NAC successfully eliminated cancer cells in this "clonal extinction" patient group. These results were later corroborated by single-nucleus DNA-sequencing (SNDS) and fluorescence-activated cell sorting (FACS), which identified multiple cancer cell clones in pre-treatment tumours but found only diploid cells in post-treatment samples. Genomic analysis of pre-treatment clones identified shared chromosome breakpoints and mutations in specific cancer-related genes (e.g., *MET*, *MYC*, *PTEN*), suggesting a shared evolutionary origin among pre-treatment clones in this group. The study also performed single-nucleus RNA sequencing (SNRS) to analyse the transcriptome in cells from pre- and post-treatment samples. SNRS revealed that post-treatment cells exhibited gene expression patterns typical of normal cells, thus confirming the genomic findings. The absence of cancer cells in post-treatment samples, as evidenced by both genomic and transcriptomic analyses, highlights the effectiveness of NAC in these patients.

In contrast, the remaining patients exhibited residual mutations after treatment ("clonal persistence" group), albeit at reduced frequencies, suggesting the presence of resistant genotypes that survived NAC (127). In this second group, new mutations also emerged in genes involved in pathways related to cell proliferation, apoptosis, solute transport, and cytoskeleton regulation. Moreover, targeted deep-amplicon sequencing revealed that in half of these patients, new mutations were already present at very low frequencies in the pre-treatment tumours, suggesting an adaptive resistance mechanism.

Resistant cells in the clonal persistence group had specific CNAs (mainly chromosomal deletions) and gene mutations absent in the clonal extinction group, highlighting the genetic basis for chemotherapy resistance (127). Subsequent transcriptomic analysis using SNRS identified gene expression signatures in resistant cells that differed from those in the clonal extinction group, including pathways related to ECM degradation, PI3K/AKT1/mTOR signalling, and hypoxia, all of which are known to contribute to chemotherapy resistance.

Furthermore, the study investigated the evolution of cancer cell phenotypes in response to NAC and found that chemoresistant cells in the clonal persistence group did not express their complete transcriptional resistance programs before treatment (127). Instead, they expressed a subset of genes indicative of partial resistance, later developing full resistance profiles under the selective pressure of chemotherapy. This suggests an adaptive response to NAC, where cancer cells dynamically adjust their genomic and transcriptomic profiles to overcome challenges imposed by the treatment.

1.5. Leveraging single-cell sequencing to explore cancer evolution

This research highlights the application of scDNA-seq and scRNA-seq for unveiling cellular heterogeneity, tracking clonal evolution over time, deciphering mechanisms of chemotherapy resistance, and revealing transcriptional programs of resistance.

1.5.3 Advancing cancer research with single-cell multiomics

Despite significant advances in sc-seq technologies, analysing only one type of "omics" data from each cell, such as DNA, RNA, or protein, offers a limited view of the complex interplay between various molecular layers that define cellular states and functions.

To bridge this gap, the field has turned to single-cell multiomics, which captures and combines information from multiple molecular layers within the same cell (116). By integrating data from genomic, transcriptomic, and proteomic analyses, single-cell multiomics provides a more comprehensive analysis, enabling the correlation of genomic variations with functional outcomes at both the transcriptome and proteome levels. This integration is indispensable for deciphering the complex mechanisms of cellular biology and unravelling disease mechanisms.

Several methodologies have been developed to facilitate single-cell multiomics analyses. Whether they are low throughput (i.e., plate-based) or high throughput (droplet-based or microfluidics), multiomics methods typically target the genome and transcriptome (e.g., G&T-seq, DR-seq, SIDR, Target-seq), the transcriptome and chromatin accessibility (e.g., scM&T-seq, scMT-seq), or the transcriptome and proteome (e.g., CITE-seq, SCITO-seq), with a few approaches capturing more than two omics layers (e.g., ScTrio-seq, iscCOOL-seq) (Figure 1.9) (116, 128-138).

The combined characterisation of these molecular layers allows scientists to determine whether changes observed in one layer are consistently reflected in the corresponding layer within the same cell. This aspect is particularly intriguing for examining somatic single-nucleotide variants (SNVs) and copy number aberrations (CNAs) affecting coding genes. If mutant genes are transcribed and observed at the transcriptional level, they may also be translated into functional proteins with abnormal functions. In this context, single-cell Genome and Transcriptome sequencing (G&T-seq) stands out as the method to simultaneously capture these molecular alterations, ultimately providing invaluable insights into the molecular mechanisms driving cancer progression and resistance to therapy (129, 132).

The key finding in the G&T-seq study was the identification of an additional copy of chromosome 11 within a subpopulation of HCC38-BL cells, a diploid B lymphoblastoid cell line used as a normal control against HCC38, a breast cancer cell line derived from the same patient that lacked this particular copy number gain (129). Furthermore, the integrated genomic and transcriptomic datasets from the same cells indicated that the trisomy of chromosome 11 was associated with increased gene expression on this chromosome. This discovery highlights the sensitivity of G&T-seq for detecting allele-specific gene expression, which is indispensable for

unravelling cellular heterogeneity within populations that might otherwise appear homogeneous.

In summary, by leveraging multiomics techniques, researchers can gain valuable insights into the mechanisms driving cellular diversity, disease progression, and response to treatments.

1.5. Leveraging single-cell sequencing to explore cancer evolution

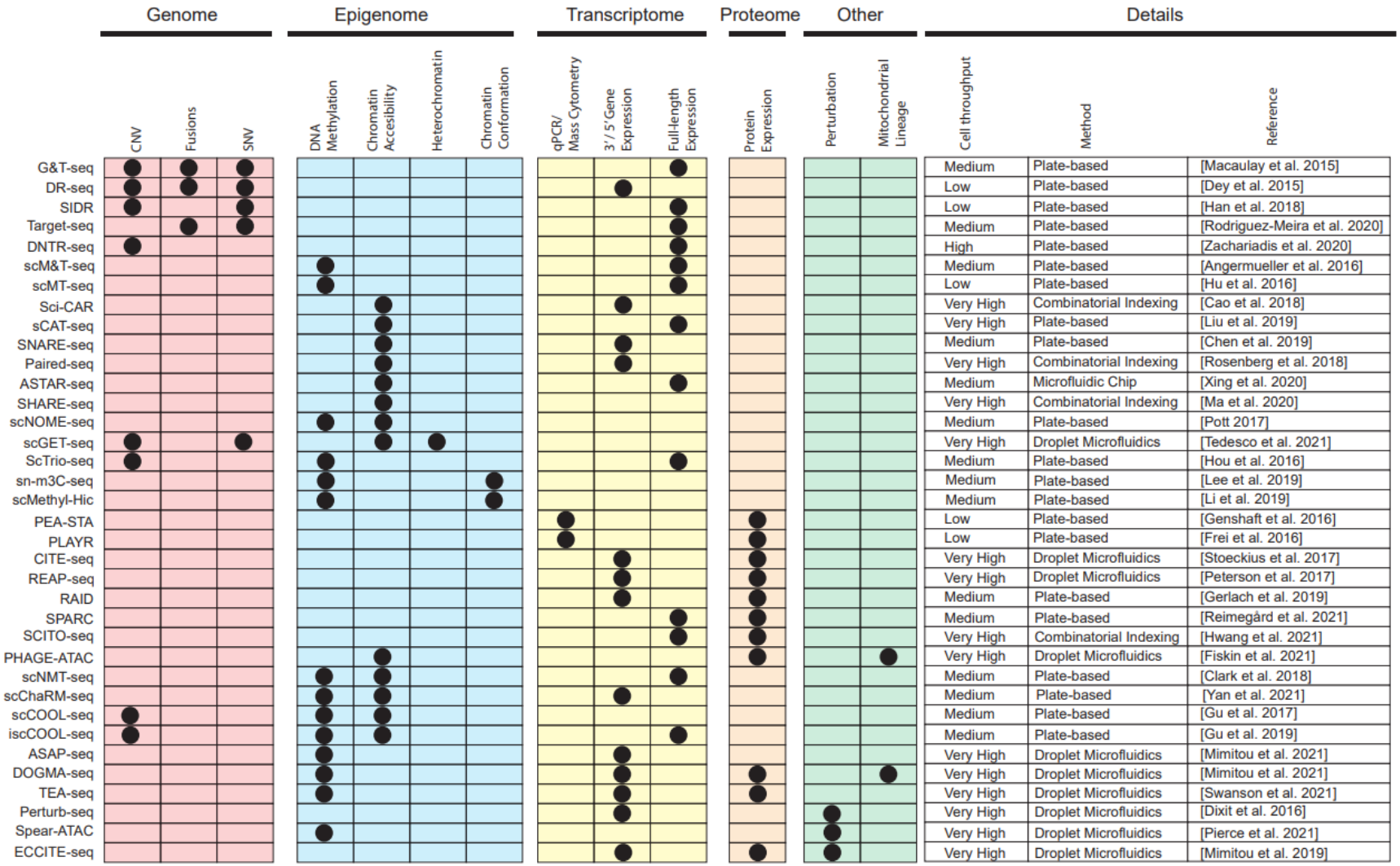


Figure 1.9. Summary of single-cell multiomics methods.

Figure reproduced from (116).

1.5.4 Limitations of single-cell sequencing

Although single-cell sequencing provides detailed insights into individual cellular functions and heterogeneity, the complex nature of these methodologies introduces several technical and biological challenges. One initial challenge involves dissociating cells from tissues, which can induce stress and alter gene expression profiles (139). Additionally, isolating viable single cells without contamination can be difficult, though methods such as fluorescence-activated cell sorting (FACS) and droplet-based approaches (e.g., Drop-seq and 10x Genomics) are effective for isolating single cells.

Another significant technical challenge is the low starting input material, which can hinder the reverse transcription of RNA and introduce variability in the amplification process (139). This issue is exacerbated by preferential amplification of certain genomic loci or transcripts over others, leading to allelic dropout events where one or both alleles of a gene fail to be detected. Such events also result in dropouts in gene expression, where genes that were actually expressed in a cell are not detected in the sequencing data. Moreover, insufficient sequencing depth may fail to capture all expressed genes, particularly those with low expression levels, and sequencing errors can lead to the incorrect identification of allelic variants.

In the context of scRNA-seq, these technical challenges can lead to an underestimation of the expression levels of certain genes, particularly those expressed at low levels (140). This can result in the incorrect characterisation of cell states or types, potentially skewing biological interpretations. Similarly, for scWGS, these limitations can lead to incorrect interpretations of clonal diversity and evolution within populations of cells. Employing sensitive and accurate whole-transcriptome amplification methods that improve the efficiency of reverse transcription can help reduce dropouts in gene expression. Additionally, using more uniform and less biased whole-genome amplification techniques can help mitigate allelic dropouts (140).

Biological challenges stem from the heterogeneity in biological samples. For instance, significant heterogeneity in gene expression among cells complicates cell type identification and classification (139). To address this, clustering algorithms are used to identify cell subpopulations based on gene expression profiles, while gene set enrichment analysis (GSEA) helps identify enriched pathways or functional categories within each subpopulation.

In addition, although scRNA-seq can detect rare cell populations that bulk RNA-seq may miss, identifying these populations is challenging due to low cell numbers typically sequenced (140). Using molecular barcodes such as unique molecular identifiers (UMIs) and full-length

transcript approaches like smart-seq2, which have higher sensitivity, can detect low-abundance transcripts, facilitating the identification of rare cell populations.

While scRNA-seq provides gene expression information at the single-cell level, it lacks spatial context. Combining scRNA-seq with spatial transcriptomics techniques, such as the 10x Genomics Visium platform or multiplexed fluorescence in situ hybridisation (FISH) methods like MERFISH (141) and STARmap, can reveal gene expression patterns within their spatial context, addressing spatial heterogeneity (117).

scRNA-seq captures a snapshot of gene expression at a single time point, but cells undergo dynamic changes in response to stimuli or environmental cues (139). Time-resolved scRNA-seq, pseudo-time analysis, trajectory inference algorithms, and integration with other omics data can capture these dynamic changes, enabling the reconstruction of cell state transitions over time. Lastly, analysing alternative splicing and gene isoforms is challenging due to data complexity. Long-read sequencing, short-read sequencing with paired-end reads, computational algorithms, and integration with other omics data can identify different isoforms and their functional implications, providing a comprehensive view of gene expression (139).

Overall, single-cell sequencing offers unprecedented insights into cellular heterogeneity and gene expression patterns but presents significant challenges that can impact the accuracy, efficiency, and reliability of the sequencing results. Addressing these challenges through optimised protocols, advanced computational methods, and integrative approaches can enhance the accuracy, reproducibility, and interpretability of single-cell sequencing data.

1.6 PhD aims and objectives

This PhD project was part of a more extensive collaboration funded by the CRUK/AIRC Accelerator Award. This award facilitates partnerships among European scientific teams to develop tools and techniques that meet critical medical needs. The collaboration included several UK (Institute of Cancer Research and Earlham Institute) and Italian research institutes (Ospedale San Raffaele, Politecnico di Milano, and Human Technopole).

The project, titled "Single-cell cancer evolution in the clinic," aimed to grow patient-derived organoids (PDOs) of metastatic colorectal cancers (mCRCs)—initially established by Vlachogiannis et al., in 2018—within microfluidic devices for drug screening purposes (109). PDOs are typically established from primary tumours, whereas those derived from metastatic sites are less common and less frequently studied (142). Given that metastatic cancer cells harbour more genetic and non-genetic changes, are more challenging to treat, and are more prone to developing resistance to cancer treatments compared to their primary tumour counterparts, PDOs derived from metastatic sites offer a novel perspective on tumour evolution and resistance.

By performing bulk and single-cell multiomics techniques on these mCRC PDOs, the collaboration sought to understand how tumours evolve when subjected to targeted drug treatments and to create novel computational models to predict tumour progression and drug resistance. These models could help personalise treatment plans based on the predicted behaviour of individual cancers, leading to more tailored and potentially more successful treatment approaches. The specific objectives of this overarching project were executed through four mutually inclusive programmes, with this PhD project falling under Programme 2 (PR2).

The primary aim of this PhD project was to elucidate the mechanisms of resistance to AKT inhibitors in mCRC PDOs using single-cell multiomics. Despite the mCRC PDOs being treatment-naïve to the AKT inhibitors, their derivation from a heavily pre-treated patient hints at the presence of resistant cells prior to AKT inhibition. Consequently, the primary hypothesis of this study was that alterations contributing to drug resistance would manifest in an interconnected manner at both the genomic and transcriptional levels even before AKT treatment in mCRC organoids.

To test this hypothesis, several PDO lines were established from F-016, the only mCRC PDO in Vlachogiannis's biobank with both a somatic mutation and an amplification affecting *AKT1*, which had responded strongly to AKT inhibitors in the drug library (109). Once the untreated,

1.6. PhD aims and objectives

Parental organoid lines were established, they were driven to resistance through prolonged exposure to two AKT inhibitors, namely MK-2206 and AZD5363. Given the significant role of intratumour heterogeneity (ITH) in contributing to varied therapeutic responses, this PhD project leveraged single-cell multiomics techniques to analyse the molecular characteristics of individual cells within PDOs. The mCRC PDO cultures were established externally by our collaborators at the Institute of Cancer Research. The single cells from these PDOs, sorted into 96-well plates, were processed, sequenced, and computationally analysed at the Earlham Institute. For single-cell multiomics, cells were processed following parallel Genome and Transcriptome sequencing (G&T-seq) (129, 132). Following the separation of nucleic acids, single-cell transcriptomes were amplified using Smart-seq2 (143), while matched single-cell genomes were amplified with the PicoPlex Gold whole genome amplification kit (Takara). Both cDNA and matching gDNA libraries were subsequently sequenced using Illumina short-read sequencing technology and computationally analysed using several bioinformatics tools.

Transcription provides insights into the cell's physiological state, behaviour, and potentially its identity and function. To understand how specific cell subpopulations responded to the two AKT inhibitors and how these responses evolved with the development of resistance, we employed the scRNA-seq data for clustering analysis, annotating colonic cell types, and comparing the gene expression patterns between control and resistant mCRC PDOs. Additionally, copy number alterations (CNAs) were inferred from the single-cell transcriptomes, which provided insights into the genomic changes accompanying resistance phenotypes.

Bioinformatics analyses of the matched single-cell genomes focused on identifying differences in CNAs between untreated and resistant mCRC organoids. By comparing the genomic CNAs to the transcriptome-based CNAs, we identified and validated changes in the clonal dynamics that contributed to the development of drug resistance.

Lastly, by integrating matched scRNA and scDNA data, we identified differentially expressed genes located within regions of CNAs, suggesting their potential role in driving the resistance mechanisms that helped bypass AKT blockade. This detailed study of mCRC cells using G&T-seq revealed new molecular targets for drug development, which could potentially counteract resistance to MK-2206 and AZD5363.

Altogether, the G&T-seq methodology, complemented by a wide array of bioinformatics tools, provided a comprehensive picture of the genetic and transcriptional adaptations that mCRC cells employed to evolve despite AKT pathway inhibition. These insights were only possible

through an integrated, single-cell multiomics approach, rather than by analysing these omics independently or by employing bulk sequencing methods.

Overall, this research underscores the power of single-cell multiomics to uncover detailed evolutionary trajectories and adaptive mechanisms in cancer under therapeutic intervention. By understanding and targeting these adaptive mechanisms, researchers can develop more effective strategies for managing treatment resistance, ultimately improving the efficacy of cancer therapies.

Chapter 2

Materials & Methods

Chapter disclosures

This chapter describes the methods employed to generate data for Chapters 3 to 5. All experiments were performed by Silvia Ogbeide unless stated otherwise in each section.

2.1 Materials and key resources

Table 2. Biological samples generated

mCRC PDO	Source	Generated by
3994-117 Parental PDO	Joint collaboration	The Institute of Cancer Research, London, SM2 5NG
3994-117 MK-2206-resistant PDO		
3994-117 AZD5363-resistant PDO		

Table 3. G&T-seq reagents

Reagent	Source	Catalogue number
Buffer RLT Plus	QIAGEN	1053393
10 M NaOH	Merk	72068
5 M NaCl	Invitrogen	AM9760G
Nuclease-free water	Invitrogen	AM9937
UltraPure 1 M Tris–HCI Buffer, pH 7.5	Invitrogen	15567027
0.5 M EDTA Solution	Promega	V4231
0.1 M Trizma Pre-set crystals, pH 8.3	Merk	T8943
2 M KCl, RNase-free	Invitrogen	AM9640G
1 M $MgCl_2$	Invitrogen	AM9530G
1 M DTT	Merk	646563
50% (vol/vol) Tween 20	Thermo Fisher Scientific	003005
Dynabead MyOne Streptavidin C1	Invitrogen	65001
5X First-strand buffer	Invitrogen	18064071
SUPERase•In RNase Inhibitor (20 U/µL)	Invitrogen	AM2696
10 mM dNTP Mix	Invitrogen	18427013
5 M Betaine solution	Merk	B0300
100 mM DTT	Invitrogen	18064071
SuperScript II Reverse Transcriptase	Invitrogen	18064071
KAPA HiFi HotStart ReadyMix	Roche	KK2602
Ethanol absolute ≥ 99.8%	VWR	437435L
Beckman Coulter Agencourt AMPure XP	Fisher Scientific	10453438

2.1. Materials and key resources

Table 4. G&T-seq primers

Oligonucleotides	Source	Sequence
Biotinylated Oligo-dT30VN Primer	IDT	5′-Biotin-TEG-AAGCAGTGGTATCAACG CAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′
IS PCR Primer	IDT	5′-AAGCAGTGGTATCAACGCAGAGT-3′
Template-switching oligo	Qiagen	5′-AAGCAGTGGTATCAACGCAGAGTACA TrGrG+G-3′

Table 5. G&T-seq buffers

Dynabead solution A		
Component	**Final molarity required (M)**	**Volume required for 50 ml (ml)**
Nuclease-free water	-	49
10 M NaOH	0.1	0.5
5 M NaCl	0.05	0.5
Dynabead solution B		
Component	**Final molarity required (M)**	**Volume required for 50 ml (ml)**
Nuclease-free water	-	49
5 M NaCl	0.1	1
Dynabead 2X 'Binding and Wash' buffer		
Component	**Final molarity required (M)**	**Volume required for 50 ml (ml)**
Nuclease-free water	-	29.4
1 M Tris-HCl (pH7.5)	0.01	0.5
0.5 M EDTA Solution	0.001	0.1
5 M NaCl	2	20.0
G&T-seq wash buffer*		
Component	**Final molarity required (M)**	**Volume required for 50 ml (ml)**
Nuclease-free water	-	21.8
0.1M Tris-HCl (pH 8.3)	0.05	25
2 M KCl	0.075	1.875
1 M $MgCl_2$	0.003	0.3
1 M DTT	0.01	0.5
50% (vol/vol) Tween 20	0.50%	0.5

*Requires supplementation with RNAse Inhibitor immediately before use

2.1. Materials and key resources

Table 6. Commercial assays

Reagent kit	Source	Catalogue number
Qubit dsDNA Quantitation, High Sensitivity kit	Invitrogen	Q32851
Agilent High Sensitivity DNA Kit	Roche	5067-4626
Nextera XT DNA Library Preparation Kit (96 samples)	Illumina	FC-131-1096
Nextera XT Index Kit v2 Set A (96 indexes, 384 samples)	Illumina	FC-131-2001
Nextera XT Index Kit v2 Set B (96 indexes, 384 samples)	Illumina	FC-131-2002
Nextera XT Index Kit v2 Set C (96 indexes, 384 samples)	Illumina	FC-131-2003
KAPA Library Quantification Kit	Roche	KK4854
PicoPLEX Gold Single Cell DNA-Seq Kit	Takara	R300670
DNA HT Dual Index Kit - 96N Set A	Takara	R400660

2.1. Materials and key resources

Table 7. Equipment

Instrument	Source	Catalogue number
Eco UV Cabinet	HCE	LAB321
Legacy UVP Crosslinker CL-1000	UVP	Discontinued since 2020
1.5 ml Eppendorf Safe-Lock microcentrifuge tubes	Merk	T9661
50 ml Falcon high-clarity polypropylene (PP) conical centrifuge tubes	Corning	352070
DynaMa-2 Magnet	Invitrogen	12321D
Fisherbrand Wizard Infrared Vortex Mixer	Fisher Scientific	11746744
Microcentrifuge	Starlab	N2631-0007
Adhesive PCR Plate Seals	Thermo Fisher Scientific	AB0558
FrameStar 96 Well Skirted PCR Plate	Azenta	4ti-0960/C
Multipette E3x electronic multi-dispenser pipette	Eppendorf	4987000029
Refrigerated centrifuge and adaptors for 96-well plates	Eppendorf	5804000060
Mini LabRoller Rotator	Labnet	H5500
Biomek FX Automated Workstation	Beckman Coulter	A31842-5
Biomek NX Automated Workstation	Beckman Coulter	A31841
Mosquito HV liquid handler with 5 plate position deck	SPT Labtech	NA
Low-elution magnet plate	Alpaqua	A000350
96S Super Magnet	Alpaqua	A001322
ThermoMixer C	Eppendorf	5382000031
ThermoTop	Eppendorf	5308000003
Eppendorf SmartBlock PCR 96	Eppendorf	5306000006
C1000 Touch Thermal Cycler with 96-Well Fast Reaction Module	Bio-Rad	1851196
Invitrogen Qubit 4 Fluorometer	Fisher Scientific	15723679
Agilent 2100 Bioanalyzer	Agilent	G2938C
LightCycler 480 Instrument II	Roche	05015278001
NovaSeq 6000 Sequencing System	Illumina	NA

2.2 Methods

2.2.1 Establishment and maintenance of patient-derived tumour organoids from human gastrointestinal cancers

To evaluate the predictive value of organoids in co-clinical trials, Vlachogiannis and colleagues established a living biobank of PDOs from metastatic gastroesophageal and colorectal cancer biopsies from patients enrolled in phase I or phase II clinical trials (109). This enabled a comprehensive comparison of anti-cancer drug responses between patients and their corresponding PDO counterparts.

The PDOs analysed in this thesis were derived from the organoids established by Vlachogiannis *et al*., specifically from a colorectal cancer liver metastasis biopsied via image-guided ultrasound from a patient with stage III CRC (sample ID: 3994-117; publication ID: F-016) (109). This patient was enrolled in the Feasibility of a Molecular Characterisation Approach to Treatment (FOrMAT) trial (144). The liver specimen was obtained from the patient at the time of disease progression to FOLFIRI, a second-line chemotherapy combination of irinotecan, folinic acid (CF) and fluorouracil (5-FU). FOLFIRI is administered to patients with metastatic colorectal cancer who did not respond optimally to first-line treatment involving oxaliplatin combined with 5-FU, or 5-FU prodrugs such as capecitabine, (145), which the patient had received previously.

All GI PDOs, including 3994-117/F-016, were developed at the Institute of Cancer Research (ICR) in London, UK, using a protocol described by Vlachogiannis *et al*. (109). Briefly, patient samples for organoid derivation were immediately placed in ice-cold PBS after collection and transported to the laboratory for processing. Patient specimens were minced in the lab, and the tissue fragments were washed with 5 ml of 5X PBS supplemented with EDTA (Thermo Scientific Chemicals) for 15 min at room temperature. Subsequently, digestion was performed in 5 ml of 1X PBS-EDTA supplemented with 2X TrypLE Select Enzyme (Gibco) for 1 hr at 37 °C. The digested tissue suspensions were then sheared through several rounds of pipetting to facilitate cell release. Isolated cells were collected in Advanced DMEM/F-12 (Gibco), centrifuged at 1,200 rpm for 5 min at 4 °C, and resuspended in 120 µl of growth factor reduced (GFR) Matrigel (Corning). Matrigel-cell suspensions were plated into a single well of a 24-well plate (Corning) and incubated at 37 °C with 5% CO_2 for 20 min to allow Matrigel polymerisation. Finally, cells were overlaid with 500 µl of complete culture medium. The culture medium consisted of Advanced DMEM/F-12 supplemented with 1X B-27 (Gibco), 1X N-2 (Gibco), 0.01% BSA (Roche), 2 mM L-Glutamine (Gibco) and 100 units/ml penicillin-streptomycin (Gibco). Additionally, various growth factors and other additives were included

in the growth media to promote organoid development. These are described in Table 8. The organoid medium was refreshed every other day.

When PDOs reached confluency (approximately after two weeks, depending on the organoid), they were collected with 1X PBS-EDTA containing 1X TrypLE Select Enzyme, incubated for 20 minutes at 37 °C and mechanically sheared by pipetting. The resulting cell suspensions were washed with HBSS (Gibco), pelleted at 1,200 rpm for 5 min at 4 °C and resuspended in GFR Matrigel before reseeding at the desired ratio. Alternatively, organoid pellets were resuspended in FBS (Gibco) containing 10% DMSO (Merck) and cryopreserved at -80 °C in multiple vials for future experiments.

Table 8. Growth factors and culture media additives for the development of GI PDOs

Additive	Concentration	Source	Catalogue number
EGF	50 ng/ml	PeproTech	AF-100-15
Noggin	100 ng/ml	PeproTech	250-38
R-Spondin 1	500 ng/ml	PeproTech	120-38
Gastrin	10 nM	Merck	G9145
FGF-10	10 ng/ml	PeproTech	100-26
FGF-basic	10 ng/ml	PeproTech	100-18B
Wnt-3A	100 ng/ml	Bio-Techne	5036-WN
Prostaglandin E_2	1 µM	Bio-Techne	2296
Y-27632	10 µM	Merck	Y0503
Nicotinamide	4 mM	Merck	N0636
A83-01	0.5 µM	Bio-Techne	2939
SB202190	5 µM	Merck	S7067

2.2.2 Generation of mCRC PDOs resistant to AKT inhibition

Bulk whole-genome sequencing of the 3994-117/F-016 mCRC PDO revealed an amplification of *AKT1*, accompanied by an E17K somatic mutation (109) (Figure 2.1A). Notably, this PDO was the only one in the organoid biobank established by Vlachogiannis *et al*., to exhibit a strong response to 1 μM of the AKT inhibitor MK-2206 (Figure 2.1B).

To create mCRC PDOs resistant to AKT inhibition, our collaborators at the ICR cultured 3994-117 in the presence of AZD5363 (also known as capivasertib, Selleck Chemicals) or MK-2206 2HCl (Selleck Chemicals) (146), until drug resistance was observed. These pan-AKT inhibitors (AKTi) induce autophagy and apoptosis, and have demonstrated their efficacy in clinical trials involving various solid tumours, including colon cancer (147-149). Recently, capivasertib was approved by the FDA in combination with fulvestrant for the treatment of adults with hormone receptor-positive, HER2-negative breast cancer that is either locally advanced or metastatic, and has one or more biomarker alterations in *PIK3CA*, *AKT1* or *PTEN* (150). This approval makes capivasertib the first AKT inhibitor available on the market.

Briefly, to generate the MK-2206 and AZD5363 resistant lines, the established Parental PDO was dissociated using the passaging procedure described in the previous section, after which the cell suspension was seeded into three wells—serving as technical replicates—of a 12-well plate for each treatment condition (Figure 2.2). Three days after seeding, the complete organoid media was replaced with fresh media containing DMSO (for the vehicle control) or 1 μM of either MK-2206 2HCl or AZD5363 (both dissolved in DMSO) for the drug-treated PDO lines. The media was refreshed every other day until drug resistance was observed in the treated PDOs, indicated by the organoids returning to full confluency after an initial period of cell death. This was observed after 35 days for the MK-2206 and 55 days for AZD5363 treated PDO. Additionally, the cell viability of organoids was assessed by adding 10% of CellTiter-Blue Reagent (Promega) directly to the PDOs, followed by incubation. Viability readings were then obtained using the EnVision plate reader (PerkinElmer).

To ensure the accurate generation of resistant PDO lines, the entire seeding and treatment process was repeated two additional times, resulting in a total of nine wells per condition. Each of these iterations started with a fresh sample of Parental cells, allowing for the assessment of biological variability in the PDOs. Once established, MK-2206-resistant (MK1-resistant) and AZD5363-resistant (AZD1-resistant) organoids were dissociated and collected as described in the previous section. The resulting organoid pellets were resuspended in FBS (Gibco) containing 10% DMSO (Merck) and cryopreserved at -80 °C for future experiments.

2.2. Methods

In contrast to the extended culturing period of the resistant organoids, the untreated Parental PDO was cultured for only two weeks due to the wells becoming very confluent. To ensure that clonal dynamics were driven by the drug treatment and not merely by the duration of PDOs in culture, continuous passages of the Parental control were analysed using low-pass WGS at the end of the experiment. The copy number alterations (CNAs) of the two-week Parental control were compared to those from multiple subsequent passages. The CNA profiles remained nearly identical before and after, indicating that despite the prolonged culture of the control organoid over multiple passages, there was no expansion of specific populations over previous seeds. This finding suggests that the observed clonal dynamics in the drug-treated organoids were likely a result of the drug treatment rather than the extended culture time.

It is also important to highlight that the bulk and single-cell experiments described in the following sections were carried out over a two-year period. The bulk and high-throughput sequencing experiments occurred before the start of this PhD project, while the G&T-seq experiment was conducted over a year later. For each of these experiments, the cryopreserved PDOs were revived and expanded in complete growth media without the inhibitors, and at different ratios for bulk and single-cell sequencing experiments. Consequently, practical considerations required using different passages over the two-year period during which these sequencing experiments were performed. These passages were kept similar to maintain consistency. However, the stability of CNVs previously observed in the parental control suggested that this should not have introduced significant differences.

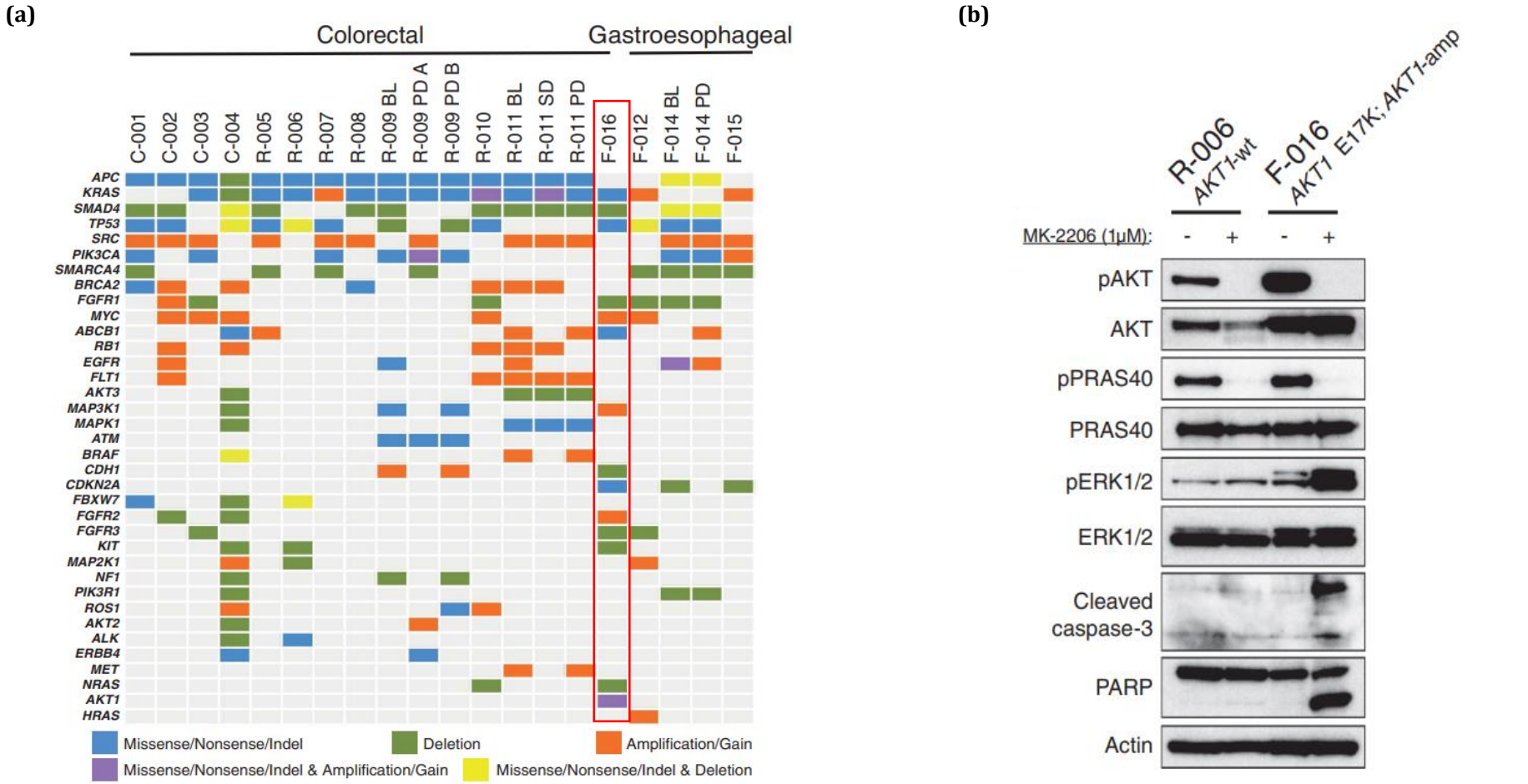


Figure 2.1. Molecular characterisation of 3994-117/F-016 mCRC PDO.

(a) *Molecular landscape of F-016.* ***(b)*** *Pathway analysis after AKT inhibition (4 hours) with MK-2206 in AKT1 wild type (R-006) and mutant (F-016) PDOs. Figure reproduced from (109).*

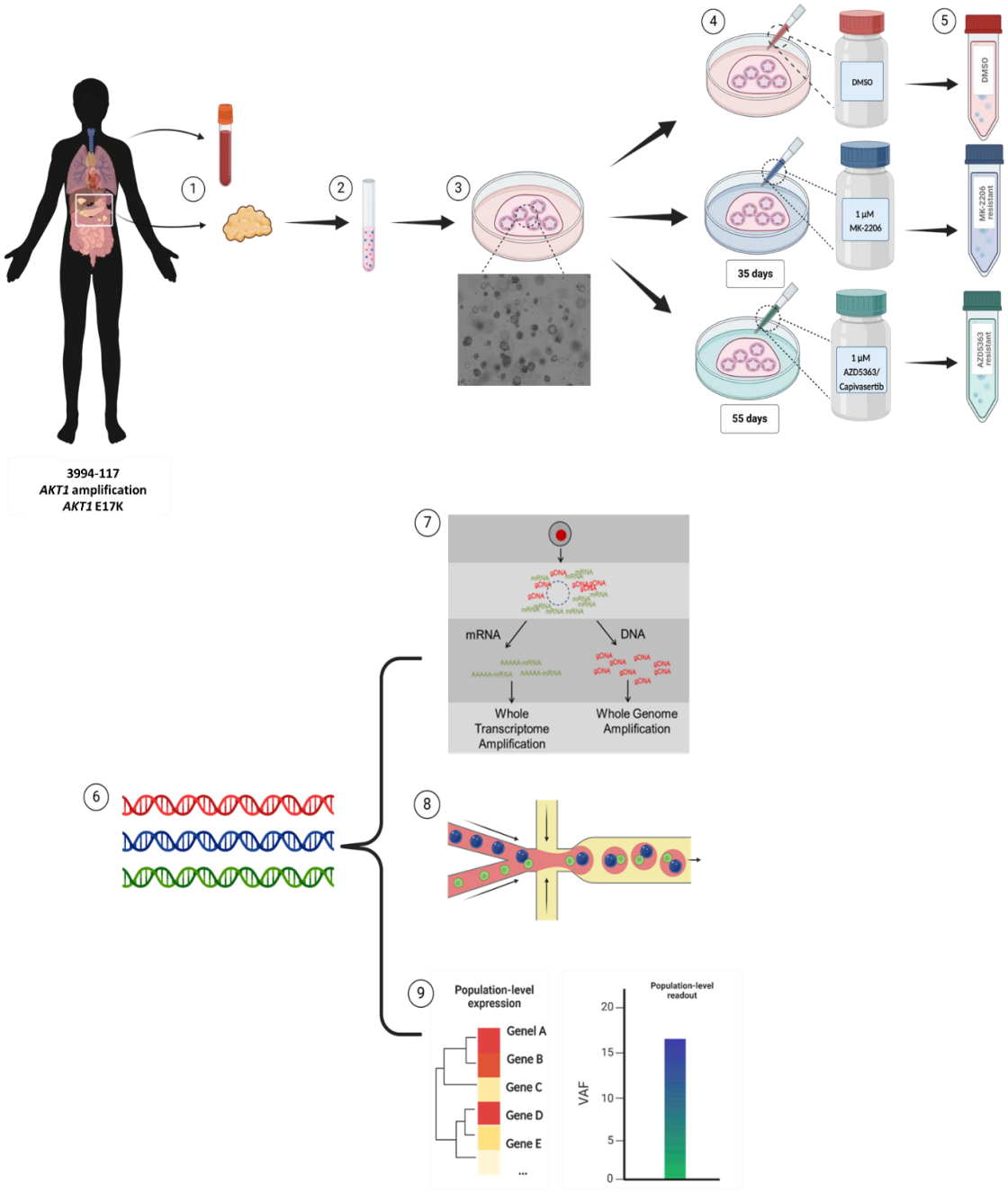


Figure 2.2. Experimental outline.

(1) *Collection mCRC biopsy and a matched blood sample from donor.* ***(2)*** *Tissue dissociation of biopsy into cell suspensions.* ***(3)*** *Establishment of mCRC PDO cultures.* ***(4)*** *Generation of mCRC PDOs resistant to AKT inhibition.* ***(5)*** *Dissociation of mCRC PDOs into cell suspensions for fluorescence-activated cell sorting (FACS).* ***(6)*** *cDNA/DNA library preparation for* ***(7)*** *single-cell genome and transcriptome sequencing (G&T-seq), (8) droplet-based single-cell sequencing, and* ***(9)*** *bulk sequencing. Figure created with BioRender.com.*

2.2.3 Single-cell sequencing of mCRC PDOs

A. Sample preparation

At the ICR, PDOs were dissociated as previously described by Erika Yara (Senior Scientific Officer, Dr Andrea Sottoriva's group). The resulting cell suspensions were washed, resuspended in PBS, and filtered through a 40 μm Flowmi cell strainer. Cell stock concentrations were adjusted to 700-1,200 cells/μl in preparation for subsequent plate- and droplet-based single-cell sequencing approaches. Cell viability was determined using the trypan blue dye exclusion test and quantified with the Countess Automated Cell Counter (Invitrogen), which confirmed that all samples had viability exceeding 90%.

B. Physical separation of mRNA and genomic DNA from single cells

For plate-based single-cell genome and transcriptome sequencing (G&T-seq) (129, 132), the CellenOne image-based single-cell sorter was used to dispense single cells, 50 cells and empty drops (representing positive and negative control wells) into individual wells of 96-well plates preloaded with 5 μl of RLT plus lysis buffer. After cell sorting, plates were sealed, briefly centrifuged at 1,000 rpm at 4 °C, and stored at -80 °C until they were shipped to the Earlham Institute for processing.

Genomic DNA (gDNA) and mRNA were separated following the G&T-seq protocol (132) (Figure 2.3). First, MyOne Streptavidin C1 Dynabeads were washed with Dynabead solutions A and B, mixed in a 1:1 ratio with 100 μM of biotinylated oligo-dT30VN primers and incubated at room temperature with gentle rotation for 30 min. Following incubation, oligo-dT30VN-conjugated beads were washed with Dynabead 1X "Binding and Wash" buffer and then resuspended in 1 ml of bead resuspension buffer (1× Superscript First-strand buffer, 1 U/μl RNase inhibitor, nuclease-free water (NF-H_2O)). Next, the sample plate was thawed on ice, and 10 μl of the "bead mix" was arrayed into each well of the plate. The plate was then incubated on a Thermomixer at 1,300 rpm for 30 min at room temperature. The gDNA and mRNA were physically separated using the Biomek FX Automated Liquid Handler equipped with a low-elution 96-well plate magnet. During this process, the mRNA-oligo-d(T) beads complex is pulled to the side of the well by the magnet, while the gDNA remains in the supernatant, which is then aspirated and transferred to a new plate. The gDNA fraction was briefly centrifuged and stored at -20 °C for future processing. On the other hand, the mRNA plate was immediately processed.

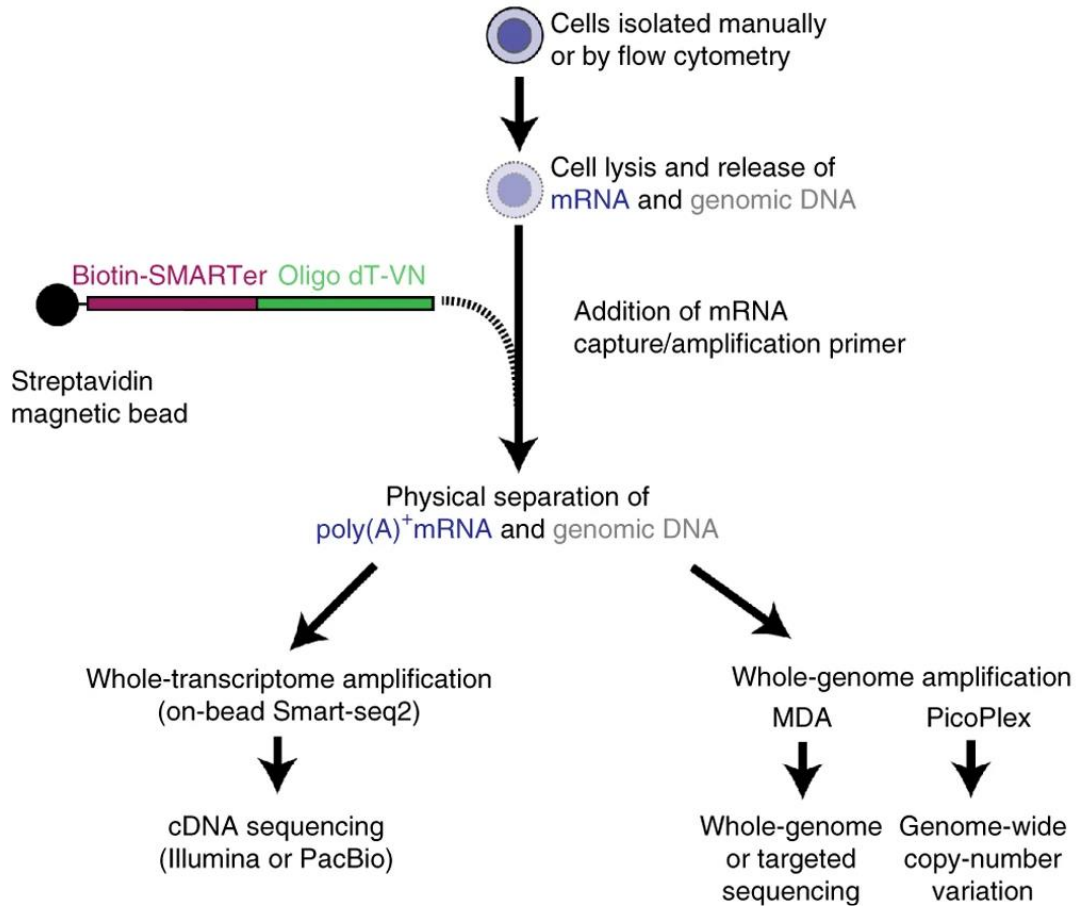


Figure 2.3. G&T-seq protocol overview.

Single cells are isolated manually or by fluorescence-activated cell sorting (FACS) into individual wells within a microwell plate (e.g., 96-well plate) containing RLT lysis buffer. This process leads to the release of polyadenylated (poly(A)+) mRNA and genomic DNA (gDNA) into the media upon centrifugation. Physical separation of mRNA from gDNA is achieved using biotinylated oligo-dT primers, which selectively bind to and capture poly(A)+ mRNA, leaving behind the gDNA, which is subsequently transferred into a new microwell plate for later processing. Nucleic acid amplification is a critical step to generate sufficient sequencing material from the low-input mRNA and gDNA. While whole-transcriptome amplification (WTA) is performed following the Smart-seq2 protocol (151), whole-genome amplification (WGA) can be performed using either PicoPLEX amplification, typically for studies focusing on genome-wide copy number alteration (CNA) analyses, or multiple-displacement amplification (MDA), which is used for identifying single-nucleotide variants (SNVs). Sequencing libraries can then be constructed from the amplified cDNA and gDNA. During this step, indices are integrated to enable the pooling of multiple cells and samples for subsequent Illumina short-read sequencing. Any of these steps can be carried out using dispensing robots to automate the processes and enhance accuracy. Figure reproduced from (129).

C. Smart-seq2 whole transcriptome amplification, library preparation and RNA-sequencing

To generate full-length cDNA libraries, a modified version of the Smart-seq2 protocol (151, 152) was performed. For this purpose, 5 µl of reverse transcription master mix (10 µM dNTP mix, 100 µM TSO, 1 M $MgCl_2$, 5 M Betaine, 5× Superscript First-strand buffer, 100 mM DTT, 200 U/µl SuperScript II Reverse Transcriptase, 20 U/µl RNAse inhibitor, NF-H_2O) was added to all wells containing mRNA on beads. The plate was then loaded onto a Thermomixer, and the reaction took place with the following settings: 42 °C for 2 min at 2,000 rpm, 42 °C for 60 min at 1,500 rpm, 50 °C for 30 min at 1,500 rpm and 60 °C for 10 min at 1,500 rpm. After the RT step, 7.5 µl of PCR master mix (2× KAPA HiFi HotStart ReadyMix, 10 µM IS PCR primers, NF-H_2O) was added to the samples. PCR amplification was performed in a Thermal cycler using the following cycling programme: 98 °C for 3 min; 20 cycles of 98 °C for 20 s, 67 °C for 15 s and 72 °C for 6 min; 72 °C for 5 min; and holding at 4 °C.

The amplified cDNA was purified on a Biomek NX with a 96-well plate super magnet. Equilibrated AMPure XP beads were added to the samples at a 0.8:1 ratio. After a 5 min incubation at room temperature, the supernatant was discarded, and the beads were washed twice with 50 µl of 80% ethanol. The beads were then left to air dry for 10 min before resuspending in 20 µl of NF-H_2O. The purified cDNA was eluted from the beads, and finally, cDNA concentration was determined using the Qubit dsDNA High Sensitivity kit, while the size distribution was assessed using the Agilent 2100 Bioanalyzer with a High Sensitivity chip.

To construct Nextera XT libraries, cDNA plates were normalised to 0.2 ng/µl. Libraries were prepared using 1/12.5 of the volume recommended in the standard protocol, with Nextera XT Index Kit v2 sets A to C. The Mosquito HV liquid handler was employed to facilitate the process. After the library prep, the 96 cDNA libraries generated from each PDO plate were pooled into 1.5 ml microcentrifuge tubes and manually purified using AMPure XP beads at a 0.6:1 cDNA-to-bead ratio. The quality of PDO-specific library pools was assessed as previously described, in combination with the KAPA Library Quantification Kit on the LightCycler 480 system. Before sequencing, the individual library pools were diluted at equimolar concentrations and then pooled together to achieve a final concentration of 2.5 nM. In this manner, a total of 576 cDNA libraries (192 libraries per organoid) were sequenced in two sequencing runs on a single lane of an SP v1.5 flow cell with the Illumina NovaSeq 6000 system in paired-end mode, generating 150 bp reads. This approach aimed to generate approximately 1.3 million read pairs per library on average.

D. 10x Genomics high-throughput scRNA-sequencing

Single-cell libraries were generated using the Chromium Single Cell 3' Library kit v3.1 (Single Index) protocol (10x Genomics). The final libraries were sequenced on individual lanes of an SP v1.5 flow cell with the NovaSeq 6000 system (150 bp, paired-end mode).

Note: Sample preparation and sequencing were conducted by Javier Fernandez Mateos and Erika Yara at the ICR (Dr Andrea Sottoriva's group).

2.2. Methods

E. PicoPlex Gold whole genome amplification, library preparation and WGS

The gDNA, isolated during the DNA-mRNA separation step of the G&T-seq protocol (see "Physical separation of mRNA and genomic DNA from single cells") was purified on the Biomek FX platform using AMPure XP beads (0.6x ratio) without eluting the DNA. For PicoPlex Gold whole genome amplification (WGA), beads were resuspended in 5 µl of cell extraction master mix (4.8 µl Cell Extraction Buffer, 0.2 µl Cell Extraction Enzyme) according to the protocol, along with 5 µl of NF-H2O. The plate was then placed in a Thermal cycler with the following temperature and time parameters: 75 °C for 10 min, 95 °C for 4 min, followed by holding at 4 °C. Next, 10 µl of pre-amplification master mix (8.7 µl PreAmp Buffer, 1.3 µl PreAmp Enzyme) was added to each well, and the plate underwent a second cycling reaction: 95 °C for 3 min, 16 cycles of 95 °C for 15 s, 15 °C for 50 s, 25 °C for 40 s, 35 °C for 30 s, 65 °C for 40 s, 75 °C for 40 s, and then maintained at 4 °C. The resulting DNA was purified and eluted in 20 µl of NF-H2O using AMPure XP beads in a 1:1 ratio. For library amplification, each well received 25 µl of amplification master mix (20 µl Amplification Buffer, 2.5 µl Amplification Enzyme, 2.5 µl of NF-H2O), along with 5 µl of indexing primers (DNA HT Dual Index Kit - 96N Set A). The amplification reaction was performed using the following cycling programme: 95 °C for 3 min, 4 cycles of 95 °C for 30 s, 63 °C for 25 s and 68 °C for 1 min; 11 cycles of 95 °C for 30 s and 68 °C for 1 min; and holding at 4 °C. Indexed libraries were pooled at equal volumes (10 µl), purified at a 1:1 ratio and eluted in 20 µl of NF-H2O. The purified libraries were quantified using the KAPA Library Quantification Kit on the LightCycler 480 system, and then diluted to achieve a final concentration of 2.5 nM.

A total of 96 DNA libraries, including 30 libraries per PDO (plus 6 control libraries), were sequenced on a single lane of an SP v1.5 flow cell with the NovaSeq 6000 system (150 bp, paired-end mode). This approach aimed to generate approximately 3 million read pairs per library on average.

2.2.4 Bulk WGS of mCRC PDOs

For bulk WGS of the mCRC PDOs and matched blood control, total DNA was extracted using the AllPrep DNA/RNA Mini Kit (Qiagen). DNA libraries were prepared from 100 ng inputs according to the NEBNext Ultra II FS recommendations from New England Biolabs, which included a 20 min enzymatic fragmentation at 37 °C. PCR enrichment of adaptor-ligated DNA was carried out using NEBNext Multiplex Oligos for Illumina with 96 Unique Dual Index Primer Pairs (New England Biolabs). The following cycling conditions were employed: 98 °C for 30 s, 5 cycles of 98 °C for 10 s and 65 °C for 75 s; 65 °C for 5 min; and holding at 4 °C. The DNA libraries were sequenced on an S4 flow cell using the NovaSeq 6000 system (150 bp, paired-end mode).

Note: Sample preparation and sequencing were conducted by Javier Fernandez Mateos and Erika Yara at the ICR.

2.2.5 Bioinformatics and statistical analyses of mCRC PDOs

All computational steps were performed on the Norwich Bioscience Institute (NBI) High-Performance Computing (HPC) cluster using the SLURM workload management system version 23.02.7, along with R v4.1.2 and Python v3.10.3. The software packages used to process and analyse single-cell and bulk data, as well as the specifics of the analyses performed, are described in the "Computational methods" sections of the scRNA-seq and scWGS chapters. These sections detail the various bioinformatics and statistical methods used in the processing, quality control, and analysis of bulk and single-cell RNA-seq and WGS data.

Chapter 3

scRNA-seq profiling of mCRC PDOs

Chapter disclosures

Preprocessing of 10x scRNA-seq data using the Cell Ranger pipeline was performed by Lucrezia Patruno (University of Milano-Bicocca, Milan, Italy), while the subsequent bioinformatics analyses using the high throughput data was performed by Silvia Ogbeide.

3.1 Introduction

Despite advances in chemotherapy and targeted treatments, the five-year survival rate for colorectal cancer (CRC) remains very low (153). Metastatic CRC (mCRC) is generally considered incurable, owing to the complexity associated with treating cancer that has spread to distant organs. However, there are notable exceptions, such as in cases of oligometastatic disease, where metastases are limited in number and confined to surgically resectable locations, such as the liver or lung.

When curative surgery is not an option to remove CRC metastases, the standard treatment regime typically involves a combination of cytotoxic chemotherapy and targeted therapy (153). This approach is mainly palliative, aiming to alleviate the symptoms and improve the quality of life rather than cure the disease, as approximately 90% of mCRC patients will develop resistance to chemotherapy. Resistance to targeted therapies is particularly common, with disease progression frequently observed within 3-12 months of initiating treatment with anti-EGFR monoclonal antibodies (153).

Drug resistance in CRC, whether intrinsic or acquired, primary or multidrug, not only reduces the effectiveness of anti-cancer drugs but also leads to CRC becoming refractory to treatments (154). Addressing this urgent issue requires the development of new therapeutic strategies and interventions and a thorough understanding of the mechanisms employed by cells to evade anti-cancer treatments (154).

The living biobank of patient-derived organoids (PDOs) established by Vlachogiannis *et al*. from metastatic biopsies of colorectal, gastrointestinal cancer (GOC) and cholangiocarcinoma and subsequently submitted to high-throughput screening of FDA-approved or candidate drugs, serves as a prime example of how these 3D models can be used to identify effective treatments for metastatic cancers (109). Although the mutational profiling of these PDOs provided valuable insights, it was somewhat limited, focusing on a panel of 151 cancer-related genes. Moreover, this research primarily sought to assess whether the PDOs responded to the treatments without investigating why tumours did not respond or only partially responded to the therapy.

Furthermore, the study did not explore the characteristics of resistant metastatic tumours that were initially sensitive to drug treatments. This oversight is particularly important in the context of mCRC, where a high rate of therapeutic failure is observed. Understanding the transition from a sensitive to a resistant state is crucial, as it could help identify new strategies to prevent or overcome the resistance seen in many mCRC patients, ultimately improving treatment outcomes and patient prognosis.

3.2 Aims

The aims of this chapter were to:

1. Perform genome and transcriptome sequencing (G&T-seq), specifically Smart-seq2 scRNA-seq, on cells derived from mCRC PDOs resistant to two AKT inhibitors: MK-2206 and AZD5363.
2. Analyse single-cell transcriptomes at multiple levels to identify gene expression changes characteristic of drug resistant phenotypes.
3. Validate Smart-seq2 findings using high-throughput 10x scRNA-seq data.
4. The hypothesis behind this work was that given that these mCRC PDOs were derived from a heavily pretreated donor and had prolonged exposure to AKT inhibitors, resistance mechanisms—likely involving multiple pathways—would be observable within the transcriptomes of resistant cells. Employing Smart-seq2 and supplemented by high-throughput 10x scRNA-seq data, this chapter details the transcriptional profiling of the two AKTi-resistant mCRC PDOs using a comprehensive suite of bioinformatics tools. This dual approach focused on identifying gene expression signatures indicative of drug resistance mechanisms.

3.3 Methods: Computational analysis of scRNA-seq data

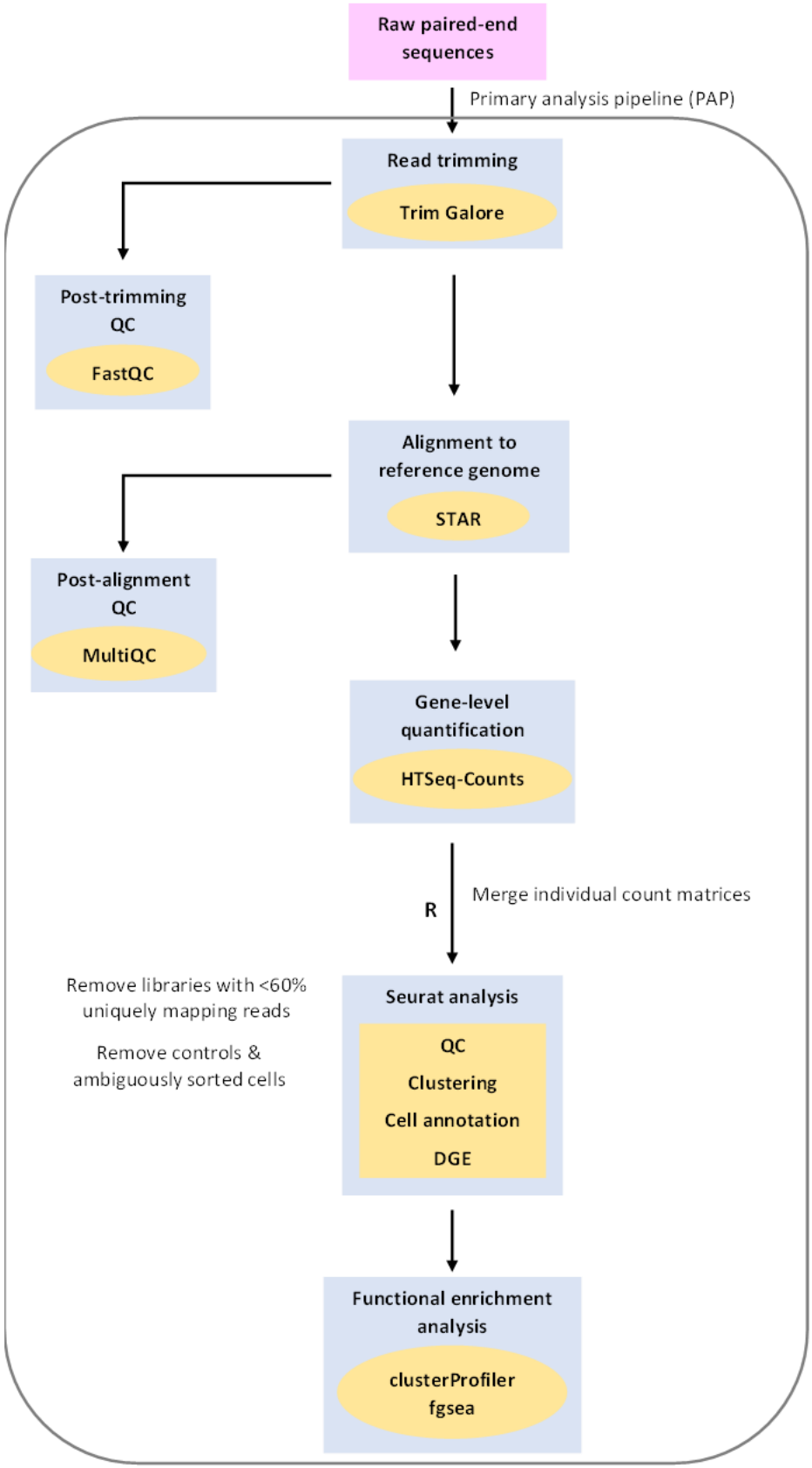


Figure 3.1. Bioinformatics workflow for the analysis of Smart-seq2 scRNA-seq data

Data processing begins with quality control (QC) of raw sequencing reads using Trim Galore and FastQC, followed by the alignment of the processed reads to the hg38 human reference genome

with STAR. Gene-level quantification is performed using HTSeq-Counts, generating per-cell count matrices that are later merged into a large matrix, serving as the primary dataset for scRNA-seq analysis using several R tools. Seurat identifies marker genes, classifies cells, and performs differential gene expression analyses at multiple levels, while functional annotation of genes is carried out using clusterProfiler and fgsea.

Table 9. List of software packages employed for the analysis of scRNA-seq data

Software	Access/citation
ClusterProfiler v4.8.3	(155)
Clustree v0.5.0	(156)
ComplexHeatmap v2.16.0	(157)
FastQC v0.11.9	(158)
Fgsea v1.26.0	(159, 160)
ggplot v2.3.4.4	(161)
HTSeq-count v0.6.1	(162)
Msigdbr v7.5.1	(163-165)
MultiQC v1.7	(166)
org.Hs.eg.db v3.17.0	(167)
Python v3.10.3	(168)
Qualimap v2.2.1	(169, 170)
R v4.1.2	R Core Team (2022)
RStudio v2023.6.2.561	R Core Team (2022)
Scillus v0.5.0	(171)
Seurat v4.3.0	(172-174)
STAR v2.6.0a	(175)
STAR v2.7.10a	(175)
tidyverse v2.0.0	(176)
Trim Galore v0.5.0	(177, 178)

3.3.1 Primary analysis of Smart-seq2 scRNA-seq data

As part of their sequencing service, the Genomics Pipeline (GP) computational team at the Earlham Institute performed an initial inspection of the Smart-seq2 scRNA-seq data from the Parental, MK1-, and AZD1-resistant mCRC patient-derived organoids (PDOs) using their primary analysis pipeline (PAP). This pipeline incorporates several bioinformatics tools to evaluate the quality of demultiplexed libraries, including FastQC (158), FastQ Screen (179) and Centrifuge (180). The outputs generated by these tools were parsed by MultiQC (166) and compiled into a comprehensive report summarising the quality of individual libraries.

A. Smart-seq2 scRNA-seq data processing: read trimming, alignment, and gene-level quantification

All computational steps described in this section were performed on the Norwich Bioscience Institute (NBI) High-Performance Computing (HPC) cluster under the SLURM workload management system version 23.02.7, along with R v4.1.2 and Python v3.10.3. The software packages and tools used to process and analyse scRNA-seq data are detailed in Table 9. Default parameters were employed for all computational tools unless stated otherwise in the text.

As previously described in the text, for each of the three mCRC organoids—Parental, MK1-resistant, and AZD1-resistant—two RNA-sequencing runs were performed through the Smart-seq2 arm of the G&T-seq protocol, with one 96-well plate sequenced per PDO in each run. Therefore, a total of 192 scRNA-seq libraries were sequenced for each experimental condition.

For each sequencing run, data processing of raw paired-end reads began by trimming off Nextera adapters and low-quality terminal bases using Trim Galore v0.5.0. The quality of trimmed reads was later assessed with FastQC v0.11.9 and MultiQC v1.7.

Before mapping reads with the STAR v2.6.0a aligner, an indexed reference genome was created using the human GRCh38 (hg38) assembly and gene transfer format (GTF) files from Ensembl release 104 (181). Trimmed reads were then aligned to the reference genome by running STAR in the 2-pass mapping mode (*twopassMode = Basic*), imposing a restriction that would allow only a maximum of one locus, i.e., one alignment per read (*outFilterMultimapNmax = 1*).

Next, the number of alignments mapped to genomic features was quantified using the Ensembl GRCh38.104 GTF annotation with HTSeq-Counts v0.6.1. The per-library count matrices generated by HTSeq-Counts were merged in R to create a large cell-by-gene count matrix consisting of 288 cells for each sequencing run (96 cells for each PDO). A custom Perl script,

courtesy of Dr Wilfried Haerty (Earlham Institute), was then employed to replace Ensembl gene IDs with corresponding gene names.

STAR log files and HTSeq-Count matrices were parsed using MultiQC v1.7 to assess the mapping rate and feature assignment performance of the sequenced libraries. This produced a comprehensive report that summarised the performance metrics for each library. The raw data used by MultiQC to generate summary plots (named "multiqc_general_stats.txt") was further processed in R for library quality visualisation using ggplot v2.3.4.4.

Additionally, Qualimap v2.2.1 *rnaseq* was used to assess the distribution of reads (coverage) across different regions of transcripts, employing the Ensembl GRCh38.104 GTF annotation file. Default settings were used for the analysis with the exception of the output format, which was specifically set to HTML (*outformat HTML*). The output files generated with this command were subsequently parsed using MultiQC v1.7 to produce a combined plot of coverage profiles across transcript lengths for all libraries derived from the Parental, MK1-resistant, and AZD1-resistant PDOs, respectively. The "multiqc_general_stats.txt" output, which provided the 5'-3' bias ratio for each library, was processed in R for visualisation using ggplot2 v3.4.4.

After exploring quality control metrics, the following criteria were set to filter out low-quality libraries for the resulting cell-by-gene count matrix:

i) Libraries with fewer than 1,000 reads as they were insufficient for accurate feature assignment.
ii) Libraries where the percentage of uniquely mapped reads was below the set threshold of 60%.
iii) Wells identified by the cell-sorting instrument as either empty or containing multiple cells, rather than a single cell.
iv) Positive and negative control libraries.

3.3.2 Seurat analysis of Smart-seq2 scRNA-seq data

A. Quality assessment of single-cell transcriptomes and batch correction

Seurat v4.3.0 served as the primary tool for analysing scRNA-seq data. The cell-by-gene count matrices and corresponding metadata files from the first and second sequencing runs, consisting of Parental, MK1-resistant, and AZD1-resistant cells (288 cells per run), were used to create two Seurat objects, namely "run1" and "run2". For "run1", an initial quality control (QC) step was conducted where only cells with more than 200 genes—as recommended in the "Seurat's guided clustering" tutorial (182)—and those with mitochondrial reads constituting less than 30% were retained. Additionally, genes were only included if they were expressed in at least 3 cells. Similar filtering parameters were applied to "run2", except the cutoff for mitochondrial read content was adjusted to 20%. These adjustments in mitochondrial read thresholds were implemented after analysing data trends, which indicated that these cutoffs were the most suitable for ensuring data quality and integrity in both Seurat objects.

The steps outlined in the "Seurat's introduction to scRNA-seq integration" tutorial (183) were followed to integrate the two Seurat objects, i.e., to filter out batch effects while still retaining and aligning the core biological signals shared between the two datasets. For this, the two objects were first combined into a list. Within this list, the objects were subjected to a series of steps required prior to integration, such as normalisation and variable feature detection.

During normalisation, raw counts were converted to a fraction of the cell's total expression, multiplied by a scale factor of 10,000, and naturally log-transformed (log1p). The top 3,000 most variable features were then identified by executing the *FindVariableFeatures* function with the variance-stabilising transformation method specified (*selection.method = "vst"*). These features typically comprise genes that capture the majority of biological and technical variability within each dataset.

Next, *SelectIntegrationFeatures* was employed to select consistently variable genes across the datasets, ensuring that subsequent analyses would focus on informative genes across both datasets. This step was followed by feature-level scaling and principal component analysis (PCA) of each dataset to reduce dimensionality while retaining the most significant components of variation within each dataset.

Reciprocal Principal Component Analysis (RPCA) was employed in the next step of the workflow, utilising a reciprocal procedure where each dataset was projected onto the reduced PCA space of the other (174, 184). This method was executed by using *FindIntegrationAnchors*

with *reduction = "rpca"* and *k.anchor = 20*. In this way, cells from "run2" were projected into the PCA space of "run1" to identify the nearest neighbours within this shared PCA space. This procedure was reciprocated by projecting cells from "run1" into the PCA space of "run2", and again identifying nearest neighbours. The mutual nearest neighbours identified in these reciprocal projections, referred to as "anchors," facilitated the alignment of similar cells across the two datasets based on their principal component scores (184). These 20 nearest neighbours formed the integration anchors, effectively identifying pairs of cells from the two datasets that likely represented the same biological state, e.g., the same cell type.

In the final step of the workflow, *IntegrateData* was applied to these anchors, resulting in a single, batch-corrected, integrated object. In the integrated analysis, technical artefacts, including differences in mitochondrial read content and total gene counts, were corrected using *ScaleData* with the parameters *vars.to.regress = c("percent.mt", "nFeature_RNA")*. Although cell cycle scores were computed on the RNA assay to determine the most likely cell cycle phase of cells, these were not regressed out to preserve potential biologically relevant findings.

Even after integration, scRNA-seq datasets remain highly dimensional, encompassing expression data for thousands of genes across numerous cells. Principal Component Analysis (PCA) was next employed to reduce the dimensionality of the dataset. PCA is a statistical method that captures linear relationships in the data by transforming a set of observations—genes, in this case—into a new set of correlated variables known as principal components (PCs) (185). After conducting PCA, an elbow plot showing variance versus principal components (PCs) was created using Seurat's *ElbowPlot* function to determine the optimal number of PCs to select for further analyses. The number of PCs chosen were selected looking at the "elbow" of the plot created, i.e., the point where the plot began to level off. Beyond this point in the plot, adding more PCs no longer results in a significant increase in the variance explained by the model.

Consequently, these PCs were selected to find clusters across resolutions ranging from 0 to 1.5 using the shared nearest neighbour (SNN) method (186) and the Louvain algorithm (186) by using the *FindNeighbors* and *FindClusters* functions. In this approach, cells are considered "neighbours" if they exhibit similar gene expression patterns, as determined by a measure of similarity and diversity, such as the Jaccard index (187). The presence of shared neighbours between pairs of cells indicates closely related transcriptional activities. Clustree v0.5.0 was subsequently employed to examine the impact of clustering at various resolutions using *clustree(integrated.seurat.object, prefix = "res.")*.

Lastly, after identifying the most appropriate clustering resolution, the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction technique was applied to visualise the clusters detected in mCRC organoids. Unlike PCA, which may struggle to capture non-linear and intricate relationships in the data (e.g., the transition from a stem cell to a mature cell within the haematopoietic system involves multiple branching points that culminate in diverse cellular fates (188)), UMAP excels as a non-linear dimensionality reduction method. This makes UMAP suitable for visualising groups of cells with similar gene expression patterns in the reduced, two-dimensional space while separating globally different clusters (189).

B. Differential gene expression analysis at multiple levels

Seurat provides various functions for differential gene expression (DGE) testing, each tailored to specific situations or goals. The DGE analyses described in this section were performed using the original, normalised RNA counts instead of the integrated data. Normalisation is crucial for ensuring that the data used in DGE analysis accurately reflects the biological variability between the groups being compared without the confounding influence of technical factors like sequencing depth (173, 174). This approach ensures that the true biological differences are accurately detected, avoids potential artefacts from data integration, maintains consistency with conventional (bulk) RNA-seq analysis, and is suitable for the statistical methods designed for count data, such as the Wilcoxon rank-sum test.

Although the integrated data is intended for dimensionality reduction and clustering, its application for DGE could lead to misleading interpretations about gene expression differences. This potential for confusion arises because the original data is adjusted (or integrated) to allow similar cells from different sequencing runs, sources or technologies to group together (173, 174). This process effectively brings their expression values to the same scale, which can obscure real expression differences. Therefore, using the normalised RNA counts is more appropriate for DGE analysis, as it avoids these complications and accurately reflects the original gene expression dynamics.

I. Cluster-level differential gene expression testing and functional annotation analysis

The *FindAllMarkers* function was applied to the normalised RNA counts to detect differences in gene expression between the previously identified clusters. The Wilcoxon rank-sum test was employed for this purpose. Genes with an average log2 fold change greater than 0.25 and detected in a minimum of 10% of cells in either cluster being compared were considered. To ensure a comprehensive analysis, the *min.diff.pct* parameter was set to infinity, ensuring the inclusion of all genes, irrespective of the magnitude of expression difference between the clusters. Moreover, setting *only.pos* = *FALSE*, ensured the inclusion of up and downregulated genes in the results. *p*-value adjustment was performed using the Bonferroni correction method based on the total number of genes in the dataset.

After the analysis, several plotting functions were employed to visualise the expression of cluster-specific genes. These included Seurat's *DotPlot* function and *plot_heatmap* from Scillus v0.5.0, which generated an annotated heatmap for these cluster markers.

For gene set enrichment analysis (GSEA), the differentially expressed genes (DEGs) within each cluster were first filtered based on statistical significance (adjusted *p*-value < 0.05) and ranked based on their average log2 fold change (avg_log2FC). This approach positioned the most upregulated genes at the top and the most downregulated at the bottom of the ranked list, preserving the magnitude and directionality of the gene expression changes. Subsequently, gene symbols were converted to their corresponding Entrez IDs using the *mapIds* function from the genome-wide annotation for the Human database, provided by org.Hs.eg.db v3.17.0, with the following parameters: *keys = genes*, *column = "ENTREZID"*, *keytype = "SYMBOL*", *multiVals = "first"*.

For GSEA analysis, the *fgsea* function from Fast Gene Set Enrichment Analysis (fgsea) v1.26.0 was employed with default parameters. This analysis focused on the Curated (C2), Ontology (C5), and Hallmark (H) gene sets from the Molecular Signatures Database (msigdbr) v7.5.1 package. Once results for each cluster were compiled into a single data frame, a filtering criterion was applied to retain pathways with adjusted *p*-values < 0.05. ComplexHeatmap v2.16.0 was then employed to visualise significant pathways for each cluster based on the Normalised Enrichment Score (NES), a metric computed by the GSEA algorithm.

II. Global differential gene expression testing and over-representation analysis

To explore gene expression differences between mCRC PDOs rather than between cell clusters, the active identity of the Seurat object was first set to the "PDO" column in the metadata. Subsequently, *FindMarkers* was employed to compare the expression of each AKTi-resistant PDO with the untreated Parental control, employing the Wilcoxon rank-sum test. The remaining parameters were consistent with those used previously for identifying cluster markers (see "Cluster-level differential gene expression testing and functional annotation analysis"). For visualising differentially expressed genes, ggplot v2.3.4.4 was used to generate a volcano plot.

The functional analysis was extended from a per-cluster approach to a PDO-based comparison to include the list of statistically significant differentially expressed genes (adjusted p-value < 0.05) identified in mCRC PDOs. For this analysis, gene symbols from the list of differentially expressed genes were converted into their corresponding Entrez ID equivalents using the biological Id Translator (*bitr*) function of clusterProfiler v4.8.3 with default parameters set. In this over-representation analysis (ORA), genes were not ranked based on avg_log2FC; instead, the analysis focused on evaluating the enrichment of specific terms and gene sets within the list of differentially expressed genes. For this purpose, the *enricher* function of clusterProfiler v4.8.3 was employed, specifically targeting gene sets from the "H", "C2", "C5", and "C6" categories as provided by the msigdbr v7.5.1 package.

C. Cell-type annotation using the Human Gut Cell Atlas

The Human Gut Cell Atlas (HGCA) (190) was chosen as a reference for annotating cells derived from mCRC PDO. The complete scRNA-seq gut cell atlas, consisting of over 400,000 cells, can be accessed from the GCA's site (191). Before being employed as a reference, the raw HGCA dataset was converted into a Seurat object and processed according to Seurat's quality control and integration workflows to create an integrated dataset. Additionally, the object was subset to only include epithelial cells in the gut. The raw HGCA data was processed by Salvatore Milite (PhD) from the Human Technopole research institute (Milan, Italy).

Cell type annotation of mCRC cells was performed using Seurat's *FindTransferAnchors* with the following parameters: *query.assay = "integrated"*, *reference.assay = "integrated"*, *normalization.method = "LogNormalize", reference.reduction = "pca", dims = 1:40*. This function identified anchors between the reference and queried datasets, which were then used to align the datasets. This alignment corrected for technical differences while preserving biological variability. For each cell in the queried dataset, a prediction score was computed for each cell type in the reference based on shared anchors and cellular expression patterns. These scores reflect the degree of similarity between a queried cell and each specific cell type in the reference. Lastly, by employing *TransferData*, queried cells were assigned the label of the highest-scoring reference cell type. In this scoring system, a high prediction score (ranging between 0 and 1) indicates a strong correlation between the expression profile of the queried cell and that of the reference cell type.

To evaluate the accuracy of the annotations, gene signatures characteristic of intestinal cell types were computed for all annotated cells, using a list of well-established gut cell markers compiled from various sources (Table 10) (190, 192-195). For each annotated cell, the signature corresponding to a specific intestinal cell type was calculated by averaging the expression of all expressed genes included in the list, employing an adaptation of the *Plot_sign* function described elsewhere (196). Lastly, *FeaturePlot* was used to visualise the distribution and expression patterns of these signatures by projecting them onto a UMAP plot.

Table 10. Marker genes for various cell types found in the gut

Cell type	Marker genes
Epithelial Signature	EPCAM
Intestinal Epithelial	CDX2, VIL1
Mesenchymal	VIM, THY1
Immune	PTPRC
Cancer Stem Cells	CD133, CD24, CD44, ABCG2, ALDH1, ALDH1A1, CD166, LGR5, CD66c, DCLK1, NES, BMI1
CLDN10 Positive	DLK1, PDX1, RBPJ, SOX9, CPA1, CLDN10
Paneth Cells	DEFA5, DEFA6, REG3A, REG1A, OLFM4, LYZ
Proliferating Cells	LGR5, BMI1, ASCL2, SMOC2, RGMB, OLFM4, SLC12A2
TA Cells	MKI67, TOP2A, PCNA, SOX9, OLFM4, PROM1, MSI1, EPHB2
Enterocytes Cells	RBP2, ANPEP, FABP2, CD36
Enterocyte Progenitors	PPP1R14D, MVP, MAD2L1, MELK, CCNA2, UBE2C, PLK1, GPSM2, QSOX1, TUBA1A
Best4 Enterocytes Cells	BEST4, OTOP2, CA7
Enterochromaffin Cells	TPH1, NPW, TAC1, CHGB
Goblet Cells	CLCA1, SPDEF, FCGBP, ZG16, MUC2
Distal Progenitors	CKB, AKAP7, GPC3
Proximal Progenitors	FGG, BEX5
Tuft Cells	TMEM45B, CHI3L1, DAPP1, SERPINI1, SLC26A2, CDKN1A, MALAT1, KRT23, CHDH, EEF2K, ENC1, EPHA4

3.3.3 Computational analysis of 10x Genomics scRNA-seq data

As previously stated in the "Generation of mCRC PDO lines resistant to AKT Inhibition" section in Chapter 2, the Parental, MK1-resistant, and AZD1-resistant PDOs subjected to either G&T-seq or 10x Genomics scRNA-seq came from different, albeit very close, passages, as these experiments were performed over a two-year period. If the experimental protocols for expanding the resistant PDO lines for either high-throughput single-cell sequencing or G&T-seq were followed, then similar subpopulations would have emerged in these organoids. Thus, any differences observed between the Smart-seq2 and 10x scRNA-seq datasets are likely due technological differences, rather than actual differences in gene expression resulting from the expansion of organoids across multiple passages.

A. 10x scRNA-seq data processing using Cell Ranger

The 10x Genomics scRNA-seq data processing for the Parental, MK1-, and AZD1-resistant PDOs was carried out by Lucrezia Patruno (University of Milano-Bicocca, Milan, Italy) using the Cell Ranger analysis pipeline (197). This process involved several steps: (1) demultiplexing sequencing reads in Illumina's raw base call (BCL) format into FASTQ files, (2) read alignment to the GRCh38 reference genome, (3) barcode processing and filtering, and (4) gene quantification. The main output files generated by Cell Ranger, which served as input for Seurat analysis, included a gene-by-cell UMI count matrix for filtered cells, a list of filtered cell barcodes and a gene annotation file. These files were organised into separate folders for each organoid and uploaded to the Accelerator project's shared drive.

3.3.4 Seurat analysis of mCRC PDOs (10x scRNA-seq)

A. Quality control and data integration

The Seurat analysis of 10x Genomics scRNA-seq data was performed following steps similar to those employed in the Smart-seq2 data analysis (see "Quality assessment of single-cell transcriptomes and batch "), with a few modifications.

First, the *Read10X* function was employed to import the gene-by-cell count matrix generated by the Cell Ranger pipeline. This matrix served as the basis for creating Seurat objects, with *min.cells = 3* applied to filter out genes expressed in fewer than three cells and *min.features = 200*—as recommended in the "Seurat's guided clustering" tutorial (182)—used to exclude cells expressing fewer than two hundred genes. This process was performed for each PDO dataset, resulting in three Seurat objects.

Subsequently, the quality of each Seurat object was independently inspected to retain cells based on specific criteria. The following cutoffs, determined to be the most appropriate for maintaining data quality and integrity in the three Seurat objects, were established after visualising the data trends (Supplementary Figure 4 and Supplementary Figure 5). For the Parental organoid, filters were applied to retain cells expressing fewer than 7,500 genes, where each gene was expressed in at least 3 cells, and ensuring that mitochondrial genes constituted less than 20% of the total gene expression. In the MK1-resistant organoid, cells expressing fewer than 8,500 genes were retained, with the remaining filtering parameters matching those employed to filter the Parental dataset. For the AZD1-resistant organoid, cells with fewer than 8,500 genes were retained, where each gene should be expressed in at least 3 cells, and mitochondrial genes should make up less than 25% of the total gene expression.

As with the Smart-seq2 data, the steps outlined in the "Seurat scRNA-seq integration" workflow (183) were followed to integrate the three 10x scRNA-seq datasets. The primary distinction in this analysis was in the dimensionality reduction phase: following PCA, the first 25 PCs were identified as capturing the majority of the variation present in the data.

B. Projection of Smart-seq2 cluster labels onto the 10x scRNA-seq dataset

To validate the Smart-seq2 clustering analysis findings with a larger dataset, cell cluster labels from the Smart-seq2 data were projected onto the integrated 10x Seurat object. This was performed by first using *FindTransferAnchors* to identify anchors between the reference (Smart-seq2) and queried (10x) Seurat objects with the following parameters: *query.assay = "integrated"*, *reference.assay = "integrated"*, *normalization.method = "LogNormalize"*, *reference.reduction = "pca"*, *dims = 1:9*. For each cell in the queried dataset, a prediction score was computed for each cell cluster in the reference based on these anchors. These scores, ranging from 0 to 1, reflected the similarity in gene expression between a query cell and a specific reference cell cluster. *TransferData* was then used to assign the highest-scoring Smart-seq2 cluster label to each cell in the 10x Seurat object based on these anchors and prediction scores. Finally, *AddMetaData* incorporated the transferred labels into the "Smart-seq2_Seurat_clusters" metadata column of the 10x object. This step effectively annotated the cells in the 10x dataset with cluster information derived from the Smart-seq2 dataset.

C. Cluster-level gene expression analysis on Smart-seq2 cluster projections

Cluster-level gene expression analysis was conducted by setting "Smart-seq2_Seurat_clusters" as the active identity in the 10x object using *SetIdent*. To facilitate a direct comparison and identification of shared differentially expressed genes between the 10x and Smart-seq2 datasets, the Wilcoxon rank-sum test was applied with identical parameters across both datasets (see "Cluster-level differential gene expression testing and functional annotation analysis"). Shared differentially expressed genes between the 10x and Smart-seq2 datasets, meeting the criteria of an adjusted p-value < 0.05, and an avg_log2FC greater than or equal to absolute 0.5 (|0.5|), were identified using Venn diagrams (198).

D. Cell-type annotation using the Human Gut Cell Atlas (10x scRNA-seq)

The parameters employed to annotate the cells in the Smart-seq2 dataset using the Human Gut Cell Atlas as a reference were also applied to label the 10x dataset (see "Cell-type annotation using the Human Gut Cell Atlas").

E. Projection of Smart-seq2 cell type labels onto the 10x scRNA-seq dataset

To assess the abundance of colonic cell types identified in the Smart-seq2 dataset labelled with the HGCA within a larger dataset, cell type labels from the Smart-seq2 data were projected onto the 10x Seurat object. This process followed the same parameters as those used for projecting Smart-seq2 clusters onto the 10x data (see "Projection of Smart-seq2 cluster labels onto the 10x scRNA-seq dataset"), with the only difference being the reference metadata column selected for annotation. In this case, the metadata column containing the cell type annotations was used instead of the one containing the cluster identities. Subsequently, the corresponding cell type labels were transferred to the "Smart-seq2_Gut_Cell_Types" metadata column of the 10x object.

F. Global differential gene expression testing

Differential gene expression analysis between AKTi-resistant and Parental mCRC organoids was conducted on the 10x dataset using the Wilcoxon rank-sum test with parameters identical to those applied to the Smart-seq2 data (see "Global differential gene expression testing and over-representation analysis"). Shared differentially expressed genes between the 10x and Smart-seq2 datasets, meeting the criteria of an adjusted p-value < 0.05, and an avg_log2FC greater than or equal to |0.5|, were identified using Venn diagrams (198).

3.4 Results

3.4.1 Primary analysis of Smart-seq2 data identifies bacterial contaminants in human cDNA libraries

To ensure scRNA-seq analyses were performed on high-quality data, the raw reads corresponding to sequencing libraries underwent extensive processing. The initial inspection of libraries focused on identifying the species represented in the sequences. Table 11 presents an excerpt from the PAP MultiQC report for the first scRNA-seq run, focusing on the Centrifuge module, which shows the top 2 most abundant species in 25 representative libraries: Index 1 for Parental, Index 2 for MK1-resistant and Index 3 for AZD1-resistant libraries. Notably, for the first (and the second) sequencing run, the PAP report revealed the presence of bacterial sequences attributed to *Variovorax* species alongside human cDNA libraries. Although cDNA sequences from the intended human target were predominant in most samples, a significant number of libraries still showed the bacterial contaminant as their foremost or second most abundant species.

Table 11. Top 2 most abundant species in mCRC libraries as extracted from the PAP report from the first set of plates sequenced by scRNA-seq.

Sample Name	1st Name	1st %	2nd Name	2nd %
SOGTseqIndex1H1	*Homo sapiens*	41.1	synthetic construct	3.0
SOGTseqIndex1H2	*Homo sapiens*	43.6	*Variovorax paradoxus*	6.3
SOGTseqIndex1H3	*Homo sapiens*	22.1	*Cyprinus carpio*	9.8
SOGTseqIndex1H4	*Homo sapiens*	87.0	synthetic construct	2.7
SOGTseqIndex1H5	*Homo sapiens*	87.4	synthetic construct	2.1
SOGTseqIndex1H6	*Homo sapiens*	85.8	synthetic construct	2.3
SOGTseqIndex1H7	*Homo sapiens*	83.3	synthetic construct	1.8
SOGTseqIndex1H8	*Homo sapiens*	82.2	synthetic construct	1.6
SOGTseqIndex1H9	*Homo sapiens*	86.6	synthetic construct	3.1
SOGTseqIndex2A1	*Homo sapiens*	83.9	*Cyprinus carpio*	2.1
SOGTseqIndex2A2	*Homo sapiens*	56.9	*Variovorax paradoxus*	13.2
SOGTseqIndex2A3	*Homo sapiens*	36.3	*Variovorax paradoxus*	19.5
SOGTseqIndex2A4	*Variovorax paradoxus*	34.7	*Variovorax boronicumulans*	7.8
SOGTseqIndex2A5	*Homo sapiens*	59.4	*Variovorax paradoxus*	10.0
SOGTseqIndex2A6	*Variovorax paradoxus*	35.5	*Variovorax boronicumulans*	7.8
SOGTseqIndex2A7	*Homo sapiens*	47.4	*Variovorax paradoxus*	15.5
SOGTseqIndex2A8	*Variovorax paradoxus*	34.2	*Variovorax boronicumulans*	7.8
SOGTseqIndex2A9	*Homo sapiens*	68.1	*Variovorax paradoxus*	8.1
SOGTseqIndex3B1	*Homo sapiens*	55.5	*Variovorax paradoxus*	12.9
SOGTseqIndex3B3	*Homo sapiens*	72.8	*Cyprinus carpio*	3.9
SOGTseqIndex3B4	*Homo sapiens*	35.5	*Variovorax paradoxus*	20.6
SOGTseqIndex3B5	*Homo sapiens*	66.6	*Variovorax paradoxus*	8.1
SOGTseqIndex3B6	*Homo sapiens*	54.3	*Variovorax paradoxus*	14.2
SOGTseqIndex3B7	*Homo sapiens*	54.5	*Variovorax paradoxus*	13.3
SOGTseqIndex3B8	*Homo sapiens*	43.4	*Variovorax paradoxus*	15.7
SOGTseqIndex3B9	*Homo sapiens*	47.8	*Variovorax paradoxus*	14.5

3.4.2 scRNA-seq quality control pipeline selects high-quality cells for downstream analyses

Following the initial inspection, single-cell libraries underwent a series of processing steps that included mapping reads to the human reference genome and subsequently assigning them to genomic features (genes).

The first sequencing run revealed significant differences in read depth among the PDO libraries (Figure 3.2A). The Parental PDO had the lowest average total reads, recording approximately 1.1 million reads per cell (±755,620 SD). The AZD1-resistant PDO followed it with 1.6 million reads per cell (±818,587 SD), and the MK1-resistant PDO had the highest average at 1.8 million reads per cell (±1,263,283 SD). In contrast, the disparity in read depth between the Parental, MK1-resistant, and AZD1-resistant PDOs was less pronounced in the second sequencing run, with averages of 1.8 million (±1,430,615 SD), 1.7 million (±1,174,742 SD), and 1.8 million (±1,240,364 SD) reads per cell, respectively. This represents a marginal difference of nearly 78,000 reads between the highest (AZD1-resistant) and lowest (Parental) averages in the second sequencing run, a stark contrast to the 704,267 read difference observed in the first run.

Given the extensive characterisation of the human genome sequence, 70-90% of RNA-seq reads are expected to map onto the reference genome (199). However, the bacterial contamination adversely affected the mapping rates across all samples in the first sequencing run. The MK1-resistant libraries were most severely affected, with an average of 39.2% of reads per cell (±34.6 SD) mapping to a single genomic location (Figure 3.2B-C, left panel). This starkly contrasted with the AZD1-resistant and Parental libraries, which had uniquely mapping read averages of 71.1% (±18.4 SD) and 87.7% (±10.0 SD), respectively. While libraries in the second sequencing run were also affected, the mapping performance was relatively uniform across all samples, with uniquely mapping read averages ranging from 64.3% (±14.6 SD) in MK1-resistant libraries to 69.3% (±12.1 SD) in Parental libraries (Figure 3.2B-C, right panel).

The low mapping percentages observed across libraries are likely due to the extensive contamination by *Variovorax* species. This bacterial contamination also impacted the efficiency of read assignment to features or genes. In the first sequencing run, the MK1-resistant organoid was notably affected, with an average of 30.3% of reads per cell (±26.7 SD) unambiguously assigned to features (Figure 3.3A, left panel). The AZD1-resistant and Parental PDOs followed with 49.8% (±14.8 SD) and 67.4% (±8.16 SD) of reads per cell assigned to features, respectively. In contrast, the second sequencing run showed a more consistent (albeit still low) feature assignment across samples, with average percentages of reads per cell assigned to features

ranging from 46.1% (±12 SD) in MK1-resistant libraries to 49.6% (±10.6 SD) in Parental libraries (Figure 3.3A, right panel). Importantly, when considering only the uniquely mapped reads from both runs, a significant fraction was correctly assigned to features in both sequencing runs (Figure 3.3B).

The analysis of read coverage distribution across the length of mapped transcripts revealed a consistent trend in scRNA-seq libraries, where the coverage increased from the 5' end towards the middle of the transcript, peaked in the middle region, and then decreased towards the 3' end (Figure 3.4A). This indicates that both the 5' and 3' ends had lower coverage compared to the rest of the transcript. This type of plot helps assess the uniformity of read coverage along the genes, which is particularly important for methods like Smart-seq2 that aim to capture full-length transcripts (143). Additionally, there was a slight preference for the 5' end over the 3' end, with mean 5'-3' bias ratios of 1.22 (±0.09 SD), 1.21 (±0.06 SD), and 1.24 (±0.15 SD) for the Parental, MK1-resistant, and AZD1-resistant libraries in the first sequencing run, and 1.27 (±0.12 SD), 1.28 (±0.21 SD), and 1.30 (±0.27 SD) for the same PDOs in the second run (Figure 3.4B). The 5' or 3' biases in RNA-seq data can occur due to library preparation protocols, primer binding efficiency, reverse transcriptase activity favouring either end, sequencing platform characteristics, and biological factors like RNA degradation and secondary structures (200, 201). These findings underscore the importance of considering regional biases in transcript coverage when interpreting RNA-seq data to ensure accurate biological conclusions.

Out of the 576 libraries sequenced (192 for each PDO) across the two sequencing runs, a total of 376 libraries (65.3%) met the scRNA-seq quality standards outlined in the Methods section (see "Smart-seq2 scRNA-seq data processing: read trimming, alignment, and gene-level quantification"). This included 143 cells (74.5%) from the Parental PDO, 93 cells (48.4%) from the MK1-resistant PDO and 140 cells (72.9%) from the AZD1-resistant PDO.

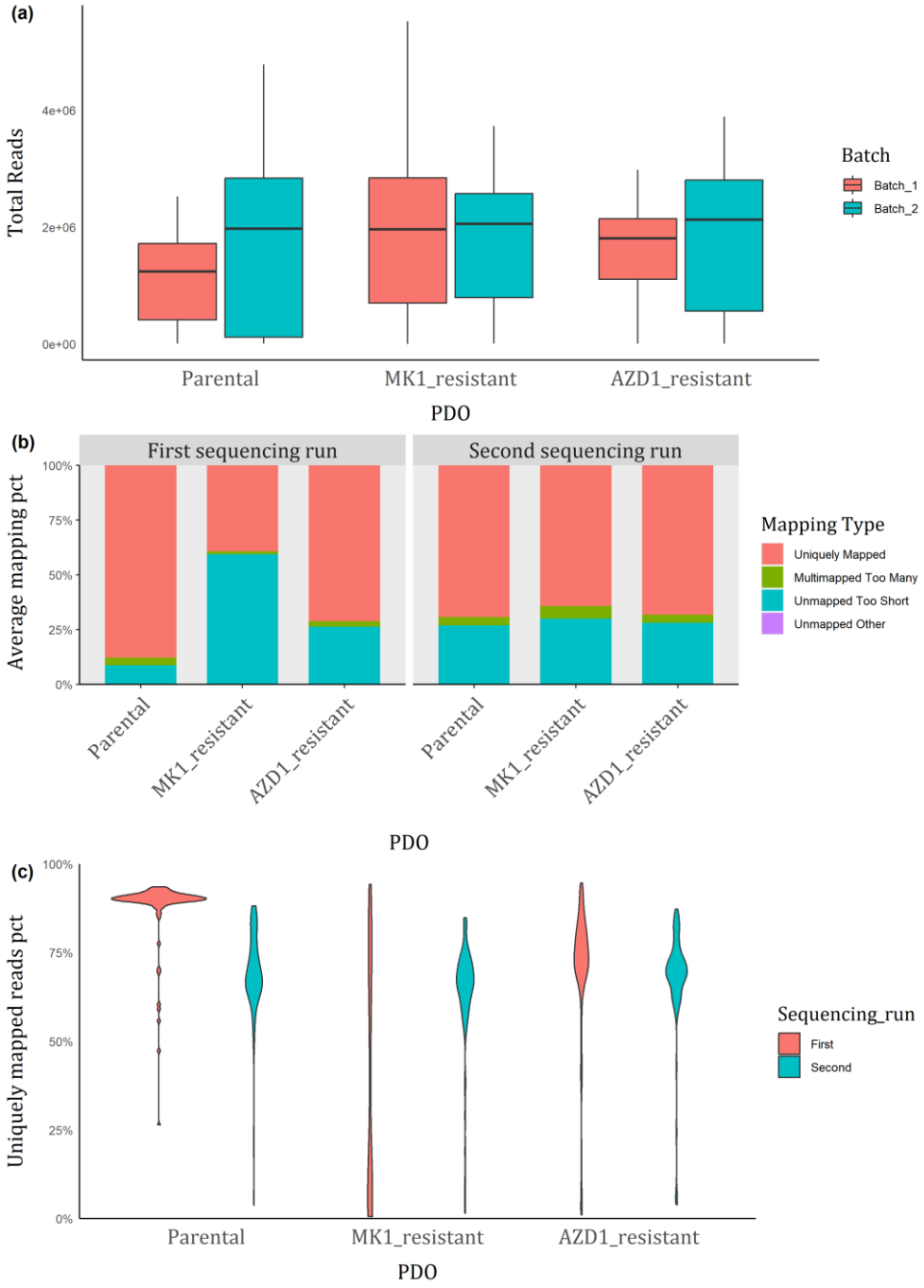


Figure 3.2. Smart-seq2 scRNA-seq mapping quality metrics of mCRC PDO libraries across two sequencing runs.

(a) *Distribution of total reads in scRNA-seq libraries derived from the Parental, MK1-resistant, and AZD1-resistant mCRC PDOs across two sequencing runs. Each box plot displays the median value (central line), the interquartile range (25th and 75th percentiles as the box boundaries), and 1.5× the interquartile range (whiskers).* ***(b)*** *Average mapping percentages across four distinct mapping categories (uniquely mapped reads, multi-mapped reads, short, unmapped reads, and other unmapped reads) for each PDO across two sequencing runs.* ***(c)*** *Distribution of uniquely mapped reads per PDO over two sequencing runs. The width of the violin plots represents the density of data points (i.e., individual libraries) at various levels of uniquely mapped reads.*

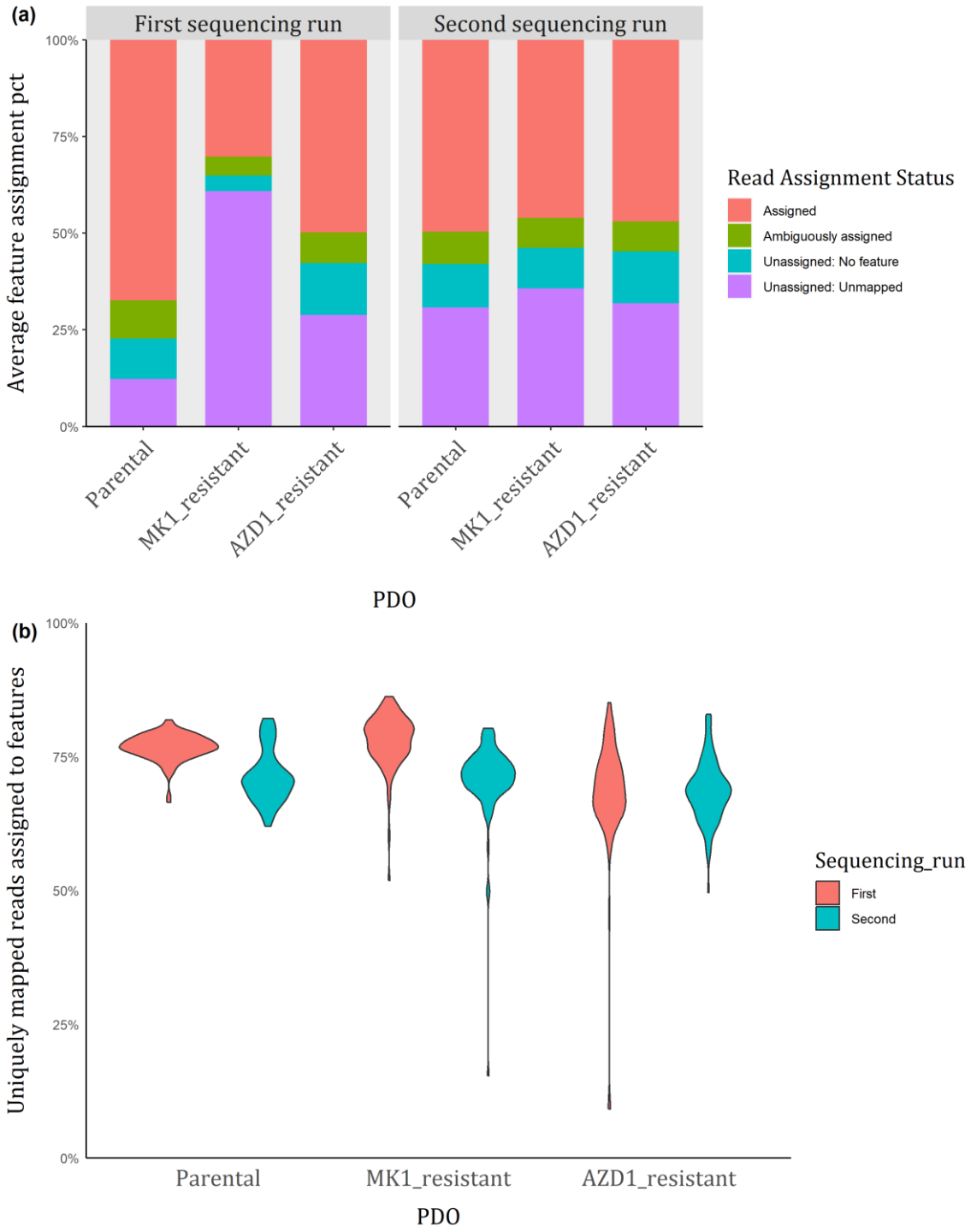


Figure 3.3. Smart-seq2 scRNA-seq feature assignment metrics of mCRC PDOs across two sequencing runs.

(a) *Average feature assignment percentages for mCRC PDOs, categorised into four groups: reads correctly assigned to genes, reads ambiguously assigned, reads unassigned to features, and unmapped reads not unassigned to genomic features.* ***(b)*** *Distribution of uniquely mapped reads unambiguously assigned to genomic features across two sequencing runs. Colour-coding distinguishes the sequencing runs, with coral representing the first and sky-blue denoting the second batch of sequenced cells.*

3.4. Results

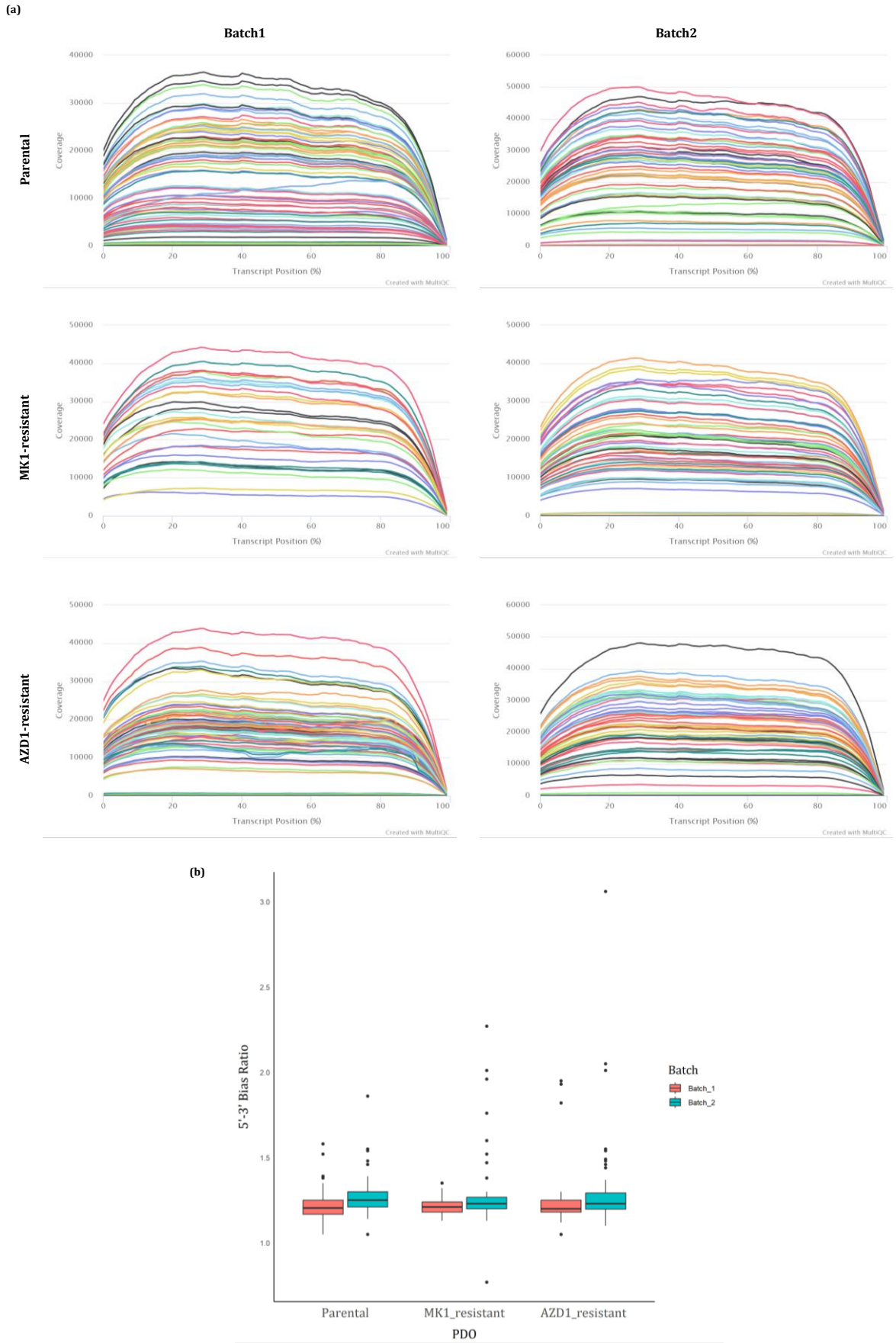


Figure 3.4. Read coverage profile across mapped transcripts for scRNA-seq libraries from mCRC PDOs across two sequencing runs.

(a) *Each plot represents the distribution of sequencing reads along the length of mapped transcripts. The x-axis indicates the position along the transcript (from 5' to 3') as a percentage, while the y-axis shows the number of reads mapping to each position. The colours in the plot represent individual single-cell RNA-seq libraries derived from Parental, MK1-resistant and AZD1-resistant mCRC PDOs.* ***(b)*** *Box plots represent the distribution of 5'-3' bias ratios across PDO libraries and sequencing runs. Each box plot displays the median value (central line), the interquartile range (25th and 75th percentiles as the box boundaries), and 1.5× the interquartile range (whiskers).*

3.4.3 Seurat analysis of single cells derived from mCRC PDOs

A. Preliminary exploration of single-cell transcriptomes selects high-quality cells for subsequent analyses

After completing the initial Smart-seq2 data processing steps, Seurat was employed as the primary tool for analysing single-cell transcriptomes. The initial stages of the Seurat workflow incorporated additional quality control measures to filter out "low-quality" cells that might not have been removed during the processing steps. Figure 3.5A-F presents Seurat quality control metrics for the Parental, MK1-resistant, and AZD1-resistant PDOs for the two sequencing runs before removing low-quality cells. The distribution of data points revealed variability within and across mCRC PDOs.

Before filtering low-quality libraries, the average number of genes captured per cell in the first plate of Parental cells sequenced was 6,504, which dropped to 5,422 genes per cell in the following run (Figure 3.5A). The MK1-resistant PDO recorded an average of 6,738 genes per cell in the first run, which decreased to 5,806 genes per cell in the subsequent run. In the case of the AZD1-resistant PDO, 7,177 genes per cell were detected in the initial sequencing, followed by 5,527 genes per cell in the next run. A similar trend was observed in the total RNA counts (Figure 3.5B). A correlation coefficient of 0.74 (Figure 3.5E) revealed a strong positive association between these two metrics, suggesting that capturing a higher number of mRNA molecules leads to the identification of more genes.

Other metrics of note include the percentage of mitochondrial- and ribosomal-related genes, both of which were the highest for the Parental PDO across the two sequencing runs (Figure 3.5C-D). Despite this, a correlation coefficient of 0.02 between mitochondrial gene percentage and RNA counts (Figure 3.5F) suggests a negligible association between the total RNA counts in cells and the proportion of those transcripts attributed to mitochondrial genes. This indicates that, in this case, the expression of non-mitochondrial genes was largely independent of mitochondrial gene activity—an important observation given that elevated mitochondrial gene content can signal cell stress or reduced cell quality.

Low-quality transcriptomes were subsequently filtered out after inspecting data trends. Figure 3.6A-F showcases Seurat quality control metrics after implementing the quality filters described in the methods section (see "Quality assessment of single-cell transcriptomes and batch correction"). From the initial pool of 376 cells loaded into Seurat, 361 satisfied the quality criteria. Breaking it down by PDO type, 134 cells (equivalent to 69.8% of the initial 192 cells sequenced) originated from the Parental PDO, 87 cells (45.3%) from the MK1-resistant PDO, and 140 cells (72.9%) from the AZD1-resistant PDO.

3.4. Results

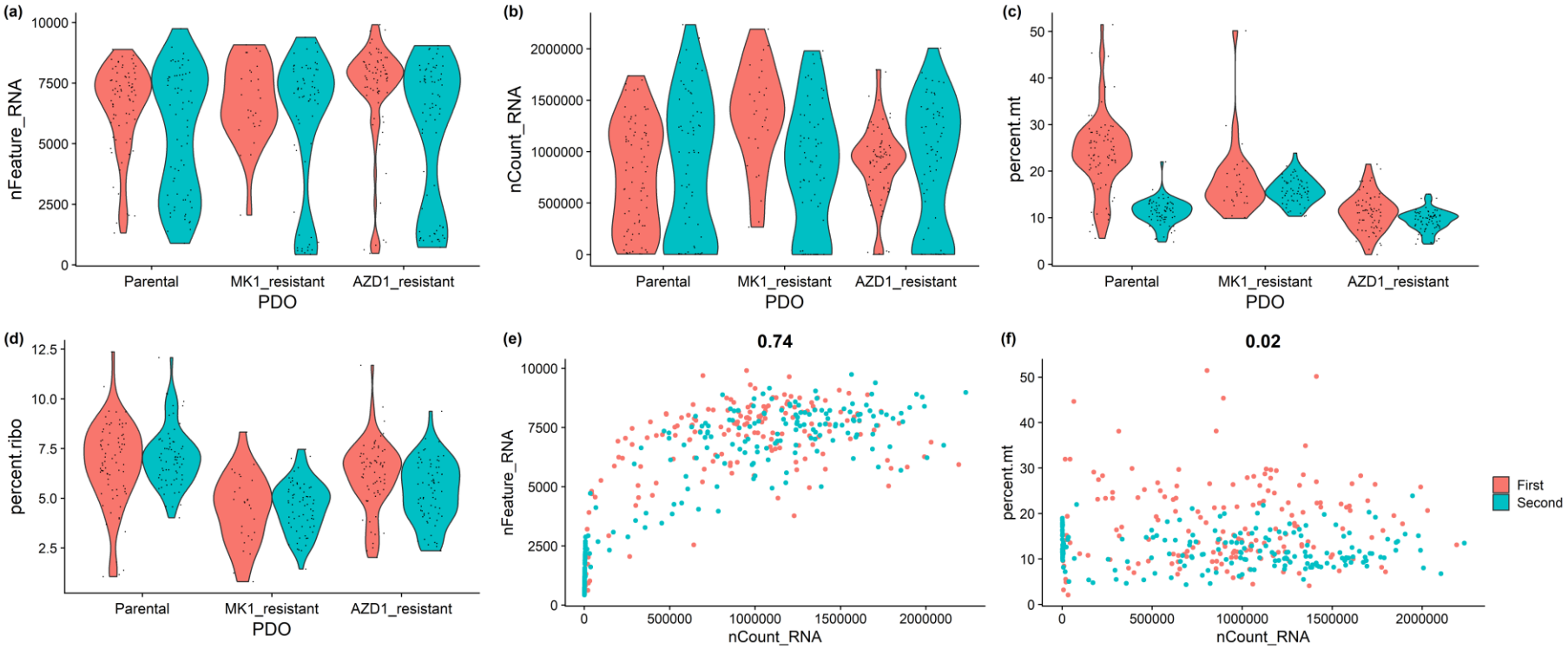


Figure 3.5. Quality metrics of 376 scRNA-seq libraries from three mCRC PDOs before excluding low-quality cells.

(a)-(f) present violin plots illustrating the distribution of single-cell transcriptomes derived from the Parental, MK1-resistant, and AZD1-resistant PDOs based on various Seurat quality control metrics before removing low-quality cells. These metrics include ***(a)*** *number of genes detected (nFeature_RNA),* ***(b)*** *number of RNA molecules (nCount_RNA),* ***(c)*** *mitochondrial RNA percentage (percent.mt), and* ***(d)*** *ribosomal RNA percentage (percent.ribo) for each cell. All metrics were evaluated across each PDO and over two sequencing runs. Scatter plots depict the relationship between RNA content versus* ***(e)*** *number of genes detected or* ***(f)*** *mitochondrial content for each cell. 376 cells in total: 143 Parental, 93 MK1-resistant, and 140 AZD1-resistant cells.*

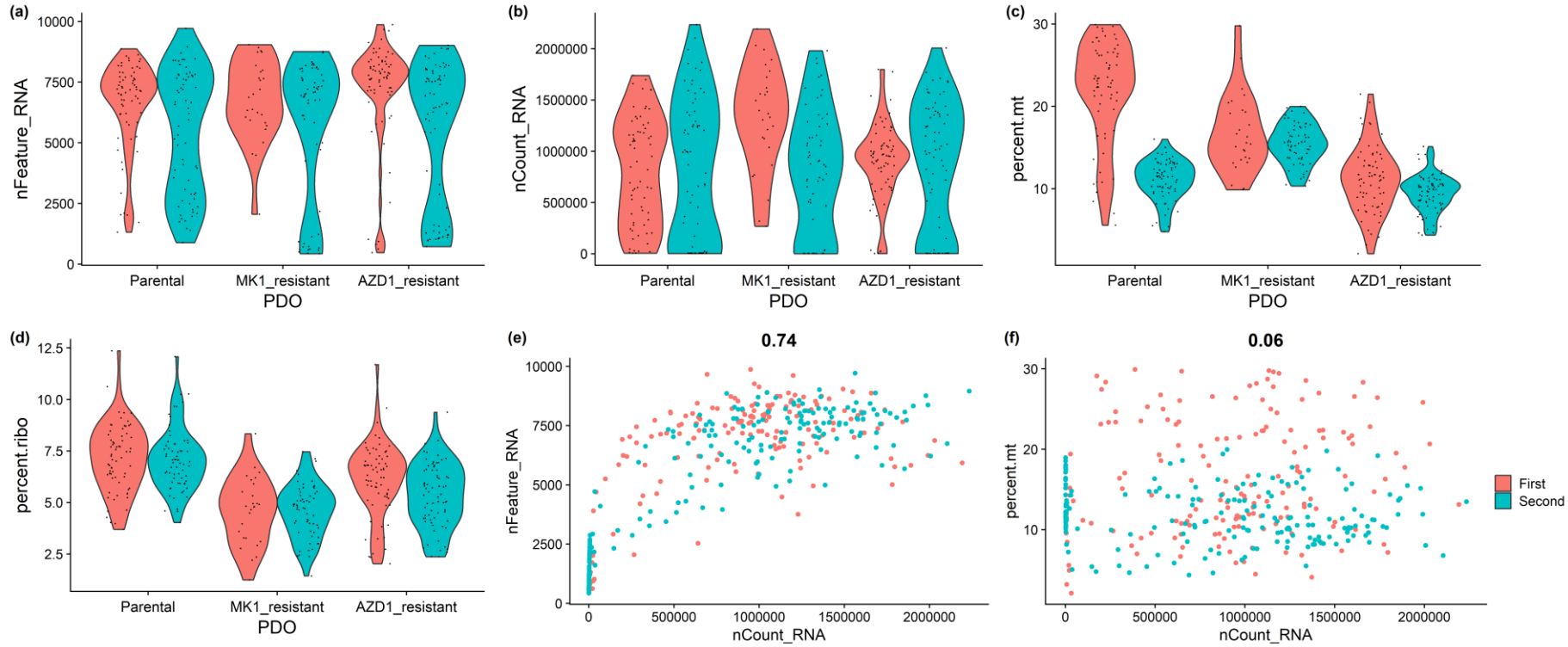


Figure 3.6. Quality metrics for 361 scRNA-seq libraries from three mCRC PDOs after excluding low-quality cells.

(a)-(f) present violin plots illustrating the distribution of single-cell transcriptomes derived from the Parental, MK1-resistant, and AZD1-resistant PDOs based on various Seurat quality control metrics after removing low-quality cells. These metrics include ***(a)*** *number of genes detected (nFeature_RNA),* ***(b)*** *number of RNA molecules (nCount_RNA),* ***(c)*** *mitochondrial RNA percentage (percent.mt), and* ***(d)*** *ribosomal RNA percentage (percent.ribo) for each cell. All metrics were evaluated across each PDO and over two sequencing runs. Scatter plots depict the relationship between RNA content versus* ***(e)*** *number of genes detected or* ***(f)*** *mitochondrial content for each cell. 361 cells in total: 134 Parental, 87 MK1-resistant, and 140 AZD1-resistant cells.*

B. Batch correction coupled with dimensionality reduction reveals transcriptionally distinct cell clusters in mCRC PDOs

The quality control analysis described in the previous section evidenced gene detection discrepancies across mCRC PDOs and sequencing batches. To address this shortcoming, the Seurat integration workflow was employed to integrate the two sequencing batches, leveraging its capability to align scRNA-seq datasets from multiple sources and sequencing technologies using the Reciprocal Principal Component Analysis (RPCA) based approach.

Seurat's RPCA-based integration can discern true biological signals, which should be consistent across datasets, from batch-specific noise (174). This is achieved by identifying anchors, i.e., pairs of cells conserved across different sources and representing the same cell type or state. These anchors act as reference points for alignment. Mutual neighbourhood constraints are subsequently applied on the anchors to ensure reference points remain close to similar points in other datasets, thus producing an integrated dataset. Figure 3.7A illustrates the top 3,000 genes exhibiting the highest variability across cells in each dataset. These features, including shared ones such as *MT4*, *DKK4*, *S100A3*, *C6orf15*, *EDN3*, and *RBP1*, were selected for integrating the two scRNA-seq datasets.

Continuing with the topic of variability, cell cycle effects can be a significant source. To address this, the Seurat's cell cycle scoring method was adopted, which relies on the expression of canonical cell cycle-associated genes to categorise cells into G1, S, or G2/M cell cycle phases (202), without regressing out or eliminating this signal. Adopting this approach preserved the inherent cellular diversity in the metastatic CRC PDOs, which typically comprise undifferentiated stem cells and differentiated cell types (203).

Principal Component Analysis (PCA) was next employed to reduce the dimensionality of the dataset. In PCA visualisations, such as the one depicted in Figure 3.7B, each gene is represented with specific loadings in the PC space. These loadings or weight coefficients indicate how much each gene contributes to a particular PC. Genes with higher loadings (in magnitude, irrespective of its positive or negative sign) on a specific PC contribute more to the variance captured by that PC. Moreover, the genes that make up the loadings of a PC can offer insights into underlying cellular functions or states that are being highlighted by that PC. In this case, PC1 consists of genes associated with extracellular matrix components (e.g., *ECM1*) (204), lipid metabolism (*CD36*, *MGLL*) (205, 206), cell adhesion (*TACSTD2*, *FNDC3A*) (207, 208) and cell signalling (*TGFA*, *EDN3*, *KIT*, *WNT5A*) (209-212). Conversely, PC2 encompass genes regulating the G2/M cell cycle phases (*CDK1*, *UBE2C*, *BIRC5*, *TOP2A*, *MKI67*) (202), suggesting this PC might capture variation related to the cell cycle phase of cells.

3.4. Results

PCs aim to capture the maximum variance in the data; however, not all PCs are equally informative. The primary PCs often represent genes conveying relevant biological information, such as in the examples above. In contrast, later PCs may capture noise, which can hinder downstream analysis. The scree plot in Figure 3.7C, commonly called the “elbow plot”, shows the variance (presented as standard deviation) captured by the first 20 PCs. This visualisation assisted in identifying the optimal number of PCs for further analyses. In this case, the first 9 PCs captured the most variation, as indicated by the plot’s “elbow”, which represents the point where adding more PCs would have a minimal increase in captured variance.

Finally, cells were clustered at a resolution of 0.5 after examining the effects of various resolutions on the number of clusters generated. As depicted in Figure 3.8, this resolution provided a balance between achieving clear separation among transcriptionally distinct cell clusters and avoiding over-partitioning of the data. As a result, four distinct cell clusters were identified in mCRC PDOs. Notably, Cluster 2 remained consistent, not dividing with increasing resolution, suggesting a unique expression profile compared to the other clusters.

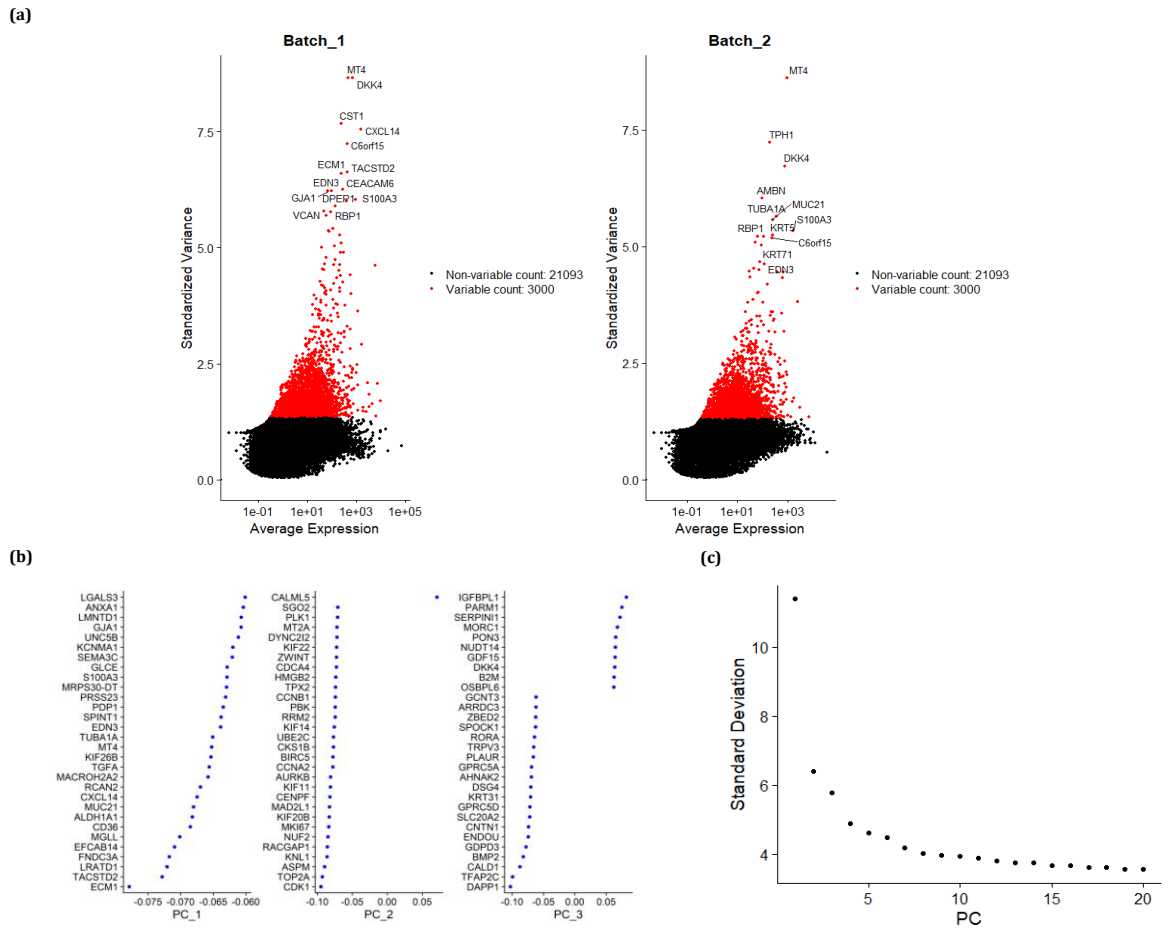


Figure 3.7. Seurat integration aligns two scRNA-seq datasets using variably expressed anchor genes, followed by principal component analysis to identify the main axes of variation for further analyses.

(a) *Scatter plots reflect the standard variance of genes against average expression, highlighting the top 3,000 genes (red) with the most significant variability in gene expression in the first (left) and second (right) sequencing runs. Top 15 genes in each batch are annotated.* ***(b)*** *Representation of genes in the first three principal component spaces in the integrated dataset. The y-axis displays genes, while the x-axis represents the contribution (or weight) of each gene to the principal component.* ***(c)*** *Scree plot shows the variation captured by the first 20 principal components. This visualisation aids in selecting the optimal number of principal components for cell clustering, with PCs 1-9 chosen based on the plot's "elbow", where the line begins to flatten.*

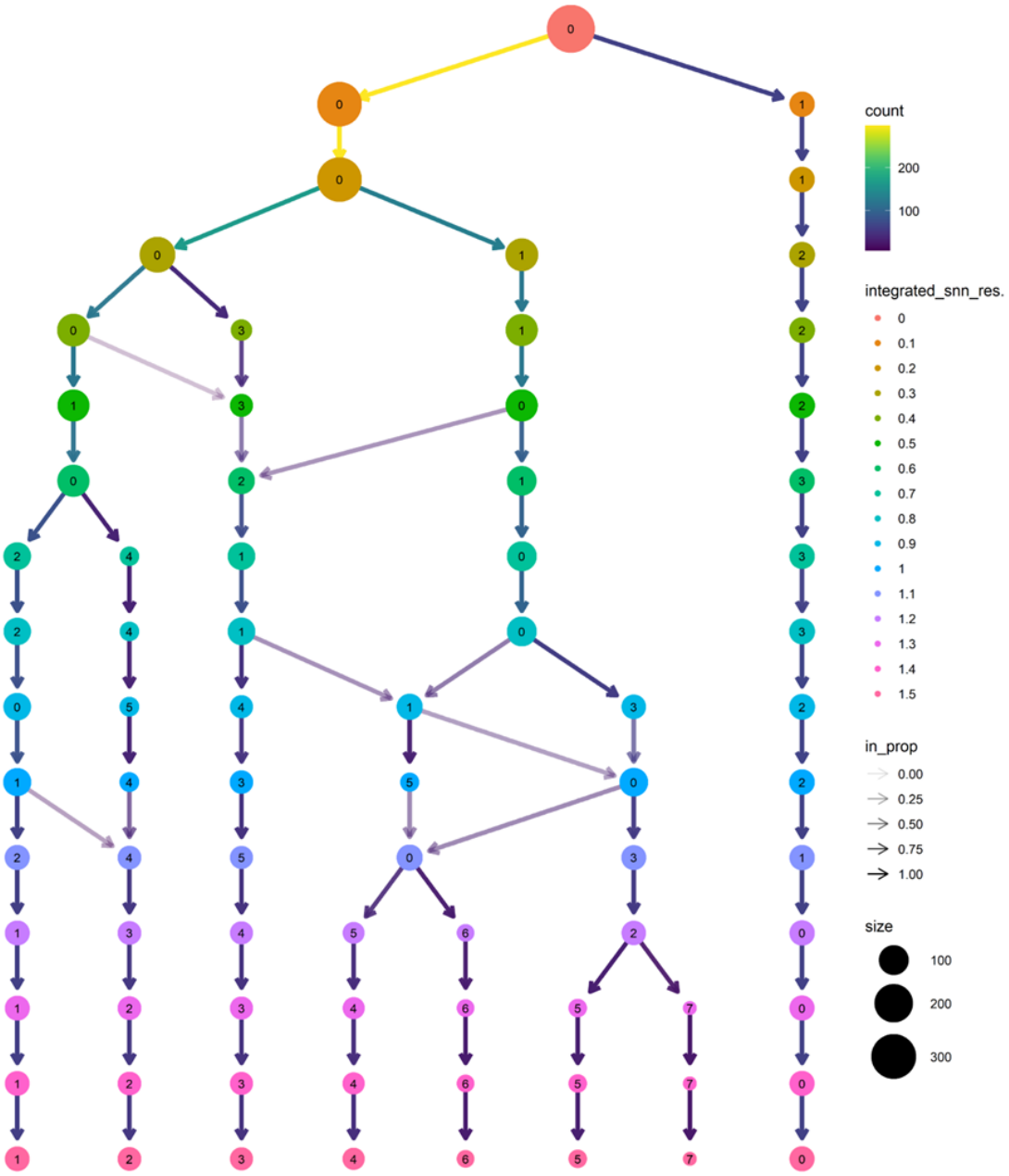


Figure 3.8. Clustree visualisation demonstrating the effect of increasing resolution on the stability of cell clustering in scRNA-seq data.

Clustree depicts how clusters of cells evolve as the resolution parameter is adjusted. Each node represents a cluster, with the node colour transitioning from red to pink as the resolution parameter increases. Nodes are sized in proportion to the number of cells they contain. Arrows represent cluster splits with increasing resolution. The "in_prop" legend measures cluster stability across resolutions. A value closer to 1 suggests high stability, where the majority of cells from a lower-resolution cluster remain together at a higher resolution. Conversely, lower values indicate the fragmentation of a cluster into smaller subsets with increasing resolution, thus making the cluster composition less stable.

C. Differential cluster abundance analysis across mCRC cells

In the UMAP plots presented in Figure 3.9A-D, each data point represents a single cell colour-coded based on gene expression similarity (Figure 3.9A), PDO of origin (Figure 3.9B), sequencing batch (Figure 3.9C) and predicted cell cycle phase (Figure 3.9D). In addition, Figure 3.9E indicates the number of cells in each cluster across the three mCRC PDOs, while Figure 3.9F shows the distribution of cell cycle phases in these samples. Several observations can be drawn from these visualisations:

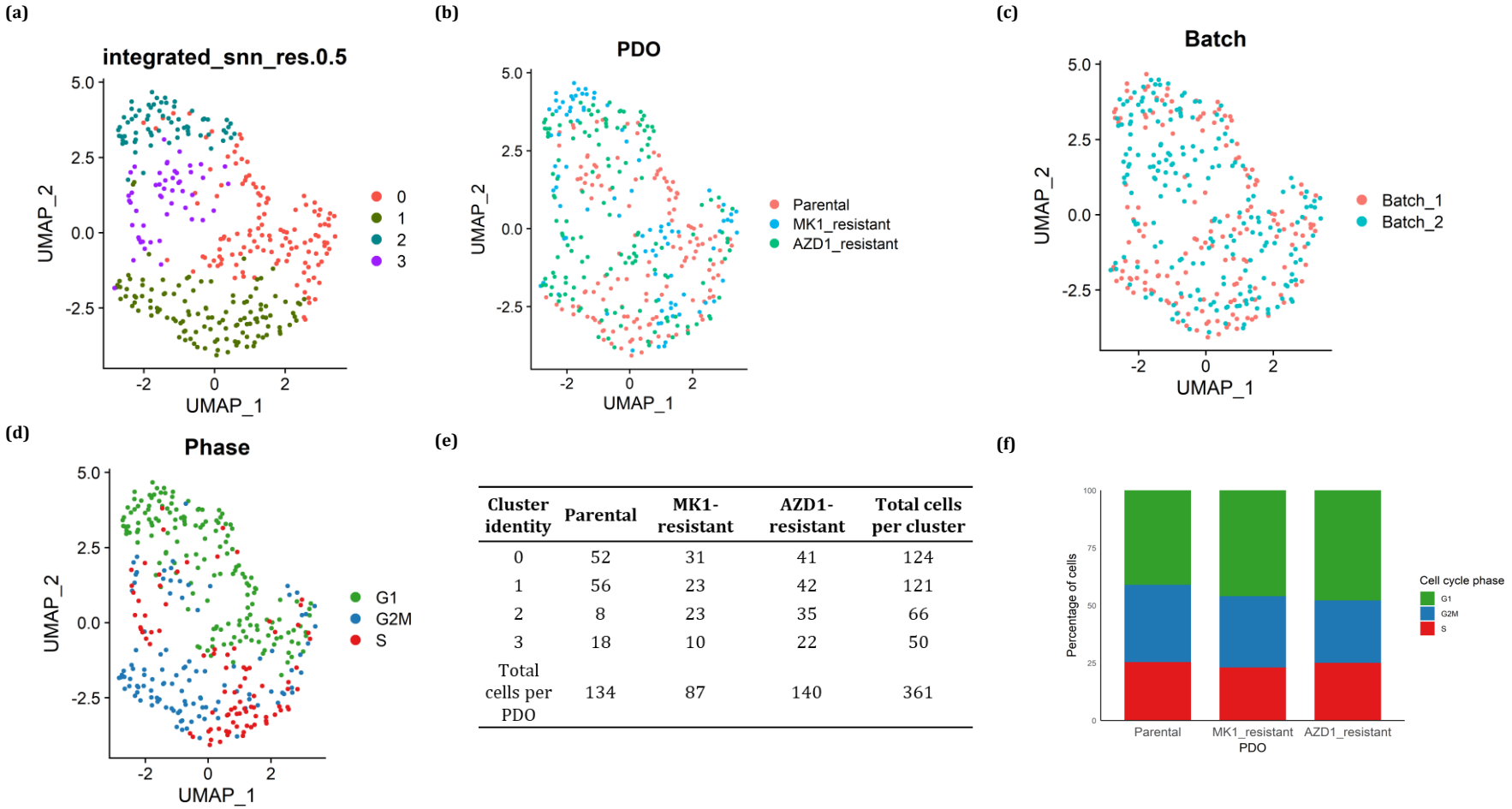
Looking at the cell cycle distribution of cells, it is apparent that some cell cycle phases were more prevalent in certain clusters (Figure 3.9 and Supplementary Figure 1). For example, most cells in Cluster 0 and Cluster 2 were found in the G1 phase, with a few cells in the S and G2/M phases. On the other hand, Cluster 1 consisted of a mix of cells in S and G2/M phases, while all cell cycle phases were equally represented in Cluster 3. Therefore, while there was some cell cycle-driven bias in the clustering of mCRC, cells did not exclusively separate by their cell cycle phase, meaning that biological differences beyond cell cycle stage also influenced cell clustering. Nevertheless, there was a consistent representation of all cell cycle phases across the PDOs (Figure 3.8F).

Secondly, Clusters 0 through 3 were consistently observed across all mCRC PDOs, though their frequencies varied. Notably, Clusters 0 and 1 were the most prevalent, with 124 and 121 cells, respectively. Although Cluster 3 consisted of 50 cells from all mCRC PDOs, most of these cells (92%) originated from the second batch of PDO plates sequenced, suggesting a potential batch effect affecting this cluster.

Comparing cell counts per cluster between the Parental control and AKTi-resistant PDOs revealed differential impacts across clusters. In Cluster 0, the number of cells decreased in resistant conditions, with 31 cells observed in the MK1-resistant PDO and 41 cells in the AZD1-resistant, compared to 52 cells in the Parental PDO. A similar reduction was observed for Cluster 1, particularly in the MK1-resistant PDO, where cell counts dropped to 23 from the 56 cells observed in the Parental PDO. In contrast, Cluster 2 was the only cluster that experienced a significant increase in the resistant PDOs, with cell counts rising to 35 cells in the AZD1-resistant PDO and 23 cells in the MK1-resistant, up from only 8 cells in the control PDO. Changes in Cluster 3 were minimal, with a slight reduction observed from 18 cells in the control to 10 cells in the MK1-resistant PDO.

It is important to acknowledge the potential uncertainty in these observations. As previously mentioned, contamination by *Variovorax* species significantly compromised the data from the

MK1-resistant PDO, resulting in a much-reduced number of cells available for analysis compared to the other organoids. In contrast, the Parental and AZD1-resistant PDOs had equal starting cell counts, providing a solid foundation for comparative analysis (Figure 3.9E).



Cluster identity	Parental	MK1-resistant	AZD1-resistant	Total cells per cluster
0	52	31	41	124
1	56	23	42	121
2	8	23	35	66
3	18	10	22	50
Total cells per PDO	134	87	140	361

Figure 3.9. scRNA-seq identifies four transcriptionally distinct cell clusters in mCRC PDOs.

UMAP transforms high-dimensional scRNA-seq data into a two-dimensional space, where each data point denotes a cell. Cells are coloured by ***(a)*** *gene expression similarity,* ***(b)*** *PDO type,* ***(c)*** *sequencing batch and* ***(d)*** *cell cycle phase.* ***(e)*** *Detailed cell counts per Seurat cluster for each mCRC PDO.* ***(f)*** *Distribution of cell cycle phases in each PDO. Number of cells = 361: 134 Parental, 87 MK1-resistant and 140 AZD1-resistant cells.*

D. Cross-platform analysis of cluster abundances reveals differential effects of AKT Inhibition on mCRC cells

The cost associated with Smart-seq2 can be prohibitive for large-scale studies. Consequently, the subset of cells sequenced—in this case, 192 cells per PDO, sorted based on cell viability—might not accurately represent the diverse cell types within the original mCRC organoid cultures. Therefore, it is challenging to ascertain whether the observed differences in cell cluster abundances between the control and AKTi-resistant PDOs reflect an actual biological difference or are merely the result of the sampling limitations of Smart-seq2, which could introduce random variability. This is where the single-cell RNA-seq techniques developed by 10x Genomics come into play (213).

To address the limitations of Smart-seq2 and validate its results, the Smart-seq2 cluster annotations were mapped or projected onto the 10x Genomics scRNA-seq datasets generated from the Parental (2,245 cells in total after filtering low-quality cells), MK1-resistant (2,708 cells) and AZD1-resistant (2,289 cells) organoids. This method involved comparing each cell in the queried dataset (i.e., 10x genomics scRNA-seq data) with cells in the reference dataset (i.e., Smart-seq2 scRNA-seq data) and assigning cluster labels based on prediction scores that indicated the most likely cluster identity for each queried cell.

The Smart-seq2 clusters projected onto the 10x scRNA-seq data (Figure 3.10A) exhibited high spatial coherence, as evidenced by the proximity of cells from the same clusters in the UMAP plot without significant overlaps (Figure 3.10B). The high prediction scores observed for all clusters, ranging between 0.5-1 (Figure 3.10C), confirmed a close resemblance between the gene expression profiles of the 10x Genomics and Smart-seq2 scRNA-seq datasets. Such precision in label transfer underscores the robustness of the approach in preserving the distinct molecular signatures of each cluster during the annotation process and provides a solid basis for reliably investigating cluster abundances across the PDOs using the 10x dataset. However, similar to the Smart-seq2 dataset, variations due to cell cycle phases were not adjusted for in the Seurat analysis of the 10x data. This might explain the clear separation of clusters based on cell cycle phase, a distinction that was less pronounced in the Smart-seq2 data (Figure 3.9A).

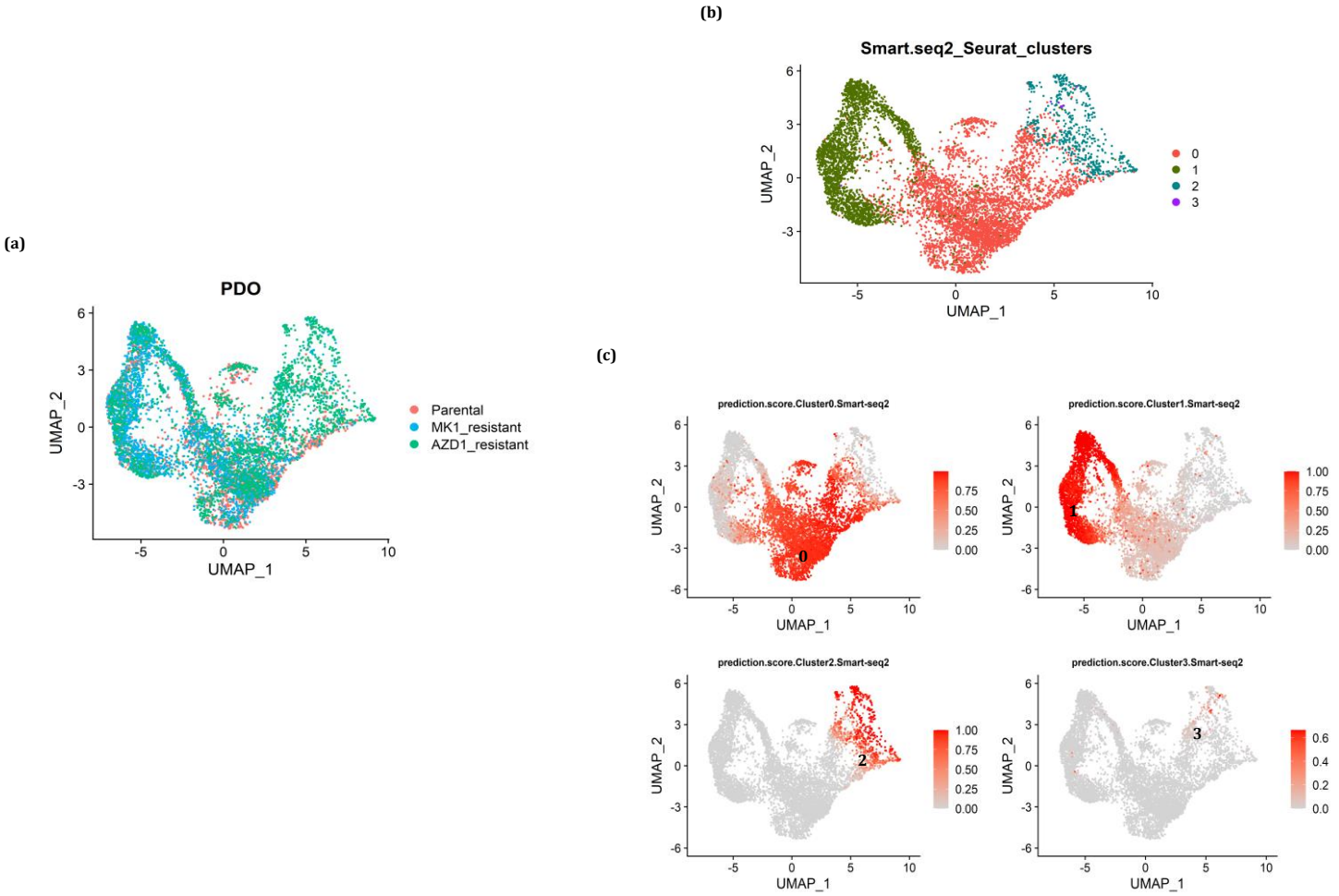


Figure 3.10. Smart-seq2 cluster projection onto 10x scRNA-seq data allows large-scale evaluation of cluster abundances in mCRC PDOs.

UMAP plots display ***(a)*** *integrated 10x scRNA-seq dataset for three mCRC PDOs,* ***(b)*** *Smart-seq2 clusters projected onto the 10x scRNA-seq data with* ***(c)*** *corresponding prediction scores indicating gene expression similarities between the datasets. Number of cells = 7,242: 2,245 Parental, 2,708 MK1-resistant and 2,289 AZD1-resistant cells.*

Similar observations to those made with Smart-seq2 were noted when examining the cell cycle distribution across the clusters projected onto the 10x data. For example, Cluster 0 and Cluster 2 predominantly consisted of cells in the G1 phase, while Cluster 1 displayed an equal distribution of cells in the S and G2/M phases (Figure 3.11A).

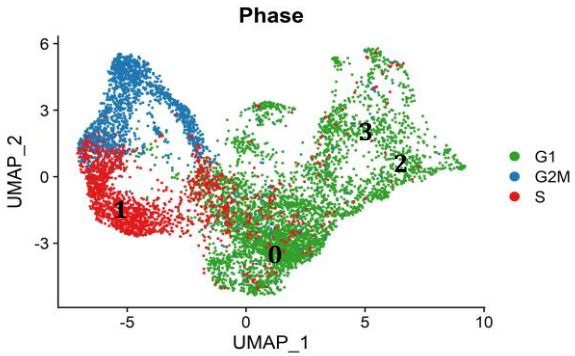
When examining the global distribution of cell cycle phases across PDOs, the Smart-seq2 data showed a uniform distribution (Figure 3.9F). In contrast, the 10x dataset revealed a distinct pattern: the MK1-resistant organoid exhibited the highest number of cells in the S and G2/M phases among all PDOs, with the G1 phase being less prevalent (Figure 3.11B). This significantly differs from the AZD1-resistant and Parental PDOs, where the G1 phase was the most common, followed by the S and G2/M phases. It is important to highlight that the cell cycle distributions observed in the 10x dataset are intrinsic to this dataset and not influenced by the Smart-seq2 cluster projections.

Figure 3.11C illustrates the variation in cluster abundances across the mCRC PDOs. The Parental and AZD1-resistant PDOs had comparable cell counts, with 2,245 and 2,289 cells each. In contrast, the MK1-resistant PDO had approximately 500 more cells than the other two (2,708 cells). Except for Cluster 3, which was unique to the MK1-resistant PDO, all clusters were represented in the three organoids, albeit at varying frequencies. In both the Parental and AZD1-resistant PDOs, Cluster 0 and Cluster 1 were the two most predominant clusters. For the MK1-resistant organoid, Cluster 1 was the most abundant, followed by Cluster 0.

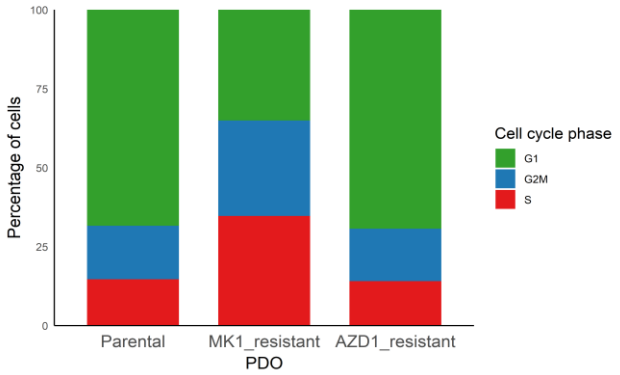
Upon examination of the cluster abundance across PDOs, Cluster 0 showed a slight decrease in the AKTi-resistant PDOs compared to the Parental control. This trend aligns with the Smart-seq2 observations, which also showed a minimal decrease in Cluster 0 in AKTi-resistant PDOs (Figure 3.9E). In contrast, Cluster 1 saw a more than a twofold increase in the MK1-resistant PDO relative to the control, whereas the same cluster showed a minor decrease in the AZD1-resistant PDO. Despite the MK1-resistant PDO having the highest cell count, which might explain the rise in Cluster 1, one would typically anticipate such an increase to be spread across all clusters. Conversely, Cluster 2 was over five times more abundant in the AZD1-resistant sample compared to the control, while it remained stable in the MK1-resistant PDO. These findings are in agreement with the Smart-seq2 data, which also noted a significant increase in Cluster 2 within the AZD1-resistant sample (Figure 3.9E).

In summary, employing a larger and potentially more representative dataset like the 10x scRNA-seq enabled a more accurate assessment of cell cluster abundances across mCRCs PDOs and a more reliable understanding of the cellular composition within the organoids.

(a)



(b)



(c)

Smart-seq2 cluster	Parental	MK1-resistant	AZD1-resistant	Total cells per cluster
0	1555	1214	1316	4085
1	620	1408	591	2619
2	70	79	382	531
3	0	7	0	7
Total cells per PDO	2245	2708	2289	7242

Figure 3.11. Distribution of Smart-seq2 cluster abundances and cell cycle phases in mCRC PDOs as derived from 10x scRNA-seq data.

(a) *Cell cycle phase distribution across projected clusters. (b) Distribution of cell cycle phases in each organoid. (c) Cell counts in each Smart-seq2-projected cluster on the 10x scRNA-seq data derived from three mCRC PDOs. Number of cells = 7,242: 2,245 Parental, 2,708 MK1-resistant and 2,289 AZD1-resistant cells.*

E. Cluster-level differential gene expression analysis in mCRC PDO cells

Differential gene expression (DGE) analysis was first conducted on the Smart-seq2 dataset to investigate the transcriptional characteristics of the four cell clusters identified. This analysis produced gene expression profiles depicted in Figure 3.12A, with the top markers for each cluster illustrated in Figure 3.12B. Cluster-level DGE was also performed on the 10x dataset containing the Smart-seq2 cluster projections to validate the Smart-seq2 findings. This approach identified shared differentially expressed genes (DEGs) across both datasets that were statistically significant (having adjusted *p*-values < 0.05) and exceeded the average log2 fold change (log2FC) threshold of |0.5| (Figure 3.13A-C).

Following the DGE analysis, gene set enrichment analysis (GSEA) was performed to investigate the biological implications of the transcriptional profiles identified, focusing on the Curated (C2), Ontology (C5), and Hallmark (H) categories from the Molecular Signatures Database (MSigDB) (163). Gene sets with significant enrichment (adjusted *p*-value $\leq$ 0.05) are depicted in Figure 3.14 based on their Normalised Enrichment Score (NES). The NES, which is adjusted to account for the size of gene sets, facilitates comparisons across gene sets of different sizes, and its magnitude represents the degree of enrichment. A positive NES indicates that the gene set is predominantly upregulated, whereas a negative NES suggests downregulation.

Given the central role of AKT in cellular proliferation and survival pathways (214), and considering this study's focus on the cellular response to AKT inhibitors, the expression of *AKT1*, *AKT2*, and *AKT3* was also examined (Figure 3.15). These genes encode the AKT isoforms targeted by the MK-2206 and AZD5363 inhibitors. This analysis aimed to provide insights into the expression patterns of *AKT* genes across the identified clusters.

Cluster 0 featured an increased expression of genes from both the large (*RPL10A*, *RPL18A, RPL31*, *RPL36*, *RPL37*, *RPL37A*) and small (*RPS14*, *RPS18, RPS19*, *RPS24*, *RPS27*) ribosomal subunits (215). These ribosomal protein (RP) family genes are components of ribosomes, which are essential for mRNA translation into proteins (216). The surge in ribosomal-associated gene activity in Cluster 0 suggests an increase in ribosomal biogenesis and protein synthesis. Furthermore, enrichment analysis revealed an involvement in RNA metabolic processes, indicated by gene sets "C2: REACTOME_METABOLISM_OF_RNA", "C5: GOMF_RNA_BINDING", and "C5: GOBP_RNA_PROCESSING" gene sets (Figure 3.14).

As previously shown in Figure 3.9D, the majority of cells in Cluster 0 were in the G1 phase. Consistent with this observation, Cluster 0 showed the lowest expression of all *AKT* isoforms in the Smart-seq2 dataset (Figure 3.15). Furthermore, there was a downregulation of genes

associated with cell cycle progression (*CDCA4*, *CDK1*, *MKI67) (217, 218)*, spindle-microtubule organisation (*ASPM*, *TPX2*) (219, 220), DNA topology (*TOP2A*) (221), DNA synthesis and repair (*RRM2*) (222), and chromatin modulation (*ASF1B*) (223) (Figure 3.12A-B). Downregulation of these genes was also observed in the 10x dataset (Figure 3.13A).

In addition, a significant downregulation was observed in genes associated with cell signalling (*TGFA*) (209), and in genes related to cell adhesion and interactions with the extracellular matrix, such as *CD36*, *ECM1*, *FNDC3A* and *ITGA1* (204, 205, 207, 224), with *TACSTD2* identified as the most downregulated gene in Cluster 0 (average log2FC = -1.898592483, adjusted *p*-value = 8.95×10^{-8}).

On the other hand, Cluster 1 exhibited elevated levels of S phase genes (*ATAD2*, *GMNN*, *MCM4*, *MCM7*, *PCNA*, *RRM1*, *RRM2*) and G2/M phase genes (*BIRC5*, *CCNB2*, *CDK1*, *CENPF*, *CKS2*, *MKI67*, *TMPO*, *TOP2A*, *UBE2C*) (202). The majority of these genes were also found in the 10x DGE analysis (Figure 3.13B). GSEA linked these genes to cell cycle regulatory pathways, including "C2: FISCHER_G2_M_CELL_CYCLE", "C5: GOBP_MITOTIC_CELL_CYCLE", "H: HALLMARK_G2M_CHECKPOINT", and "H: HALLMARK_E2F_TARGETS". The leading-edge subset of genes within the "H: HALLMARK_E2F_TARGETS" gene set, which contributes to the NES, revealed an upregulation of E2F transcription factor targets. These genes are involved in a range of cellular functions such as DNA replication (*DUT*, *MCM3*, *RFC3*, *TK1*), DNA repair (*BARD1*, *CHEK1*, *PRKDC*, *RAD51AP1*), G2/M checkpoints (*CENPE*, *CHEK1*, *MAD2L1*), chromatin remodelling (*CBX5*), and mitotic regulation (*PLK1*) (225). Aligned with these findings, Cluster 1 exhibited the highest expression levels of all *AKT* isoforms in the Smart-seq2 and 10x datasets. (Figure 3.15).

Cluster 1 also showed enrichment for several cancer-related gene sets such as "C2: GRADE_COLON_AND_RECTAL_CANCER_UP", "C2: VECCHI_GASTRIC_CANCER_EARLY_UP" and "C2: SHED-DEN_LUNG_CANCER_POOR_SURVIVAL_A6". The leading-edge subset for these gene sets featured genes such as *CDK1*, *CENPF*, *RRM2*, and *TOP2A,* which relate directly to cell cycle regulation. Other upregulated genes include *EIF4EBP1*, a negative regulator of mRNA translation (226)*, EZH2*, which is involved in chromatin modification and gene silencing (227), and *MELK*, playing roles in post-translational modifications, signal transduction, the cell cycle, and proliferation (228).

In terms of downregulated genes, Cluster 1 exhibited downregulation in genes associated with extracellular matrix remodelling, such as *ECM1* (204), and genes involved in cell adhesion, including *TACSTD2* and *FNDC3A* (207, 208). This finding aligns with gene sets related to cell-cell interaction, displaying negative NES scores, including "C5: GOBP_BIOLOGICAL_ADHESION"

and "C5: GOBP_CELL_CELL_ADHESION". Additionally, the gene set "C5: GOBP_EPITHELIAL_CELL_DIFFERENTIATION" was observed to be downregulated in this cluster, suggesting that Cluster 1 likely consists of rapidly proliferating but undifferentiated cells.

Turning to Cluster 2, as noted earlier, it was the only cluster that remained unsplit with increasing resolution (Figure 3.8). This observation indicates that Cluster 2 has a unique expression profile that sets it apart from the other clusters. Although Cluster 2 and Cluster 0 primarily consisted of cells in the G1 phase, Cluster 2 exhibited a pronounced downregulation in translation processes (Figure 3.14), setting it apart from Cluster 0. Additionally, G1-phase cells in Cluster 2 showed low expression levels of proliferation markers (e.g., *CDK1*, *CENPF*, *RRM2*)(202), a trend accompanied by the downregulation of cell cycle-related gene sets including "H: HALLMARK_G2M_CHECKPOINT" and "H: HALLMARK_E2F_TARGETS".

Additionally, the transcriptional profile of Cluster 2 was marked by the upregulation of genes involved in cell-cell communication (*GJA1*), cell signalling (*JAG2*, *NRP2*, *SEMA3C*, *SORBS1*), cell survival and metabolism (*FABP5*, *KCNMA1*, *PDK3*, *RCAN2*, *SHH*), detoxification (*MT4*, *PON2*, *PON3*, *QSOX1*), migration and tissue remodelling (*CEMIP*, *ECM1*, *LGALS1*), and pathways of energy and nutrient metabolism (*CD36*, *GLCE*, *SLC7A8*) (229). GSEA identified these DEGs in gene sets related to epithelial cell migration, such as "C2: WU_CELL_MIGRATION" and various "C5" gene sets, including "GOBP_CELL_CELL_ADHESION" and "GOBP_BIOLOGICAL_ADHESION".

Lastly, Cluster 3 mirrored the expression patterns observed in Cluster 1 and Cluster 2 (Figure 3.12A). The clustree plot illustrated in Figure 3.8 shows that Cluster 1 and Cluster 3 diverged from a common "ancestral" cluster at the 0.3 resolution, likely explaining their resemblance in gene expression profiles. The DGE analysis for Cluster 3 identified 24 statistically significant DEGs for this cluster (adjusted p-value < 0.05). Of these, only *TACSTD2* and *CFTR* demonstrated deregulation exceeding the average log2FC threshold of |0.5|, with values of 0.525 and -0.540, respectively. The limited number of genes might explain why the GSEA for this cluster yielded no results. Furthermore, there were no shared DEGs between the Smart-seq2 and 10x datasets for this cluster (Figure 3.13C).

The cluster-level differential gene expression analysis revealed diverse cellular activities across the four clusters identified in mCRC PDOs. By identifying unique expression patterns in critical areas such as cell cycle regulation, cell signalling, and cell adhesion, this analysis hints at potential mechanisms through which cells respond to AKT inhibition.

3.4. Results

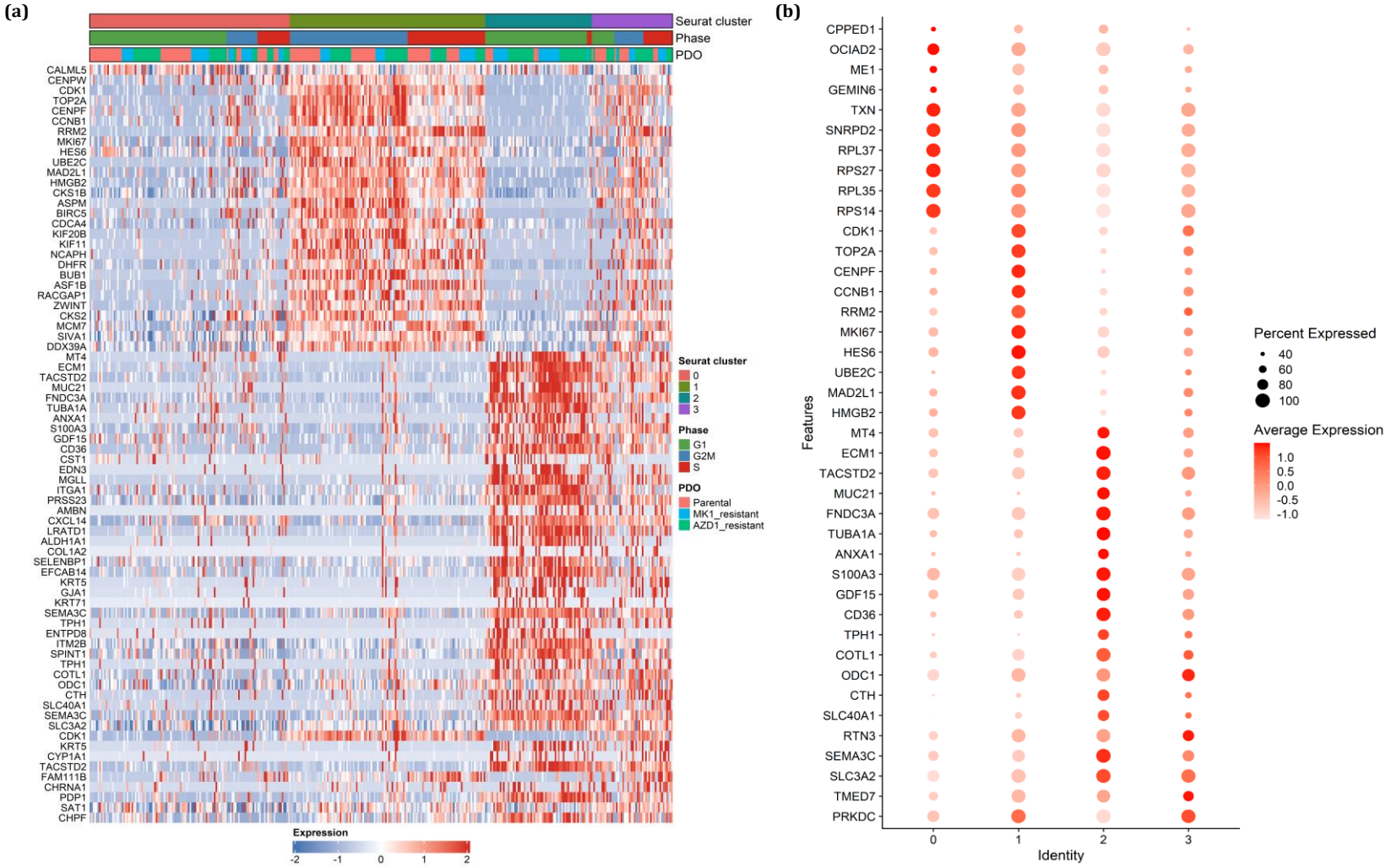


Figure 3.12. Cluster-level differential gene expression analysis in mCRC PDOs.

(a) *Heatmap of differentially expressed genes characterising clusters identified in mCRC PDOs, with gene expression levels transitioning from blue (low expression) to red (high expression). Annotations above the heatmap indicate the cell cycle stage and PDO type.* ***(b)*** *Dot plot of top 10 cluster markers.*

3.4. Results

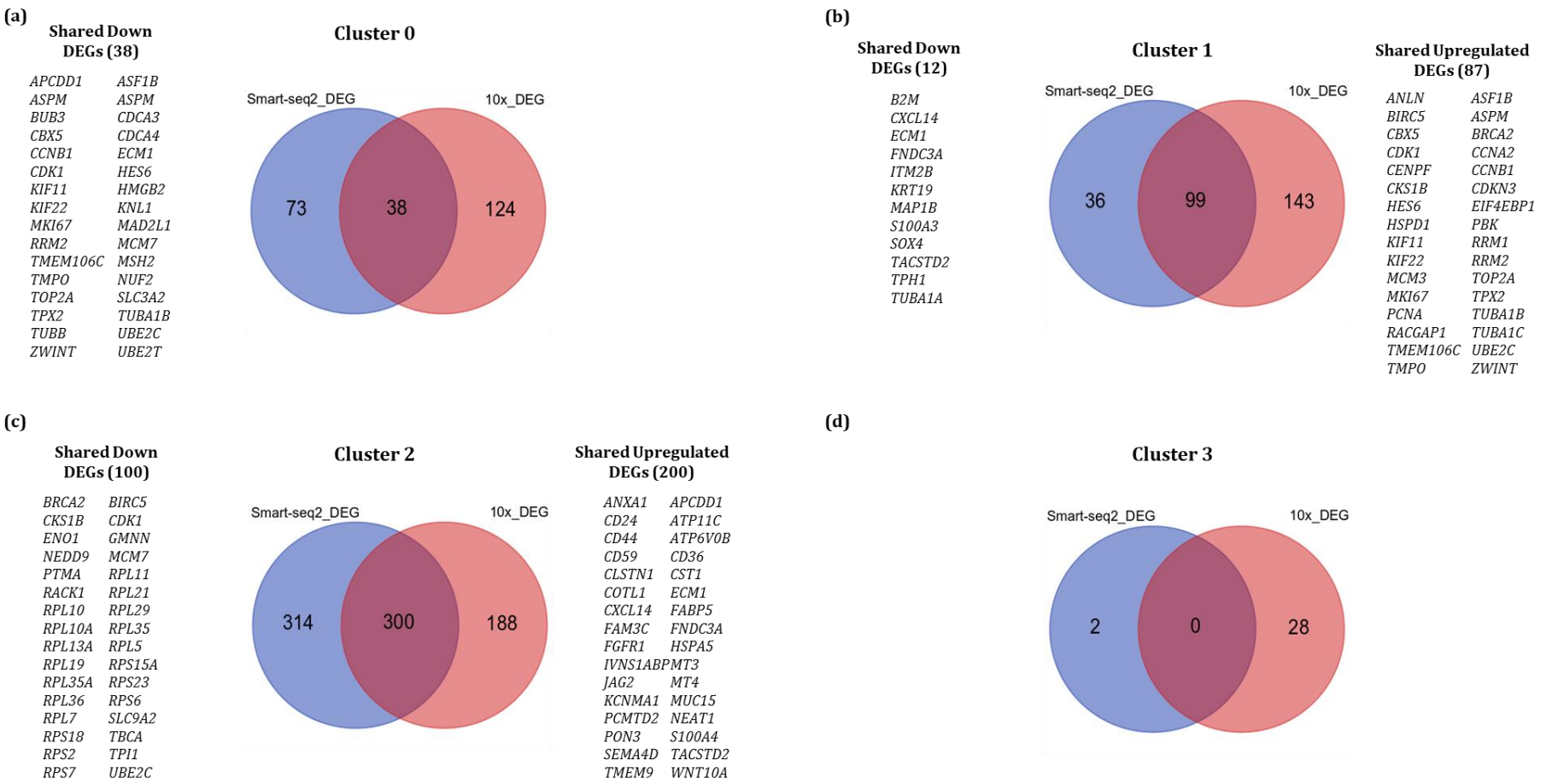


Figure 3.13. Cluster-level differentially expressed genes shared between Smart-seq2 and 10x scRNA-seq datasets

*Venn diagrams illustrate the number of differentially expressed genes in Clusters 0 to 3 (**a-d**), identified in both the Smart-seq2 dataset and 10x data containing Smart-seq2 cluster projections. Intersections in the Venn diagram indicate the DEGs shared by both datasets, with a selection of upregulated and downregulated genes listed adjacent to the diagrams.*

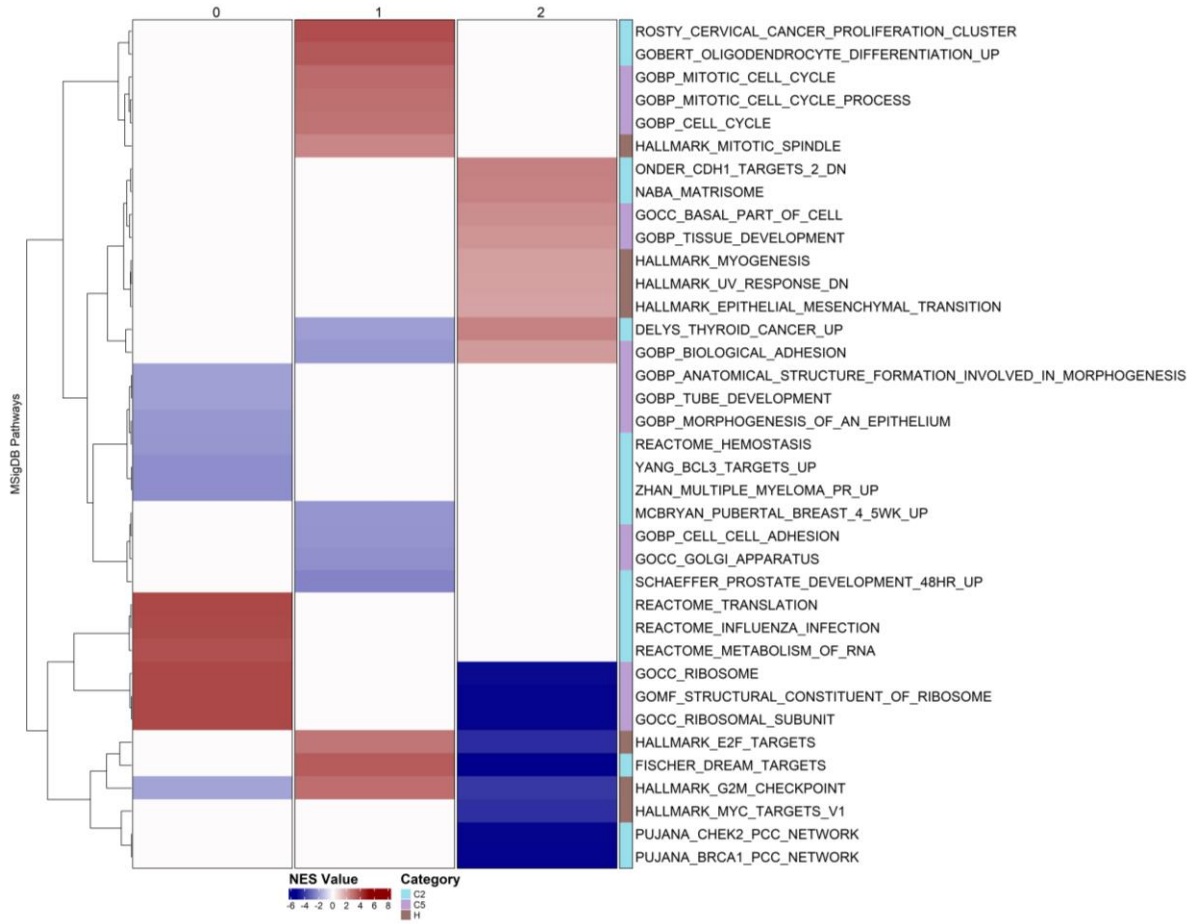


Figure 3.14. Gene set enrichment analysis of cluster markers in mCRC PDOs.

Heatmap rows show the top 3 pathways with significant enrichment (adjusted p-value < 0.05) in MSigDB categories for Clusters 0 to 3, ranked by Normalized Enrichment Score (NES). The colour scale ranges from blue (negative NES, indicating gene downregulation) to red (positive NES, indicating upregulation), with intensity denoting enrichment strength. White spaces show no category representation in a cluster. The analysis focused on the Curated (C2), Ontology (C5), and Hallmark (H) gene sets. Pathways are clustered by Euclidean distance. Note: Cluster 3 showed no significant pathways.

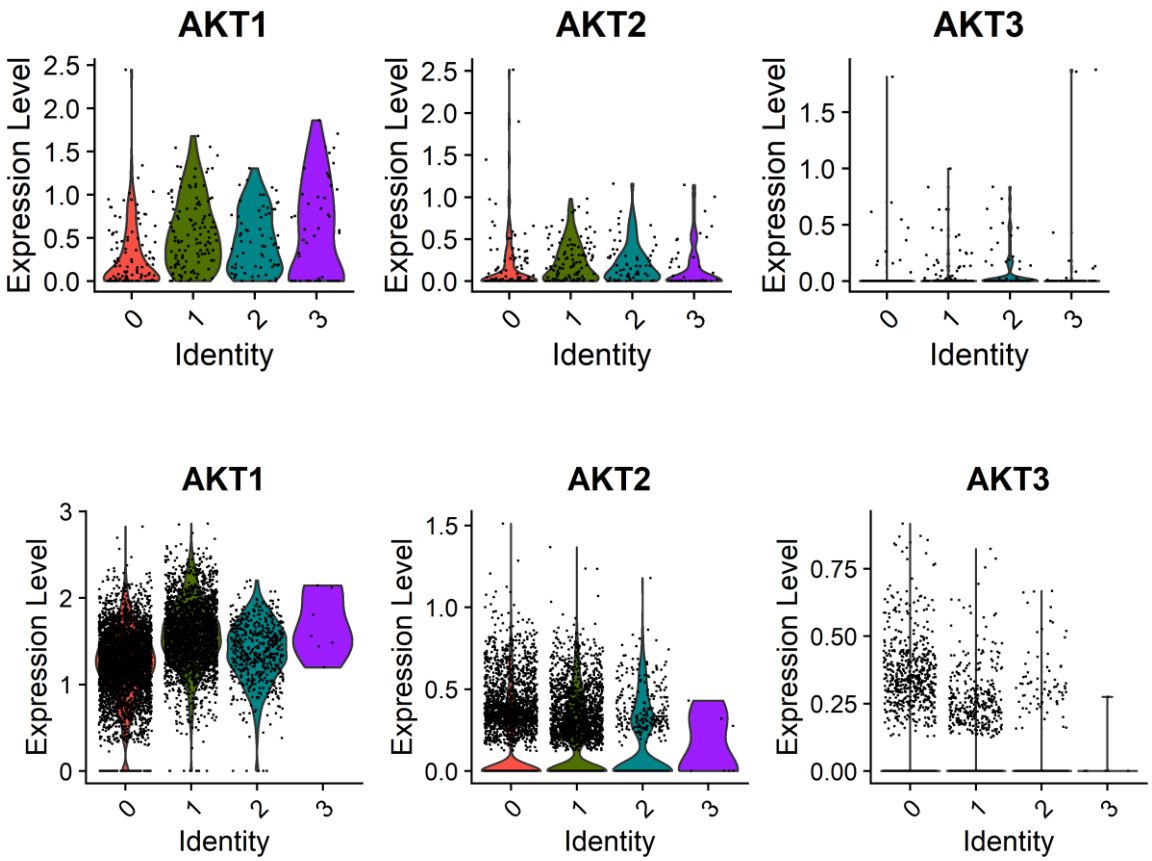


Figure 3.15. Gene expression of AKT isoforms across clusters identified in mCRC PDOs.

Violin plots illustrate the gene expression levels of genes encoding AKT isoforms across the four clusters identified in mCRC PDOs. The top panel displays gene expression data from the Smart-seq2 dataset, while the bottom panel presents the expression of the same clusters projected onto the 10x scRNA-seq dataset, providing a comparative analysis of AKT gene expression in different scRNA-seq technologies.

F. Reference-based cell type annotation uncovers the cellular diversity in mCRC PDOs

The cluster-based DGE analysis identified several genes that are known markers of various intestinal epithelial cell types. In Cluster 1, *MKI67*, *TOP2A* and *UBE2C* are markers of the highly proliferative transit-amplifying cells (190, 195). Among the DEGs identified in Cluster 2, *CD36* is expressed on the apical surface of enterocytes in the proximal colon (230); the protein encoded by *TPH1* is a key enzyme used by enterochromaffin cells for serotonin production (190, 231); and *ALDH1A* has been identified as a biomarker of normal and cancer stem cells, particularly in tissues where this gene is not typically expressed at high levels (e.g., breast, lung, oesophagus, colon and stomach) (232, 233). Building on these findings, the subsequent step involved annotating mCRC cells using the Human Gut Cell Atlas (HGCA) as a reference for cell type classification (234).

Cell type labels were assigned to individual cells, and not clusters, in the mCRC PDOs (Figure 3.16A). In this way, the mCRC PDO cells were annotated as various cell types, with Claudin-10-positive (CLDN10+) cells being the most prevalent (255 cells), followed by transit-amplifying (TA) cells (87), enterocytes (11), Paneth cells (6), and single instances of goblet and enterochromaffin cells expressing tachykinin (EC TAC1+) (Figure 3.16C-E).

In light of the observed variability in prediction scores (Supplementary Figure 2), further validation was conducted by examining the expression of marker genes characteristic of colonic cell types (Table 10). Some gene signatures displayed specificity to particular cell types. For example, *DEFA5* and *DEFA6* expression identified Paneth cells, reinforcing the accuracy of their annotation (Figure 3.16). In contrast, proliferative genes such as *MKI67*, *PCNA* and *TOP2A*, were observed across multiple cell types. These markers were not limited to TA cells but were also expressed in CLDN10+ cells within Cluster 1 and Cluster 3.

Besides CLDN10+ cells, TA cells showed a dispersed distribution across various clusters. Cluster 0 was previously characterised by enhanced protein synthesis and a general downregulation of genes involved in cell cycle progression. Cell type analysis indicated that this cluster primarily consisted of a mixture of CLDN10+ and TA cells, with a small presence of Paneth cells (Figure 3.16C). However, two key points need closer examination. Firstly, the prediction scores for the cell types identified in Cluster 0 were not exceptionally high (except for Paneth cells), which suggests a certain level of uncertainty in these annotations (Supplementary Figure 2). Secondly, the markers specific to CLDN10+ cells, and to a lesser extent, TA cells, exhibited only moderate expression levels (Figure 3.16B). This could imply

either an inherently low expression of these markers in mCRC organoids or potential limitations in their detection by the Smart-seq2 protocol.

Similar to Cluster 0, Cluster 1 was primarily composed of CLDN10+ cells and TA cells. Additionally, the clustering analysis revealed that Cluster exhibited a high proliferation rate and was mainly composed of undifferentiated cells, aligning with the recognized attributes of TA cells (235). While TA cells are more differentiated than stem cells, they retain the capacity to proliferate into more specialised colonic cell types.

Apart from CLDN10+ cells, enterocytes and a small proportion of Paneth cells were identified as the primary cell types in Cluster 2. This cluster was previously characterised by cells in the G1 phase with an elevated expression of genes associated with epithelial cell migration and adhesion. Furthermore, the strong expression of genes linked to enterocyte progenitor markers in mice, including *MAD2L1*, *MELK*, *PLK1* and *UBE2C* (236) (Supplementary Figure 3)—particularly in CLDN10+ cells within this cluster—further supports the hypothesis that CLDN10+ cells may indeed be colonic progenitors in mCRC organoids.

Cluster 3 was characterised by a diverse mixture of cell types, including TA, CLDN10+ cells, and a few Paneth cells. As previously stated in the text, Cluster 1 and Cluster 3 originated from the same cluster at the 0.3 resolution. This shared transcriptional background likely explains why both TA and CLDN10+ cells were present in Cluster 1 and Cluster 3.

The annotated Smart-seq2 dataset also enabled a detailed comparison of cell type abundances across mCRC PDOs. This analysis showed that AKTi-resistant PDOs contained twice the number of enterocytes compared to the Parental control (Figure 3.16E). In contrast, the proportions of CLDN10+ and TA cells remained similar between the AZD1-resistant and control PDOs but were reduced in the MK1-resistant organoid. Nevertheless, as with the clustering analysis, the validity of these comparisons must be considered carefully due to initial discrepancies in cell numbers between the organoids. To facilitate a more robust analysis of cell type abundances, the study was extended to include the 10x Genomics scRNA-seq dataset, providing a larger dataset for analysis.

During the annotation of the 10x Genomics scRNA-seq dataset with the HGCA, a notable discrepancy was observed: a significant number of cells were predicted as Paneth cells (Supplementary Figure 9A). This was unexpected, given the large number of Paneth cells identified in mCRC PDOs when these cells predominantly reside in the small intestine rather than the colon (237). The high prediction scores, ranging between 0.60 to 0.75, initially suggested a strong presence of Paneth-like cells in the 10x scRNA-seq dataset. However, upon

a closer examination of *DEFA5* and *DEFA6* expression, these markers were confined to a small subset of cells, isolated in a distinct region of the dataset (Supplementary Figure 9B). This observation indicated that while the prediction scores were high, the actual expression of Paneth cell-defining markers was limited, casting doubt on the accuracy of the initial annotations. To address this, the HGCA-annotated Smart-seq2 dataset was employed to label the 10x Genomics scRNA-seq dataset (Figure 3.17A). This approach yielded a more consistent distribution of prediction scores and a more reliable representation of cell types (Supplementary Figure 9C).

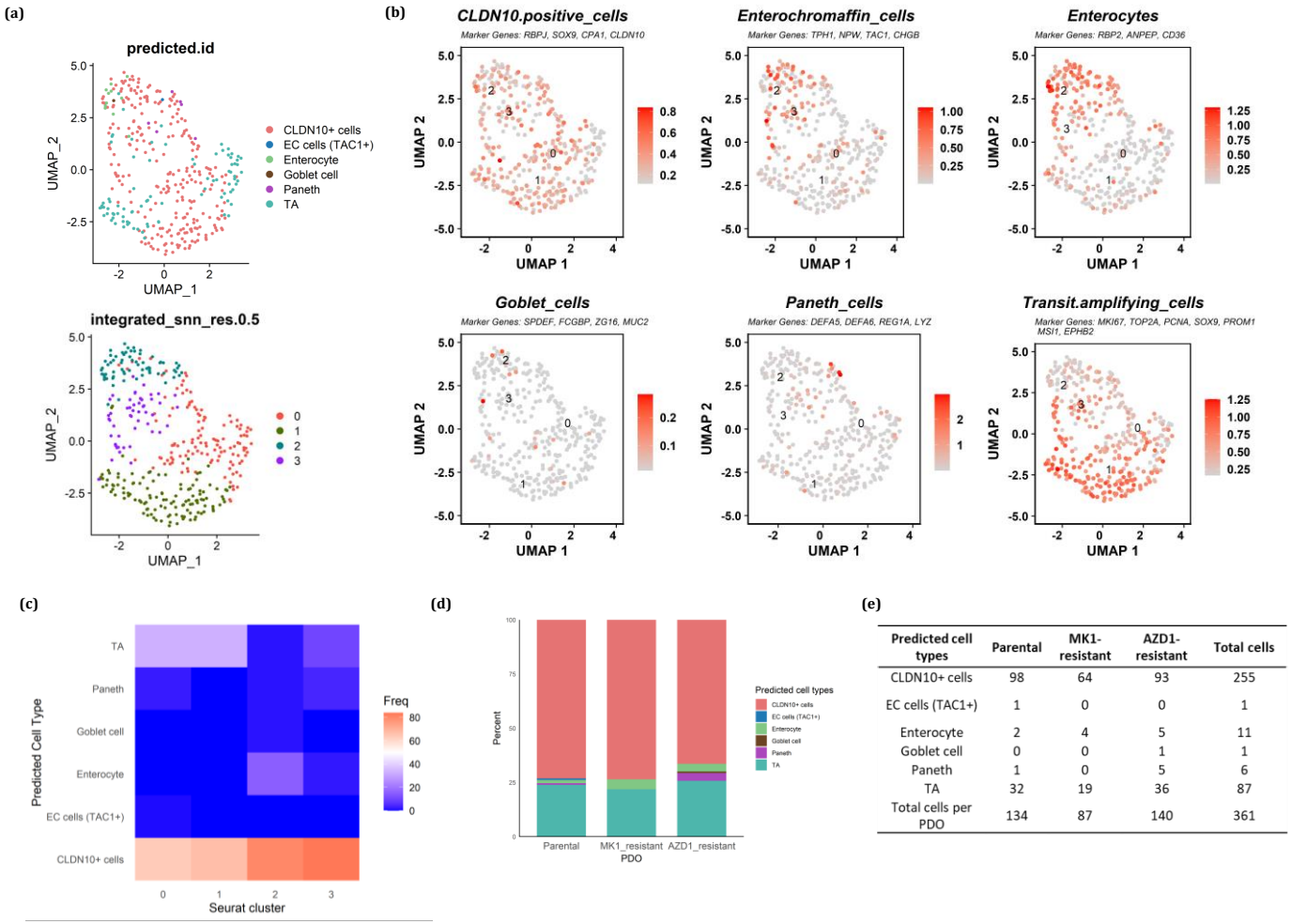
The refined analysis identified CLDN10+ cells, TA cells and enterocytes in the 10x dataset (Figure 3.17A). Notably, cells expressing *DEFA5* and *DEFA6* were not classified as Paneth cells in the 10x dataset, likely due to only six such cells being present in the Smart-seq2 reference, which is insufficient for a robust comparative expression analysis (no goblet cells or EC TAC1+ cells were identified in the 10x data for the same reason). The expression of these genes was instead associated with CLDN10+ cells (Figure 3.17B), supporting the notion that CLDN10+ cells may embody a differentiation spectrum ranging from the undifferentiated state of TA cells to, potentially, terminally differentiated such as Paneth cells.

The comparative analysis of cell type abundances between control and AKTi-resistant PDOs in the 10x dataset revealed findings consistent with those obtained from the Smart-seq2 data (Figure 3.17C-E). Specifically, the highest number of enterocytes were identified in AKTi-resistant PDOs (8 in each), while none were present in the control sample (Figure 3.17E). Notably, the MK1-resistant PDO exhibited a marked increase in immature cell populations, with almost 250 more CLDN10+ cells and over twice the number of TA cells compared to the AZD1-resistant and control organoids. However, it is essential to remember that the MK1-resistant dataset started with a higher cell count, which could influence the observed increase of these cell types. Despite this, the clustering analysis further corroborated these findings by showing that Cluster 1 of the MK1-resistant PDO presented the highest number of proliferative and undifferentiated cells across the three PDOs.

The general trend of an increased presence of undifferentiated intestinal cell types, such as TA and CLDN10+ progenitor cells, in both datasets compared to terminally differentiated cells may be attributed to the fact that over 94% of colorectal cancers exhibit mutations in genes involved in the Wnt/β-catenin signalling pathway, including *APC*, *CTNNB1* and *AXIN* (203). These mutations lead to the constitutive activation of the signalling pathway in mCRC, which creates an environment that promotes the proliferation and self-renewal of stem-like cells while minimising the presence of terminally differentiated cell types. Although the organoid culture

medium was also optimised for stem cell proliferation with additives such as Wnt-3a, a mitogen that activates the Wnt/β-catenin signalling, this supplementation is redundant for the establishment of colon organoids harbouring such mutations (although it is necessary to grow normal colon organoids) (203). However, Wnt-3a is necessary to increase the establishment rate of organoids (203, 237). Nevertheless, even with culture conditions favouring stemness, the presence of a Wnt gradient within the organoid culture still facilitated the differentiation of some stem cells into specialised cell types such as enterocytes, Paneth cells, goblet cells, and EC TAC1+ cells in the mCRC PDOs, albeit in limited numbers.

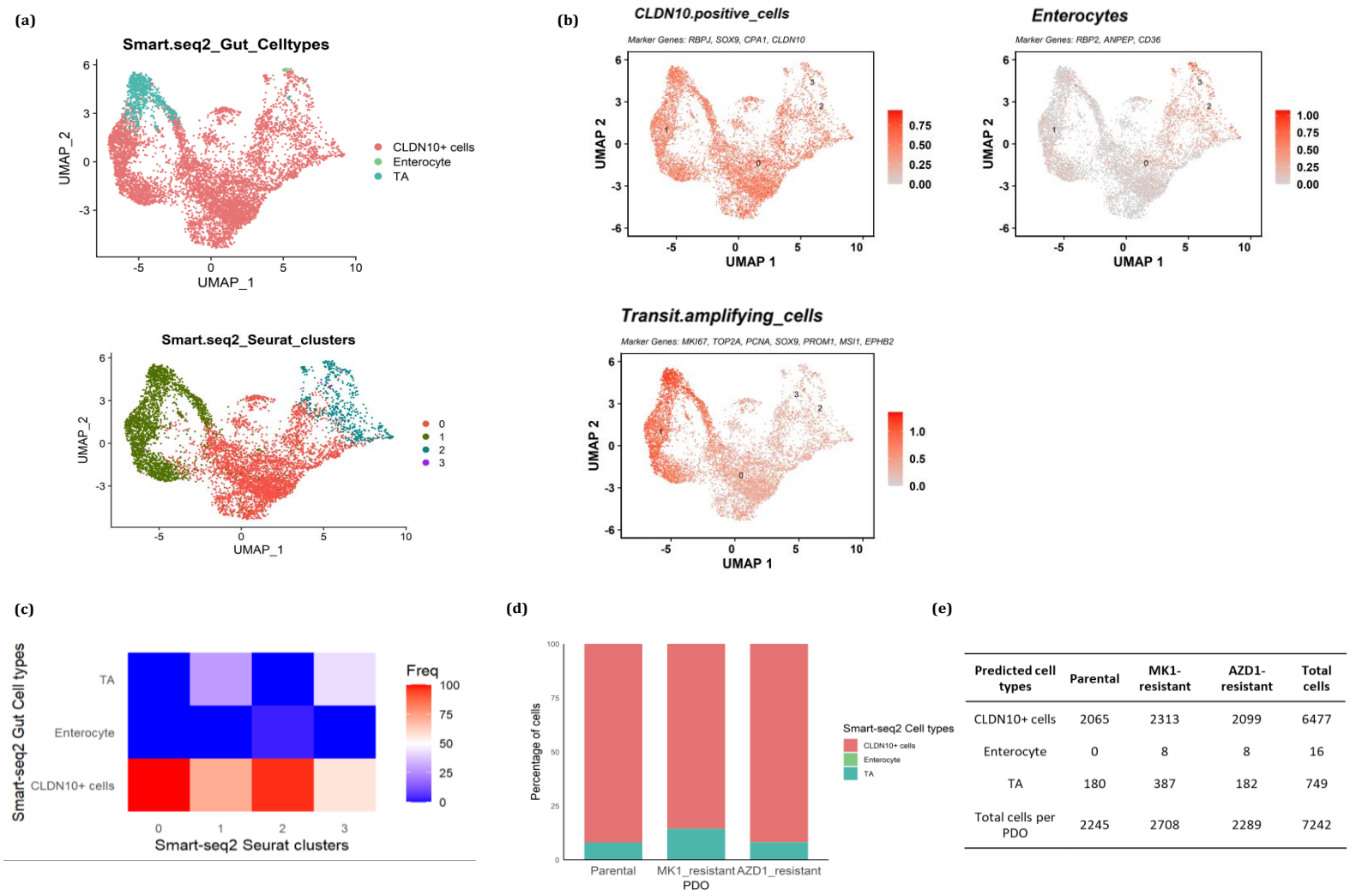
In summary, the cell-based annotation approach was invaluable for identifying subtle variations in gene expression among individual cells, revealing heterogeneity that might have been missed by assigning cell type labels to seemingly uniform clusters.



Predicted cell types	Parental	MK1-resistant	AZD1-resistant	Total cells
CLDN10+ cells	98	64	93	255
EC cells (TAC1+)	1	0	0	1
Enterocyte	2	4	5	11
Goblet cell	0	0	1	1
Paneth	1	0	5	6
TA	32	19	36	87
Total cells per PDO	134	87	140	361

Figure 3.16. Cell-type classification of Smart-seq2 data using the Human Gut Cell Atlas.

(a) *UMAP plots illustrate colonic cell types identified in mCRC PDOs, annotated with the HGCA (top), and the same cells grouped based on clustering identity (bottom).* ***(b)*** *UMAP plots illustrate the expression patterns of marker genes characterising colonic cell types.* ***(c)*** *Heatmap displays the frequency of colonic cell types across clusters identified in mCRC PDOs.* ***(d)*** *Proportion of cell types per PDO.* ***(e)*** *Cell type counts in each PDO.*



Predicted cell types	Parental	MK1-resistant	AZD1-resistant	Total cells
CLDN10+ cells	2065	2313	2099	6477
Enterocyte	0	8	8	16
TA	180	387	182	749
Total cells per PDO	2245	2708	2289	7242

Figure 3.17. 10x scRNA-seq data visualisation of colonic cell types annotated using Smart-seq2 data labelled with the HGCA.

3.4. Results

(a) *UMAP plots illustrate colonic cell types identified in 10x scRNA-seq data annotated with a Smart-seq2 dataset previously annotated with the HGCA (top), and the same cells grouped based on clustering identity (bottom).* ***(b)*** *Gene expression patterns of marker genes characterising colonic cell types.* ***(c)*** *Heatmap displays the frequency of colonic cell types across clusters identified in mCRC PDOs.* ***(d)*** *Proportion of cell types per PDO.* ***(e)*** *Cell type counts in each PDO.*

G. Differential gene expression analysis reveals transcriptional profiles associated with drug resistance mechanisms in mCRC organoids

Differential gene expression (DGE) analysis was performed on AKTi-resistant and untreated mCRC PDOs to identify gene expression patterns contributing to the development of drug resistance to AKT inhibition. To achieve this, two independent DGE analyses were performed, one on the Smart-seq2 dataset and a second analysis on the 10x Genomics scRNA-seq dataset. While the Smart-seq2 findings remained the primary focus, the 10x Genomics analysis provided an independent layer for validation. This dual approach ensured that any shared differentially expressed genes (DEGs) identified between the datasets could be attributed solely to intrinsic gene expression patterns within the organoids without being influenced by the introduction of labels or cluster information from one dataset to another, thereby enhancing the reliability of the findings.

After the DGE analysis, an over-representation analysis (ORA) was conducted on the significant DEGs identified in the Smart-seq2 dataset to assess the enrichment of predefined gene sets from the Molecular Signatures Database (MSigDB) (163). The analysis focused mainly on the Curated (C2), Gene Ontology (C5), Oncogenic signature (C6) and Hallmark (H) gene set categories.

Under predefined expression and statistical significance thresholds (average log2 fold change > |0.5|, and adjusted *p*-value < 0.05), a total of 131 DEGs were identified in the Smart-seq2 dataset when comparing the MK1-resistant PDO to the untreated Parental PDO, with 57 genes upregulated and 74 genes downregulated (Table 12 and Supplementary Table 1). In comparison, 176 DEGs were identified in the 10x Genomics dataset (Figure 3.18A). Of the 131 DEGs identified in the Smart-seq2 dataset, 44 (33.6%) were also present in the 10x Genomics dataset, including 21 upregulated and 23 downregulated genes (Figure 3.18B). Table 13 presents a selection of dysregulated genes, ordered by decreasing avg_log2FC. For the complete results of the DGE analysis from the 10x dataset, refer to Appendix 4 “Pairwise DGE analysis between MK1-resistant and Parental PDOs”.

MUC21, which was exclusively dysregulated in the Smart-seq2 dataset, showed the highest upregulation (Table 13). The protein encoded by *MUC21* is a highly glycosylated transmembrane mucin (229). Functionally, Muc21, like other mucins, protects the underlying epithelia from physical, chemical, and biological insults (238). Beyond this protective role, Muc21 is also involved in cell-cell adhesion, signal transduction, and modulating cell surface proteins (239). Previously, *MUC21* emerged as a marker gene in Cluster 2 (Figure 3.12A), a cluster characterised by genes associated with epithelial cell adhesion, mobility, and processes

like the epithelial-to-mesenchymal transition (EMT), all of which are known to promote cellular invasion (Figure 3.14).

Also upregulated in the MK1-resistant PDO were genes related to cytoskeletal reorganisation, motility, cellular shape, and extracellular-matrix (ECM) alterations. These include *ACTB*, which provides the structural framework for cell division, migration and cell signalling (229); *CFL1*, a regulator of actin filament dynamics; *TMSB4X*, involved in actin polymerisation; *KRT18*, a keratin of the intermediate-filament family of cytoskeletal proteins; and *S100A4*, which interacts with cytoskeletal and ECM proteins (229). As expected, given the expression of these genes, the ORA analysis for this AKT-resistant mCRC organoid revealed enrichment in gene sets associated with cell substrate junctions (e.g., C5: GOCC_CELL_SUBSTRATE_JUNCTION) (Figure 3.19), highlighting a potential upregulation of cellular adhesion, signalling, and interaction with the extracellular matrix. The DGE analysis further revealed elevated expression of genes encoding cell surface proteins, including the tumour-associated calcium signal transducer 2 (*TACSTD2*) (240) and basigin (*BSG*, also known as *EMMPRIN* or *CD147*) (241).

Shifting focus from genes associated with structural alterations, MK1-resistant cells demonstrated a significant upregulation of genes encoding detoxification enzymes. These include glutathione peroxidase 1 (*GPX1*), glutathione S-transferase pi 1 (*GSTP1*), and Parkinson's disease protein 7 (*PARK7*, also known as DJ-1) (229), as observed in both the Smart-seq2 and 10x datasets. Additionally, genes directly involved in glycolysis and related biosynthetic pathways were notably upregulated, including *ENO1*, *PGD*, *SLC2A1* and *SLC6A14*. Among these, the solute carrier (SLC) membrane transporter *SLC2A1*, also known as *GLUT1*, facilitates glucose uptake in malignant neoplasms (242).

Continuing with the theme of metabolic alterations, the DGE analysis also revealed the upregulation of *PDK3*. *PDK3* encodes a kinase that inhibits the pyruvate dehydrogenase (PDH) complex, which catalyses the oxidative decarboxylation of pyruvate to acetyl coenzyme A (acetyl-CoA), an entry molecule for the tricarboxylic (or Krebs) cycle (229, 243). Also upregulated were fatty acid transporters (*FABP5*) (244), protein-folding genes or chaperones (*PDIA3* and *PDIA6, HSP90AB1*) (245, 246) and Ribophorin II (*RPN2*), an integral glycoprotein of the rough endoplasmic reticulum (ER), crucial for protein processing (247).

Furthermore, the MK1-resistant PDO displayed moderate overexpression of mitochondrial complex subunits. Notably, this included *MT-ND1* and *MT-ND4*, which were also overexpressed in the 10x dataset, along with *MT-ND2*, *MT-ND3*, *MT-ND5*, and *MT-ND6*, all of which encode subunits of NADH dehydrogenase (Complex I)—a key component of the electron transport

chain. Also co-expressed with these genes was *MT-CYB*, a component of ubiquinol–cytochrome c reductase (Complex III) (248).

Regarding downregulated genes in the MK1-resistant PDO, there was a notable decrease in the expression of ribosomal-related genes compared to the Parental control (Figure 3.18A-B). This observation aligns with the Seurat Smart-seq2 quality control analysis findings, which indicated that the Parental line exhibited the highest ribosomal content among all organoids (Figure 3.6D). A similar trend was observed in the 10x analysis, confirming that this reduction in ribosomal gene expression is a biological effect rather than a technical artefact.

Among other notable downregulated genes identified in the Smart-seq2 and 10x datasets was *IGFBP2*, which primarily binds to insulin-like growth factors (IGFs). This binding regulates the bioavailability of IGFs and modulates the interactions with their respective receptors (249). On the other hand, *DEFA5* was identified as the most significantly downregulated gene in MK1-resistant cells, exclusively in the Smart-seq2 dataset. Human defensin 5, known for its role in host defence, is a highly conserved antimicrobial peptide predominantly secreted by Paneth cells (250).

Following the pattern of downregulated genes associated with intestinal cells signatures, *LGR5*, a stem cell marker, was also slightly downregulated in the MK1-resistant. This observation is especially significant considering the role of the LGR5 receptor in enhancing Wnt signalling in CRC (28).

3.4. Results

Table 12. Top 10 statistically significant upregulated and downregulated genes by avg_log2FC in the MK1-resistant PDO

Gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	pct_diff	Gene_id	Expression
MUC21	2.02E-08	2.384874	0.747	0.47	0.000487	0.277	ENSG00000204544	Upregulated
CXCL14	2.05E-07	2.240981	0.954	0.978	0.004935	-0.024	ENSG00000145824	Upregulated
TACSTD2	3.43E-08	2.199259	0.897	0.858	0.000826	0.039	ENSG00000184292	Upregulated
LGALS1	1.63E-09	1.95564	0.92	0.873	3.93E-05	0.047	ENSG00000100097	Upregulated
CEACAM6	1.15E-06	1.953189	0.828	0.784	0.027699	0.044	ENSG00000086548	Upregulated
LRATD1	4.27E-08	1.4727	0.793	0.582	0.001028	0.211	ENSG00000162981	Upregulated
FABP5	6.71E-09	1.315372	0.931	0.97	0.000162	-0.039	ENSG00000164687	Upregulated
SDR16C5	2.35E-08	1.22146	0.678	0.433	0.000566	0.245	ENSG00000170786	Upregulated
EPAS1	2.13E-10	1.173951	0.874	0.903	5.12E-06	-0.029	ENSG00000116016	Upregulated
SLC2A1	2.14E-07	1.155772	0.851	0.731	0.005162	0.12	ENSG00000117394	Upregulated
RDH10	2.25E-10	-0.88481	0.747	0.91	5.42E-06	-0.163	ENSG00000121039	Downregulated
RPL9	1.57E-20	-0.89717	0.943	1	3.78E-16	-0.057	ENSG00000163682	Downregulated
RPL13A	3.48E-19	-0.91091	0.977	1	8.39E-15	-0.023	ENSG00000142541	Downregulated
RPL22L1	5.88E-13	-0.93615	0.759	0.94	1.42E-08	-0.181	ENSG00000163584	Downregulated
RPL23	1.57E-18	-0.98915	0.908	1	3.78E-14	-0.092	ENSG00000125691	Downregulated
IGFBP2	3.22E-18	-1.00172	0.345	0.851	7.76E-14	-0.506	ENSG00000115457	Downregulated
RPS3A	8.11E-23	-1.02767	0.943	1	1.95E-18	-0.057	ENSG00000145425	Downregulated
RPL7	5.61E-17	-1.13727	1	1	1.35E-12	0	ENSG00000147604	Downregulated
RPS6	4.39E-25	-1.1452	0.977	1	1.06E-20	-0.023	ENSG00000137154	Downregulated
DEFA5	1.23E-07	-2.63208	0.276	0.627	0.002952	-0.351	ENSG00000164816	Downregulated

Note: pct.1 and pct.2 indicate the percentage of cells in the first and second group that express the gene, respectively, while pct_diff represents the difference in the percentage of cells expressing the gene between the two groups.

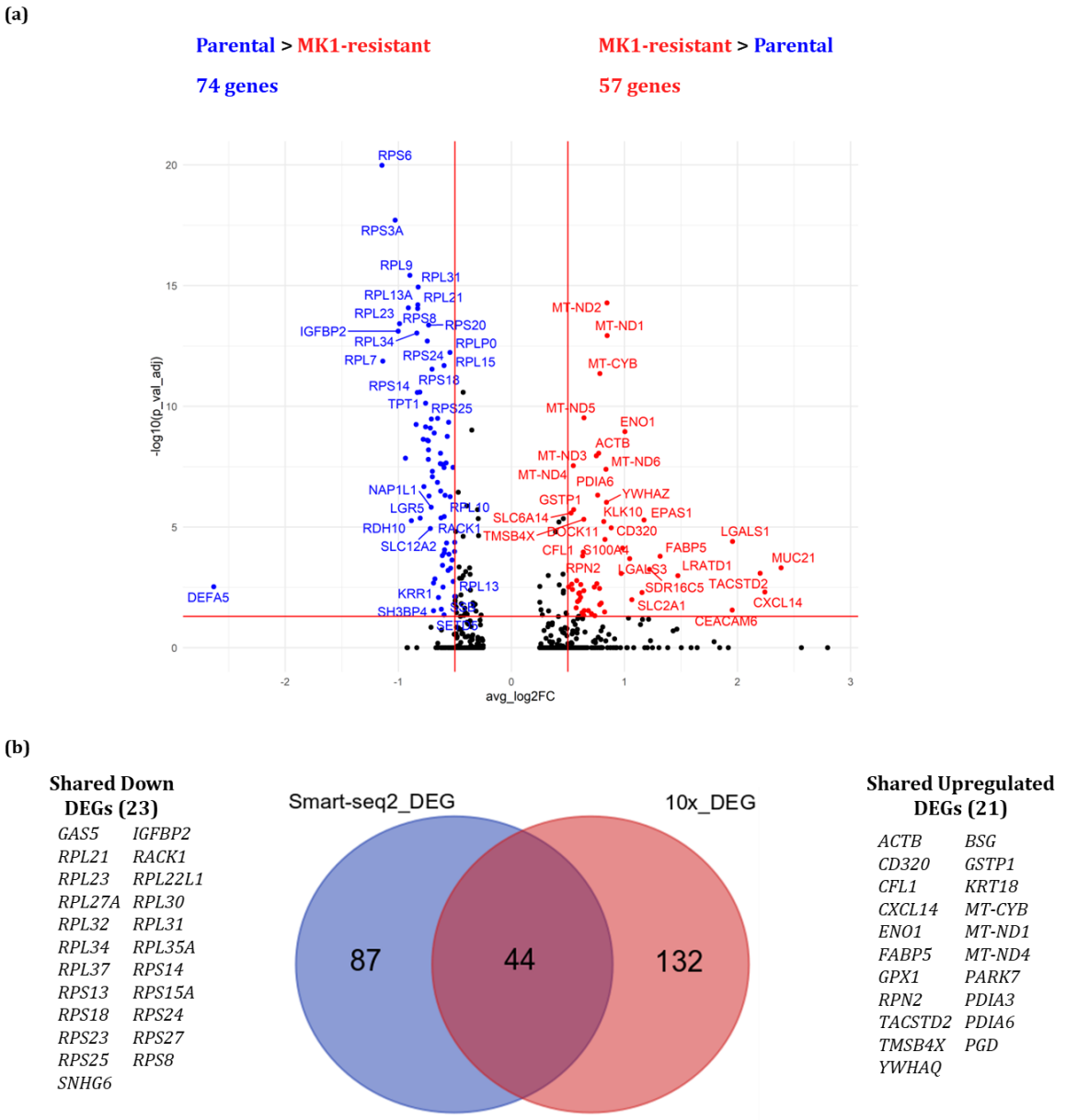


Figure 3.18. Differential expression analysis between MK1-resistant and Parental PDOs.

(a) *Volcano plot illustrates differentially expressed genes in the MK-2206-resistant PDO relative to the untreated Parental control. The x-axis represents the average log2 fold change, and the y-axis depicts the significance -log(adjusted p-value). Vertical lines at x=-0.5 and x=0.5 delineate fold change limits, while a horizontal line at y=-log10(0.05)~1.3 indicates the significance threshold. Genes are colour-coded: blue for down-regulation, red for up-regulation, and black for non-significant genes. Upregulated and downregulated genes are annotated on the plot.* ***(b)*** *Venn diagram illustrates the number of DEGs detected in the Smart-seq2 and 10x Genomics scRNA-seq datasets. This includes genes with average log2 fold change ≥ |0.5| and adjusted p-value < 0.05 in both datasets. The intersection represents DEGs shared by both datasets. All upregulated and downregulated genes are listed adjacent to the Venn diagram.*

3.4. Results

Table 13. Dysregulated genes in the MK1-resistant PDO observed in Smart-seq2 and 10x datasets

Genes	Locus	avg_log2FC	p_val_adj	Protein function
MUC21	6p21.33	2.385	4.87E-04	mucin 21, cell surface associated [Source:HGNC Symbol;Acc:HGNC:21661]
*CXCL14**	5q31.1	2.241	4.94E-03	C-X-C motif chemokine ligand 14 [Source:HGNC Symbol;Acc:HGNC:10640]
*TACSTD2**	1p32.1	2.199	8.26E-04	tumor associated calcium signal transducer 2 [Source:HGNC Symbol;Acc:HGNC:11530]
CEACAM6	19q13.2	1.953	2.77E-02	carcinoembryonic antigen related cell adhesion molecule 6 [Source:HGNC Symbol;Acc:HGNC:1818]
*FABP5**	8q21.13	1.315	1.62E-04	fatty acid binding protein 5 [Source:HGNC Symbol;Acc:HGNC:3560]
SLC2A1	1p34.2	1.156	5.16E-03	solute carrier family 2 member 1 [Source:HGNC Symbol;Acc:HGNC:11005]
PDK3	Xp22.11	1.066	1.03E-02	pyruvate dehydrogenase kinase 3 [Source:HGNC Symbol;Acc:HGNC:8811]
*ENO1**	1p36.23	1.004	1.12E-09	enolase 1 [Source:HGNC Symbol;Acc:HGNC:3350]
*PGD**	1p36.22	0.795	1.41E-02	phosphogluconate dehydrogenase [Source:HGNC Symbol;Acc:HGNC:8891]
*PDIA6**	1p36.23	0.764	4.76E-07	protein disulfide isomerase family A member 6 [Source:HGNC Symbol;Acc:HGNC:30168]
*TMSB4X**	Xp22.2	0.641	4.80E-06	thymosin beta 4 X-linked [Source:HGNC Symbol;Acc:HGNC:11881]
*PDIA3**	15q15.3	0.641	4.30E-03	protein disulfide isomerase family A member 3 [Source:HGNC Symbol;Acc:HGNC:4606]
*CFL1**	11q13.1	0.635	1.08E-04	cofilin 1 [Source:HGNC Symbol;Acc:HGNC:1874]
*PARK7**	1p36.23	0.596	5.59E-03	Parkinsonism associated deglycase [Source:HGNC Symbol;Acc:HGNC:16369]
*KRT18**	12q13.13	0.573	2.24E-02	keratin 18 [Source:HGNC Symbol;Acc:HGNC:6430]
*GSTP1**	11q13.2	0.551	1.90E-06	glutathione S-transferase pi 1 [Source:HGNC Symbol;Acc:HGNC:4638]
*BSG**	19p13.3	0.505	3.02E-03	basigin (Ok blood group) [Source:HGNC Symbol;Acc:HGNC:1116]
EEF1A1	6q13	-0.542	5.57E-07	eukaryotic translation elongation factor 1 alpha 1 [Source:HGNC Symbol;Acc:HGNC:3189]
*RACK1**	5q35.3	-0.622	4.23E-06	receptor for activated C kinase 1 [Source:HGNC Symbol;Acc:HGNC:4399]
NPM1	5q35.1	-0.699	8.29E-08	nucleophosmin 1 [Source:HGNC Symbol;Acc:HGNC:7910]
*GAS5**	1q25.1	-0.779	2.33E-09	growth arrest specific 5 [Source:HGNC Symbol;Acc:HGNC:16355]
LGR5	12q21.1	-0.806	4.26E-06	leucine rich repeat containing G protein-coupled receptor 5 [Source:HGNC Symbol;Acc:HGNC:4504]
*RPS14**	5q33.1	-0.809	2.59E-11	ribosomal protein S14 [Source:HGNC Symbol;Acc:HGNC:10387]
*IGFBP2**	2q35	-1.002	7.76E-14	insulin like growth factor binding protein 2 [Source:HGNC Symbol;Acc:HGNC:5471]
DEFA5	8p23.1	-2.632	2.95E-03	defensin alpha 5 [Source:HGNC Symbol;Acc:HGNC:2764]

*Genes differentially expressed in the Smart-seq2 and 10x Genomics scRNA-seq datasets.

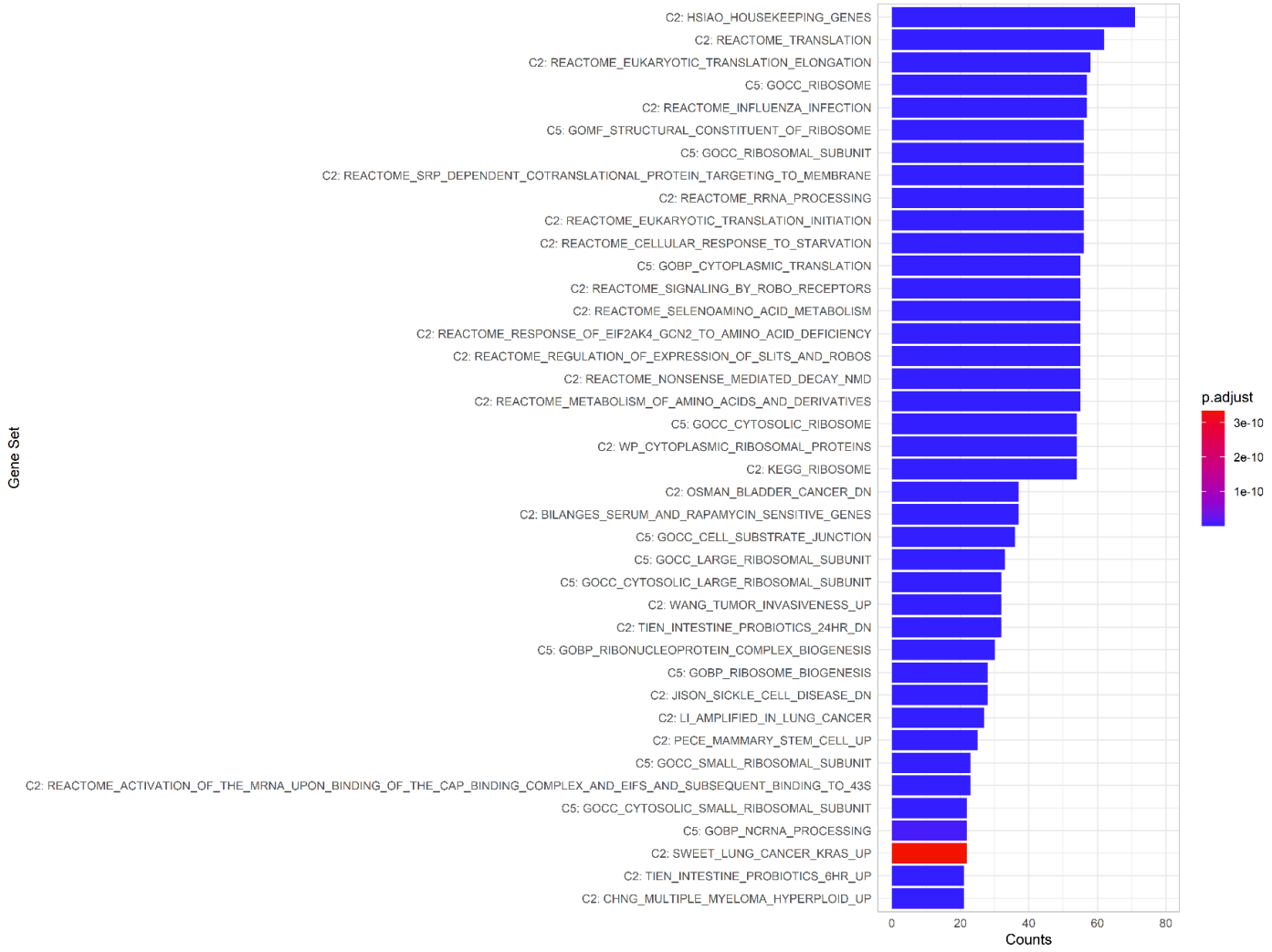


Figure 3.19. Over-representation analysis of differentially expressed genes in the MK1-resistant organoid.

Top 40 gene sets over-represented in MK1-resistant mCRC organoid. Bar length indicates the count of DEGs overlapping each gene set. Colour gradient represents the level of statistical significance, with a threshold set at an adjusted p-value < 0.05.

In the DGE analysis of the AZD1-resistant PDO, a total of 109 DEGs were identified in the Smart-seq2 dataset (avg_log2FC > |0.5|, *p*-value < 0.05), with 73 genes upregulated and 36 genes downregulated (Table 14 and Supplementary Table 2). In comparison, 193 DEGs were identified in the 10x Genomics dataset (Figure 3.20A). Of the 109 DEGs identified in the Smart-seq2 dataset, 40 DEGs (36.7%) were also present in the 10x Genomics dataset, including 32 upregulated genes and 8 downregulated genes (Figure 3.20B). Table 15 presents a selection of dysregulated genes, ordered by decreasing avg_log2FC. For the complete results of the DGE analysis from the 10x scRNA-seq dataset, refer to Appendix 4 "Pairwise DGE analysis between AZD1-resistant and Parental PDOs".

In the AZD1-resistant PDO, *C6orf15* exhibited the highest upregulation in the Smart-seq2 dataset (Table 15). *C6orf15* plays a role in several processes related to the organisation and function of the extracellular matrix (ECM), including collagen, fibronectin, and glycosaminoglycan binding (229). Besides *C6orf15*, the AZD1-resistant organoid also overexpressed other structural and ECM-remodelling genes, such as *FNDC3A*, *ECM1*, *TACSTD2*, and *TMSB4X*. Interestingly, the latter three were similarly upregulated in the MK1-resistant organoid (Figure 3.18A).

Multiple genes from the S100 Ca^{2+}-binding protein family were found to be upregulated in the AZD1-resistant PDO. Notably, *S100A4*, which exhibited upregulation in both the Smart-seq2 and 10x datasets, was similarly upregulated in the MK1-resistant organoid. The *S100A4* proteins are known to play a role in cellular motility and extracellular matrix (ECM) remodelling (229). Although *S100A3* is not as well studied as *S100A4*, it is proposed to regulate cell cycle progression and differentiation (229). Similarly, *S100A6* can regulate cell survival in a RAGE-dependent manner and contributes to maintaining cellular stability by enhancing the functions of Hsp90 and Hsp70 chaperones under conditions of cellular stress (e.g., heat shock, oxidative stress, or the presence of exogenous cytotoxic substances) (251). Calreticulin (*CALR9*), another chaperone, was also upregulated in the AZD1-resistant organoid.

Also upregulated was *ENTPD1/CD39*, a cell-surface ectonucleotidase that hydrolyses the sequential conversion of extracellular adenosine triphosphate (ATP) into adenosine monophosphate (AMP) and has a known role in regulating immune responses (229, 252). Another key regulator of immune responses identified in the AZD1-resistant PDO was the chemokine *CXCL14*, which is important in establishing immune surveillance in normal epithelia (253).

Other upregulated genes in the AZD1-resistant organoid can be broadly categorised based on their functions. These include genes encoding cell surface proteins and receptors (*EPCAM*, *BSG*, *CD44*) (229); genes involved in fatty acid (*CD36*, *FABP5*) or carbohydrate (*PDP1*) metabolism; genes linked to detoxification processes (*MGST3*); genes related to protein folding, degradation or cleavage (*CALR9*, *PCYOX1*, *PRSS23*); genes involved in signal transduction (*TGFA*, *SEMA3A, PPP1CB*); and regulators of gene expression, including several long non-coding RNAs (lncRNAs) such as *XIST*, *LINC00867*, *NEAT1*) (229).

Regarding downregulated genes in the AZD1-resistant organoid, a significant shift was observed in the expression of ribosomal protein genes. Similar to the MK1-resistant PDO, the AZD1-resistant PDO also exhibited significant downregulation of genes belonging to the ribosomal protein family, including *RPL13A*, *RPL14*, *RPL22L1*, *RPL7*, *RPL8*, *RPS14*, and *RPS9*. This pattern aligns with the results from the ORA, which indicated enrichment for gene sets associated with many ribosomal-related processes, such as eukaryotic translation (Figure 3.21). Furthermore, there was a slight decrease in the expression of the heterogeneous nuclear ribonucleoprotein A1 (*HNRNPA1*), an RNA-binding protein essential for the regulation of alternative splicing (254), nucleophosmin 1 (*NPM1*), a multifunctional protein with chaperoning functions (255), and the eukaryotic translation elongation factor 1 alpha (*EEF1A1*), which is also important in the translation machinery (256).

Multiple genes integral to cell signalling were downregulated in the AZD1-resistant organoid. This included *IGFBP2*, which was also observed to be downregulated in the MK1-resistant organoid. *IGFBP2* plays a critical role in IGF signalling by regulating the availability and activity of insulin-like growth factors, key mediators in cell growth and metabolism (249). Consequently, *IGFBP2* influences several oncogenic processes, including proliferation, migration, EMT, angiogenesis, apoptosis, and immunoregulation (249, 257). Also downregulated was *RACK1*, which encodes a scaffold protein known for its interactions with the cytoplasmic domain of various receptors, including IGF-1R (258). Other notable genes with decreased expression were *HES6*, a transcription factor involved in NOTCH signalling (259), and the interleukin-2 receptor (*IL-2R*), essential for regulating immune responses in T and B cells (260). This pattern suggests a broad downregulation of multiple signalling pathways in the AZD1-resistant organoid.

In summary, the DGE analyses of MK-2206 and AZD5363-resistant mCRC PDOs revealed unique and shared gene expression patterns that may contribute to drug resistance mechanisms. Upregulation of genes linked to cytoskeletal reorganisation, cellular motility, and extracellular matrix changes suggests enhanced cell division, migration, and signalling capabilities.

Conversely, upregulation of genes related to the regulation of immune responses could indicate mechanisms by which cancer cells evade the immune system. Additionally, increased gene expression of enzymes involved in detoxification processes, glycolysis, and the pentose phosphate pathway signals a metabolic adaptation, while changes in genes related to mitochondrial function suggest modifications in energy metabolism. In contrast, the downregulation of genes related to ribosomal function and insulin-like growth factor binding indicates shifts in protein synthesis and growth factor signalling. These findings highlight the diversity of cellular strategies employed in response to different AKT inhibition strategies.

3.4. Results

Table 14. Top 10 statistically significant upregulated and downregulated genes by avg_log2FC in the AZD1-resistant PDO

Gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	pct_diff	Gene_id	Expression
C6orf15	7.57E-12	2.225761	0.793	0.552	1.82E-07	0.241	ENSG00000204542	Upregulated
CXCL14	4.46E-12	2.02619	0.986	0.978	1.07E-07	0.008	ENSG00000145824	Upregulated
S100A3	5.14E-11	1.860376	0.957	0.955	1.24E-06	0.002	ENSG00000188015	Upregulated
LINC00867	2.27E-12	1.485194	0.736	0.433	5.47E-08	0.303	ENSG00000232139	Upregulated
CD36	4.88E-12	1.464117	0.836	0.612	1.17E-07	0.224	ENSG00000135218	Upregulated
FNDC3A	1.93E-09	1.417007	0.929	0.903	4.64E-05	0.026	ENSG00000102531	Upregulated
AP1S2	3.46E-16	1.385132	0.871	0.776	8.33E-12	0.095	ENSG00000182287	Upregulated
NEAT1	1.13E-11	1.338621	0.95	1	2.72E-07	-0.05	ENSG00000245532	Upregulated
SEMA3C	1.20E-10	1.280454	0.907	0.843	2.89E-06	0.064	ENSG00000075223	Upregulated
ECM1	5.72E-09	1.230542	0.836	0.724	0.000138	0.112	ENSG00000143369	Upregulated
CALCA	3.20E-07	-0.66291	0.129	0.396	0.00771	-0.267	ENSG00000110680	Downregulated
TESC	6.09E-08	-0.67436	0.821	0.925	0.001468	-0.104	ENSG00000088992	Downregulated
RPL13	2.09E-14	-0.6865	0.964	1	5.04E-10	-0.036	ENSG00000167526	Downregulated
PLEKHB1	3.03E-09	-0.73659	0.614	0.799	7.31E-05	-0.185	ENSG00000021300	Downregulated
EEF1A1	1.35E-22	-0.7544	1	1	3.26E-18	0	ENSG00000156508	Downregulated
SLC25A6	7.54E-15	-0.79627	0.921	0.97	1.82E-10	-0.049	ENSG00000169100	Downregulated
HES6	5.65E-09	-0.83116	0.814	0.91	0.000136	-0.096	ENSG00000144485	Downregulated
RACK1	4.22E-22	-0.8878	0.986	1	1.02E-17	-0.014	ENSG00000204628	Downregulated
RPL3	1.17E-23	-1.09246	0.979	0.993	2.82E-19	-0.014	ENSG00000100316	Downregulated
MT-RNR2	1.61E-14	-1.25719	1	1	3.88E-10	0	ENSG00000210082	Downregulated

Note: pct.1 and pct.2 indicate the percentage of cells in the first and second group that express the gene, respectively, while pct_diff represents the difference in the percentage of cells expressing the gene between the two groups.

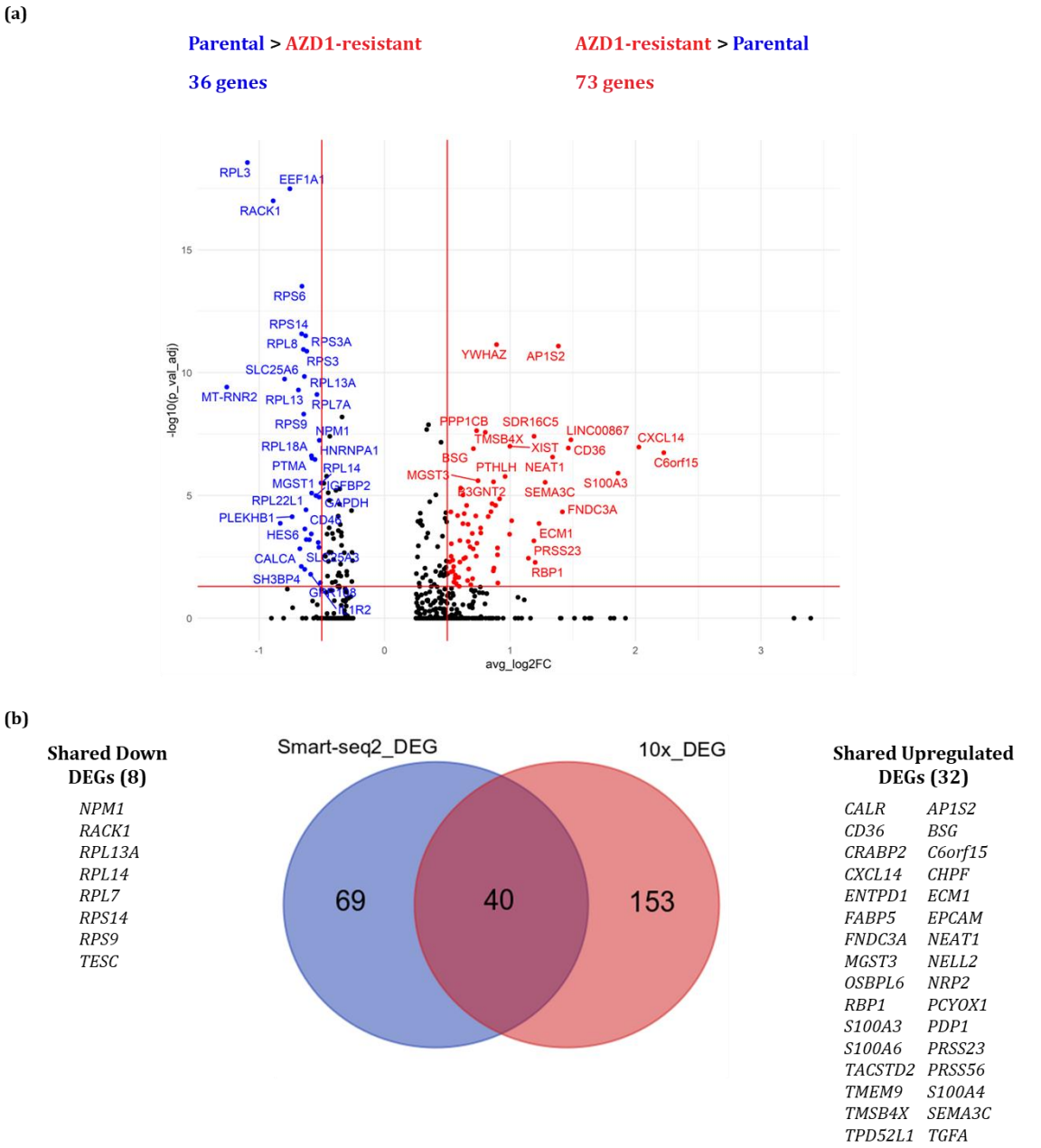


Figure 3.20. Differential expression analysis between AZD1-resistant and Parental PDOs.

(a) *Differentially expressed genes in the AZD5363-resistant PDO relative to the untreated Parental control. The x-axis represents the average log2 fold change, and the y-axis depicts the significance -log(adjusted p-value). Vertical lines at x=-0.5 and x=0.5 delineate fold change limits, while a horizontal line at y=-log10(0.05)~1.3 indicates the significance threshold. Genes are colour-coded: blue for down-regulation, red for up-regulation, and black for non-significant genes. Upregulated and downregulated genes are annotated on the plot.* ***(b)*** *Venn diagram illustrates the number of DEGs detected in the Smart-seq2 and 10x Genomics scRNA-seq datasets. This includes genes with average log2 fold change ≥ |0.5| and adjusted p-value < 0.05 in both datasets. The intersection represents DEGs shared by both datasets. All upregulated and downregulated genes are listed adjacent to the Venn diagram.*

3.4. Results

Table 15. Dysregulated genes in the AZD1-resistant PDO observed in Smart-seq2 and 10x datasets

Genes	Locus	avg_log2FC	p_val_adj	Protein function
*C6orf15**	6p21.33	2.226	1.82E-07	chromosome 6 open reading frame 15 [Source:HGNC Symbol;Acc:HGNC:13927]
*CXCL14**	5q31.1	2.026	1.07E-07	C-X-C motif chemokine ligand 14 [Source:HGNC Symbol;Acc:HGNC:10640]
*S100A3**	1q21.3	1.860	1.24E-06	S100 calcium binding protein A3 [Source:HGNC Symbol;Acc:HGNC:10493]
*CD36**	7q21.11	1.464	1.17E-07	CD36 molecule [Source:HGNC Symbol;Acc:HGNC:1663]
*NEAT1**	11q13.1	1.339	2.72E-07	nuclear paraspeckle assembly transcript 1 [Source:HGNC Symbol;Acc:HGNC:30815]
*ECM1**	1q21.2	1.231	1.38E-04	extracellular matrix protein 1 [Source:HGNC Symbol;Acc:HGNC:3153]
*TACSTD2**	1p32.1	1.147	3.61E-03	tumor associated calcium signal transducer 2 [Source:HGNC Symbol;Acc:HGNC:11530]
*TGFA**	2p13.3	0.900	2.62E-03	transforming growth factor alpha [Source:HGNC Symbol;Acc:HGNC:11765]
*PDK3**	Xp22.11	0.900	1.37E-03	pyruvate dehydrogenase kinase 3 [Source:HGNC Symbol;Acc:HGNC:8811]
*S100A4**	1q21.3	0.885	2.56E-05	S100 calcium binding protein A4 [Source:HGNC Symbol;Acc:HGNC:10494]
*FABP5**	8q21.13	0.872	9.01E-03	fatty acid binding protein 5 [Source:HGNC Symbol;Acc:HGNC:3560]
*TMSB4X**	Xp22.2	0.803	2.69E-08	thymosin beta 4 X-linked [Source:HGNC Symbol;Acc:HGNC:11881]
*ENTPD1**	10q24.1	0.730	2.97E-03	ectonucleoside triphosphate diphosphohydrolase 1 [Source:HGNC Symbol;Acc:HGNC:3363]
*BSG**	19p13.3	0.708	1.25E-07	basigin (Ok blood group) [Source:HGNC Symbol;Acc:HGNC:1116]
*EPCAM**	2p21	0.607	5.04E-06	epithelial cell adhesion molecule [Source:HGNC Symbol;Acc:HGNC:11529]
*PCYOX1**	2p13.3	0.600	6.85E-04	prenylcysteine oxidase 1 [Source:HGNC Symbol;Acc:HGNC:20588]
*S100A6**	1q21.3	0.530	1.30E-02	S100 calcium binding protein A6 [Source:HGNC Symbol;Acc:HGNC:10496]
*RPL14**	3p22.1	-0.506	3.09E-06	ribosomal protein L14 [Source:HGNC Symbol;Acc:HGNC:10305]
IGFBP2	2q35	-0.546	1.01E-05	insulin like growth factor binding protein 2 [Source:HGNC Symbol;Acc:HGNC:5471]
PTMA	2q37.1	-0.579	3.02E-07	prothymosin alpha [Source:HGNC Symbol;Acc:HGNC:9623]
EIF3F	11p15.4	-0.584	3.70E-04	eukaryotic translation initiation factor 3 subunit F [Source:HGNC Symbol;Acc:HGNC:3275]
IL1R2	2q11.2	-0.589	1.61E-02	interleukin 1 receptor type 2 [Source:HGNC Symbol;Acc:HGNC:5994]
*RPS14**	5q33.1	-0.661	2.63E-12	ribosomal protein S14 [Source:HGNC Symbol;Acc:HGNC:10387]
*TESC**	12q24.22	-0.674	1.47E-03	tescalcin [Source:HGNC Symbol;Acc:HGNC:26065]
*RACK1**	5q35.3	-0.888	1.02E-17	receptor for activated C kinase 1 [Source:HGNC Symbol;Acc:HGNC:4399]

*Genes differentially expressed in the Smart-seq2 and 10x Genomics scRNA-seq datasets.

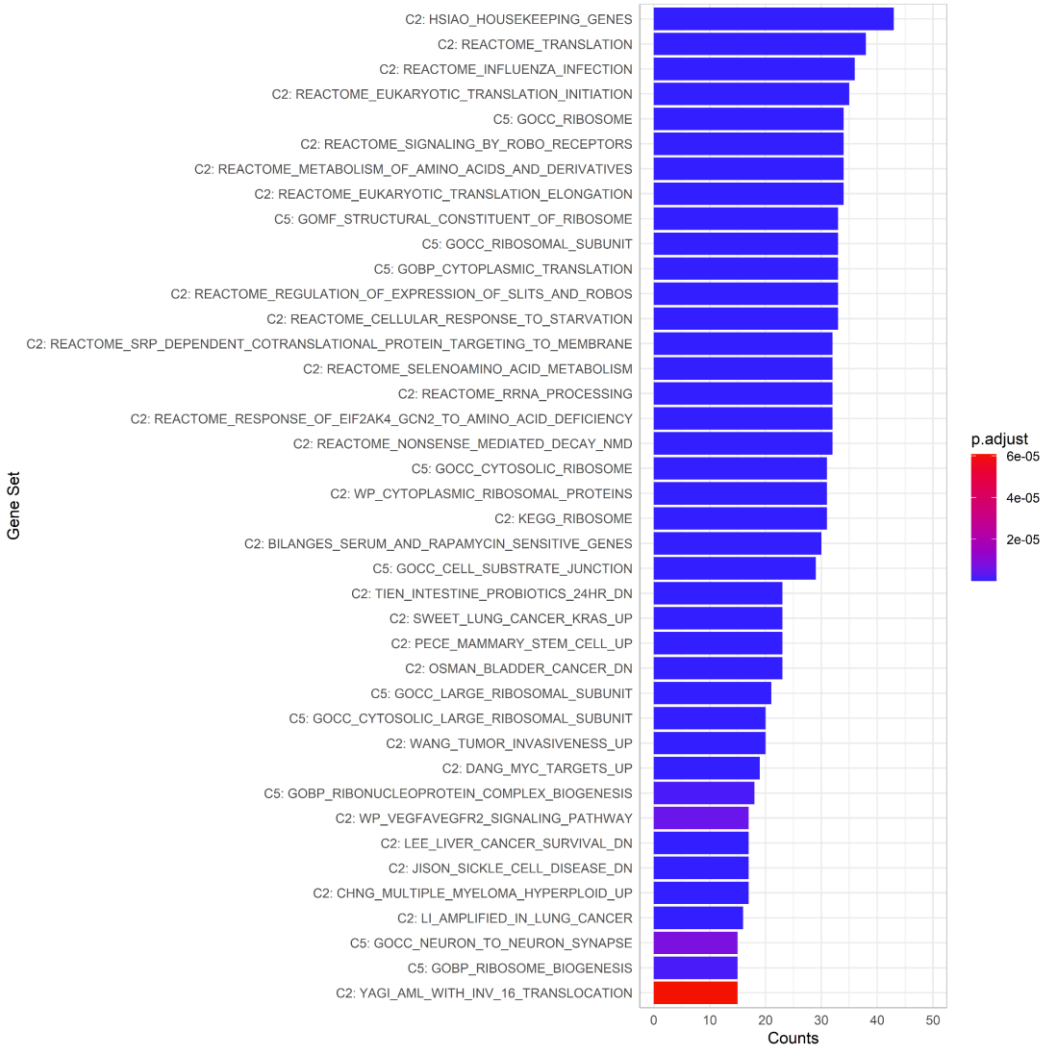


Figure 3.21. Over-representation analysis of differentially expressed genes in the AZD1-resistant organoid.

Top 40 gene sets over-represented in MK1-resistant mCRC organoid. Bar length indicates the count of DEGs overlapping each gene set. Colour gradient represents the level of statistical significance, with a threshold set at an adjusted p-value < 0.05.

3.5 Discussion

In this chapter, scRNA-seq was employed to explore various aspects of mCRC patient-derived organoids (PDOs) resistant to the AKT inhibitors MK-2206 and AZD5363. By employing a multifaceted approach, detailed information about the cellular and molecular landscape of these organoids was uncovered, providing insights into the cellular mechanisms and adaptations contributing to drug resistance.

3.5.1 Differential cluster abundance

The chapter begins with an evaluation of the quality of raw scRNA-seq data obtained via the Smart-seq2 arm of the G&T-seq protocol (129, 132). This initial quality control step identified a bacterial contaminant that affected most libraries. While some libraries were lost, the Smart-seq2 processing pipeline ensured that only high-quality libraries were selected for downstream analyses. Post-processing, Seurat (172-174) served as the primary bioinformatics tool employed for the analysis of single-cell transcriptomes. Within this analysis, further quality control of single-cell was performed, leaving a total of 361 cells: 87 from the untreated Parental PDO, 87 from the MK1-resistant, and 140 from the AZD1-resistant PDO.

A typical step in scRNA-seq analysis involves regressing out biological covariates, the most common being cell cycle effects (261). This process adjusts for the variability introduced by the cell cycle to minimise its influence on the analysis (261). Although this adjustment can help distinguish gene expression differences unrelated to the cell cycle, it poses challenges for result interpretation and carries the risk of overcorrection, potentially eliminating relevant biological signals (261). To preserve the biological signals from proliferating cells and maintain the cellular diversity characteristic of cultured organoids (237), this study computed cell cycle scores of mCRC cells without removing this information.

Four distinct clusters were identified in mCRC PDOs at a resolution of 0.5 (Figure 3.9A). Clusters 0 through 3 generally showed an equal distribution of cells from both the first and second batches of PDO plates sequenced. However, a significant proportion of cells in Cluster 3 were from the second batch, suggesting a potential batch effect in this cluster despite efforts to mitigate technical variability through integrated analysis. Additionally, while cells from all cell cycle phases were equally represented across the three PDOs, the clustering was influenced by cell cycle variation, with certain phases predominating in specific clusters (e.g., G1 phase in Clusters 0 and 2, S and G2/M phases in Cluster 1).

3.5. Discussion

Changes in cluster cell abundances were observed when comparing cell counts between the Parental and AKTi-resistant PDOs. For instance, Cluster 0 and Cluster 1 showed a slight reduction in cell counts across both resistant lines, while Cluster 2 experienced an increase, particularly in the AZD1-resistant organoid. However, bacterial contamination, which predominantly affected the MK1-resistant PDO, introduced uncertainty regarding the significance of the observed differences in cell abundances.

The strength of Smart-seq2 lies in its ability to generate full-length cDNA (151), which provides several advantages, including the detection of low-expressed genes. However, the inherent nature of the protocol presents certain challenges. Primarily, it requires cells to be sorted into wells before processing. Although this allows the selection of specific cell populations based on surface markers, it limits throughput compared to droplet-based scRNA-seq approaches such as those provided by 10x Genomics (10x Genomics, Pleasanton, CA, USA) (213).

The single-cell RNA-seq techniques developed by 10x Genomics can address the sampling limitations associated with Smart-seq2. Unlike Smart-seq2, these methods do not necessitate prior cell sorting, maintaining an unbiased representation of the cell population from the original sample. Additionally, 10x Genomics single-cell protocols offer the flexibility to adjust the input cell number for the Chromium instrument, targeting the recovery of a specific cell count. This feature enables the sequencing of multiple samples with a consistent number of cells, facilitating the comparison of cluster abundances across various experimental conditions.

To validate the findings from the Smart-seq2 clustering analysis, the 10x scRNA-seq datasets from the three PDOs were utilised, with Smart-seq2 cluster labels projected onto the 10x scRNA-seq dataset. The label transfer method, as implemented in Seurat (172-174), assigns cell labels—those representing cluster identities in this context—to queried cells in the 10x dataset by comparing their gene expression profiles with reference data from Smart-seq2, rather than assigning labels to predefined clusters. This cell-by-cell analysis approach allows for a more precise and tailored assignment of labels based on the unique features of each cell.

Analysis of cell abundances from the 10x perspective showed similarities with Smart-seq2 findings, particularly for Cluster 0 and Cluster 2. The slight decrease observed in Cluster 0 among AKTi-resistant cells indicates a potentially stable cell population that exhibited minimal reduction during treatment. Conversely, the significant increase in Cluster 1 and Cluster 2 in MK1- and AZD1-resistant PDOs, respectively, indicates that these inhibitors may have different transcriptional effects, likely due to their distinct mechanism of action and affinity for AKT1/2/3.

3.5. Discussion

The exclusive detection of Cluster 3 cells in MK1-resistant PDOs within the 10x dataset raises questions about whether its presence in Smart-seq2 data was due to technical variability or a limitation of the 10x methodology in identifying rare cell populations. Further statistical analysis is necessary to confirm these changes in cell abundance and understand the biological implications.

3.5.2 Cluster-level DGE

A dual approach to differential gene expression (DGE) analysis at the cluster level was employed to characterise the clusters identified in mCRC PDOs. This involved using Smart-seq2 data, and 10x data onto which Smart-seq2 cluster projections had been applied to identify shared differentially expressed genes (DEGs).

Cluster 0 was characterised by elevated expression of ribosomal genes and related processes, such as protein synthesis and ribosomal biosynthesis, alongside a significant reduction in cell-cycle-related genes. Coupled with the predominance of cells in the G1 phase and the lowest expression of AKT1/2/3 in this cluster (Figure 3.15), these findings suggest a limited growth potential in Cluster 0 within mCRC PDOs and a reduced dependence on the PI3K/AKT/mTOR pathway.

Furthermore, Cluster 0 exhibited substantial downregulation of genes associated with cell adhesion and those encoding proteins interacting with the extracellular matrix (ECM), including *CD36*, *ECM1*, *FNDC3A*, and *ITGA1* (204, 205, 207, 224). Alterations in these genes are known to enhance epithelial cell motility, a precursor to metastasis through mechanisms such as epithelial to mesenchymal transition (224, 262-265). *TACSTD2*, also known as *TROP-2* (208), showed the most significant downregulation in this cluster. Overexpression of this cell surface glycoprotein is common in various epithelial tumours, including CRC, and it has been identified as a potential target for cancer therapy (266-268). Despite the downregulation of genes typically associated with cancer aggressiveness, the invasive characteristics of Cluster 0 in mCRC PDO cells seem inherent, considering the origin of these cells and the consistent presence of Cluster 0 across both control and resistant mCRC PDOs. This observation suggests a sustained invasive potential despite the reduced expression of genes linked to aggressiveness.

Cluster 1 was characterised by the exclusive presence of S and G2/M phase-related genes, along with an upregulation of E2F transcription factor targets. The upregulation of these targets implicates various processes critical to the aggressive phenotype of cancer stem cells (CSCs), which are influenced by E2F transcription factors. These processes include the growth and division of CSCs, maintenance and acquisition of self-renewal capabilities, invasion and metastatic progression, and resistance to chemotherapy and radiotherapy (269). Additionally, the downregulation of genes related to epithelial cell differentiation suggests the presence of mechanisms that support stemness, aligning with the notion that resistant cells possess inherent stem-like qualities enabling them to withstand conventional cancer treatments (270). These characteristics, combined with the role of CSCs in initiating and driving disease progression, recurrence, and metastasis in CRCs (270), imply that Cluster 1 consists of highly

proliferative and undifferentiated resistant cells. This is underscored by the significant increase in Cluster 1 cells in MK1-resistant PDOs compared to control and AZD1-resistant organoids, as observed in the 10x scRNA-seq data.

The aggressive nature of Cluster 1 is further evidenced by the upregulation of genes such as *CCNA2*, *CCNB1*, *EIF4EBP1*, *EZH2*, *MELK*, *PBK*, *RACGAP1*, and *ZWINT*, which are associated with tumour progression, resistance to chemotherapy or radiation, and poor prognosis in mCRCs (271-277). Notably, MELK overexpression has been implicated in several cancer-related mechanisms, including the maintenance of CSCs, promotion of anchorage-independent cell growth, and modulation of reactive oxygen species (ROS) signalling (275, 278, 279).

Similar to Cluster 0, Cluster 2 predominantly comprised cells in the G1 phase, as indicated by the downregulation of S and G2/M phase-related genes and gene sets related to cell cycle progression. This, along with the significant downregulation of ribosomal biogenesis and protein synthesis processes, suggests a decrease in cell proliferation within this cluster. Interestingly, when determining the most appropriate resolution for clustering, Cluster 2 was the only cluster that remained undivided with increasing resolution, indicating a very distinct gene expression profile. This may be attributed to the upregulation of genes and gene sets involved in cell-to-cell communication (*GJA1*), migration and tissue remodelling (*CEMIP*, *ECM1*, *LGALS1*) (229). These findings highlight the migratory nature of this specific cluster.

Epithelial migration is a complex, energy-demanding process characterised by substantial changes in cell shape, cytoskeletal reorganisation, and the formation and disassembly of cell-cell and cell-matrix adhesions (280, 281). Correspondingly, genes associated with energy and nutrient metabolism, such as *CD36*, *GLCE*, and *SLC7A8* (229), were upregulated in Cluster 2 likely to fulfil the high energy demands of these processes.

Given their aggressive gene expression profiles and their increased presence observed in the Smart-seq2 data mapped onto the 10x dataset (Figure 3.11), Cluster 1 and Cluster 2 likely represent cells with drug resistant phenotypes that expanded in the MK1-resistant and AZD1-resistant organoids, respectively. On the other hand, the ambiguity associated with Cluster 3, especially the absence of shared DEGs between the Smart-seq2 and 10x datasets, hindered a conclusive DGE analysis. Only *TACSTD2* and *CFTR* were upregulated above the set thresholds in this cluster. *CFTR* is particularly interesting because a biallelic mutation in this gene is a well-established cause of cystic fibrosis (CF) (282). However, a lesser-known aspect of this gene is the increased risk of CRC in CF patients, which has been attributed to the role of *CFTR* as a tumour suppressor (282).

3.5. Discussion

Several other publications have recognised the power of scRNA-seq to identify clusters exhibiting drug-resistant phenotypes within heterogeneous cell populations in CRC PDOs. Chen et al. explored the mechanisms underlying oxaliplatin resistance in CRC, aiming to identify potential therapeutic targets to overcome it (283). They established two PDO models from CRC biopsies: one from a patient who had received neoadjuvant chemoradiotherapy (NACR) with oxaliplatin prior to sample collection and another from a treatment-naïve patient. The researchers characterised these PDOs by examining their morphology, histology, and drug sensitivity to oxaliplatin and 5-fluorouracil (5-FU). Additionally, they performed 10x Genomics scRNA-seq to analyse the gene expression profiles of individual cells within the PDO models, enabling the identification of differentially expressed genes, signalling pathways, and cell clusters associated with oxaliplatin resistance.

Although the two CRC PDO models retained the characteristics of their original tumours—in terms of morphology, histology, and drug sensitivity—, clear differences were observed between the control and NACR-treated organoids (283). Firstly, no effect of oxaliplatin was observed in the NACR-treated PDOs. While these oxaliplatin-resistant PDOs responded to 5-FU, the IC_{50} for this drug was approximately 116.00 μM, compared to the much lower IC_{50} of 7.6 μM and 12.38 μM for oxaliplatin and 5-FU, respectively, in the treatment-naïve CRC organoids.

Significant cellular heterogeneity within the PDOs was revealed using scRNA-Seq, particularly in the oxaliplatin-resistant models. Cells from the oxaliplatin-resistant organoids were grouped into five clusters (clusters 0, 4, 5, 6, and 7), while the treatment-naïve sample showed three clusters (clusters 1, 2, and 3) (283). Cluster 4 in the oxaliplatin-resistant organoids was particularly notable for several reasons. Firstly, this cluster exhibited a high proliferation rate, as indicated by cell cycle analysis, which showed the majority of cells in the G2/M phase. This was surprising as the organoids were derived from a tumour resistant to oxaliplatin, which typically induces cell death or dormancy (37). The presence of proliferating cells in a resistant tumour suggests a potential mechanism of treatment evasion. Secondly, Cluster 4 exhibited a distinct gene expression profile compared to other clusters (Cluster 0, 5, 6, and 7) in the resistant organoids, hinting at heterogeneity between cells. Indeed, Cluster 0, 4, and 6 were associated with nuclear division, while the remaining clusters in the oxaliplatin-resistant PDO were mainly related to DNA replication. Conversely, Clusters 1, 2, and 3 in the treatment-naïve organoids exhibited more homogenous gene expression patterns, mainly related to RNA catabolic processes and translational initiation, suggesting a less diverse cell population compared to the resistant organoids.

Moreover, KEGG pathway analysis revealed that Cluster 4 was enriched in pathways related to platinum drug resistance, aligning with the patient's treatment history (283). Cluster 4's enrichment in platinum drug resistance pathways further reinforces the hypothesis that these cells may have developed specific adaptations to evade the cytotoxic effects of oxaliplatin, and several potential drug resistance-related genes (*STMN1*, *VEGFA*, *NDRG1*) and transcription factors (*E2F1*, *BRCA1*, *MYBL2*, *CDX2*, *CDX1*) were identified in the resistant PDOs.

In another study, Chen et al. derived CRC PDOs from early-stage tumours from treatment-naïve patients (284). These organoids were later treated with oxaliplatin, and their transcriptomes were profiled by Drop-seq scRNA-seq to investigate the effect of oxaliplatin on the cellular diversity of the CRC PDOs. Researchers identified approximately 30 clusters with distinct heterogeneity in gene expression patterns and pathway signatures (284). Common pathways included those related to ribosomes, protein targeting to membranes, and mRNA catabolic processes. Unique pathways included glycolysis/gluconeogenesis, fructose and mannose metabolism, and non-homologous end-joining.

Analysis of cell percentage changes before and after oxaliplatin treatment showed altered diversity of clusters (284). Eight subtypes were completely depleted after treatment, while four new clusters emerged. Moreover, two major clusters in the control PDOs were significantly reduced, while another became dominant after treatment. These clusters were categorised into four groups based on their response to oxaliplatin: drug-induced, drug-insensitive, drug-sensitive, and drug-ultrasensitive.

In conclusion, the identification of distinct cell clusters with different gene expression profiles in CRC PDOs observed in this PhD project, supported by the studies mentioned above, suggests that not all tumour cells are equally sensitive to chemotherapy. Indeed, these studies show that understanding the specific resistance mechanisms in each cluster can guide the development of combination therapies that target multiple cluster-related pathways or molecules simultaneously to effectively eradicate tumours.

Clustering analysis is undeniably powerful for dissecting the cellular and molecular landscape of tumours, providing insights into their heterogeneity and potential therapeutic targets. However, it is important to acknowledge that clustering results can be influenced by various factors, including the choice of clustering algorithm, batch effects in the data, the inherent assumption that cells within a cluster share biological similarity, and the specific data preprocessing steps taken.

3.5. Discussion

In the scRNA-seq analysis, the quality of the clusters was evaluated based on several criteria. The clusters displayed clear separation and distinctiveness in the UMAP plots (Figure 3.9A), with minimal "leakage" between clusters in both the Smart-seq2 data and the same data projected onto the 10x dataset (Figure 3.10B). This suggests good resolution of different cell populations. All clusters were identified in both control and AKTi-resistant PDOs (Figure 3.9B) and were consistent across the two sequencing batches (Figure 3.9C), except for Cluster 3, indicating reproducibility and robustness. Aside from Cluster 3, all clusters exhibited distinct expression patterns, and cells within each cluster were highly similar in terms of gene expression (Figure 3.12A). This high intracluster homogeneity further validated the effectiveness of the clustering algorithm in grouping similar cells together.

Nevertheless, it is clear that the clustering of mCRC organoids was heavily influenced by differences in cell cycle phases. This likely occurred because cell cycle effects were not regressed out, and suggests that the differences between clusters may have been driven more by variations in cell proliferation rather than true resistance mechanisms. Regressing out cell cycle effects could have helped isolate the molecular changes directly associated with drug resistance, potentially providing a clearer separation of clusters based on resistance-related changes rather than differences driven by proliferation status. However, the decision not to regress out this covariate was based on the hypothesis that mechanisms of resistance do not solely relate to direct genetic changes but also involve broader cellular processes such as altered cell cycle dynamics and differentiation states.

Knowing that in preparation for bulk and single-cell sequencing experiments AKTi-resistant organoids were cultured in growth media without the AKT inhibitors provides important insights into how the results might be interpreted. Since the resistant organoids were cultured in regular growth media, the cells likely resumed normal proliferation patterns. This could have increased the impact of cell cycle phases on clustering, as cells were no longer under the selective pressure of the AKT inhibitors and could cycle normally.

In the context of this PhD project, where differentiation and proliferation were likely intertwined in AKTi-resistant organoids, considering the biological relevance of cell cycle effects was crucial. If non-cell cycle genes show changes in expression correlated with cell cycle phases, it may indicate that these changes are due to a different biological process that is linked to the cell cycle rather than the cell cycle itself (285). Thus, regressing out cell cycle effects in our analysis without considering this might have mistakenly attribute these changes solely to the cell cycle, missing the underlying response mechanism.

For instance, drug resistance often emerges due to the inactivation of genes regulating cell proliferation, such as *TP53* and *CDKN2A* (286). There is also evidence that cancer stem cells (CSCs), which may become enriched after chemotherapy (287-290), can arise from either adult stem cells, adult progenitor cells that have undergone mutation, or from differentiated cells/cancer cells that have acquired stem-like properties through dedifferentiation (290-292). CSCs are also able to induce cell cycle arrest (quiescent state), supporting their ability to become resistant to chemo- and radiotherapy (290, 293-295). This finding is particularly intriguing as Cluster 1 cells, which increased in the MK1-resistant organoid, exhibited a stem-like gene expression profile, along with the expression of EMT-related genes in both organoids. EMT has been associated with the generation of cells with stem-like properties (295-297). Therefore, the presence of these genes suggests that the observed resistance mechanisms may involve changes in both proliferation and differentiation states, including the possible enrichment of CSCs.

The results of the cluster-level differential gene expression analysis indicated that while cell cycle phase influenced clustering, it did not solely dictate cluster formation. Critical insights into several resistance mechanisms—such as the upregulation of genes related to chromatin modification and gene silencing, detoxification, DNA repair, cell survival and metabolism, and pathways of energy and nutrient metabolism—were still obtainable when comparing gene expression between the clusters without regressing out these effects. Therefore, by not regressing out cell cycle effects, the full spectrum of changes associated with resistance, including those related to cell cycle regulation and differentiation, was captured, providing insights into the mechanisms driving resistance to AKT inhibition in PDOs.

Regardless of the approach used for this particular research project, project-specific aims as well as data-driven evidence should guide the decision on whether regression is necessary. Future analyses could refine this approach by:

1. Performing a comparative analysis with and without regressing out cell cycle effects to evaluate the impact on clustering and the identification of resistance-related genes. This approach can help determine if cell cycle regression clarifies or obscures the biological signals of interest. If certain cell cycle regulators are consistently upregulated even after regression, this might indicate that altered cell cycle control is part of the resistance mechanism.
2. Integrating cell cycle information into the current approach by subclustering the dataset based on the identified cell cycle-driven clusters. Following this, differential expression analysis could be performed between AKTi-resistant and parental PDO cells within each

cluster. This method would control for cell cycle effects and focus on the underlying biological differences contributing to resistance.

3. Excluding cell cycle genes after identifying variable features, as recommended in bioinformatics forums (285), which ensures that these genes do not influence downstream analysis such as clustering, highlighting other important biological signals.

By considering these approaches, future analyses can better control for cell cycle effects and more accurately identify the mechanisms underlying cancer evolutionary processes.

3.5.3 Cell type classification

In the next phase of the scRNA-seq analysis, the Human Gut Cell Atlas (HGCA) was employed to classify or label mCRC PDO cells from the Smart-seq2 dataset. The HGCA, a subset of the Human Cell Atlas (HCA), focuses on defining the cellular composition of the human gut. It includes cells from diverse intestinal regions and sample types—including fresh, frozen, and formalin-fixed tissues (118, 190, 234). The HGCA also encompasses cells from various physiological conditions and developmental stages (e.g., foetal and paediatric), making it the most extensive collection of annotated intestinal cell types essential for gut function.

To examine cell type abundances across different platforms, the Smart-seq2 dataset, labelled with HGCA annotations, was subsequently employed to label the 10x scRNA-seq dataset. The primary cells identified in mCRC PDOs across both datasets were CLDN10+ cells, followed by transient amplifying (TA) and Paneth cells, with a minority of cells being enterocytes. The non-specific nature of proliferative markers typically associated with TA cells prompted further investigation into CLDN10+ cells, which are less well-documented in the literature compared to the other intestinal cell types identified.

Claudins, including claudin-10, are critical components of tight junctions in epithelial cells, playing a key role in maintaining permeability barriers and establishing cell polarity (298). Claudin-10 is localised throughout the entire crypt in murine models (299). According to the HGCA, CLDN10+ cells, which were formally identified in foetal samples, may represent pancreatic progenitors, as evidenced by their expression of *CPA1*, *DLK1*, *PDX1*, *RBPJ*, and *SOX9*—genes associated with pancreatic development (190). However, this similarity in gene expression does not definitively classify CLDN10+ cells as pancreatic in nature but rather indicates a potential similarity in developmental or functional state. In addition, the elevated expression of stem and progenitor cell markers in CLDN10+ cells across all clusters and cell cycle phases in mCRC PDOs (Figure 3.17B) suggests that these cells span a range of differentiation stages. Within this spectrum, CLDN10+ may be more differentiated than TA cells but less mature than specialised cells like Paneth cells or enterocytes.

While the abundance of CLDN10+ and TA cells remained somewhat stable between Parental and AZD1-resistant PDOs, an increase was observed in the MK1-resistant PDO. This increase in undifferentiated cells in the MK1-resistant organoid suggests a cellular adaptation to AKT inhibition, leading to a shift towards a more aggressive cancer phenotype characterised by enhanced cell proliferation despite AKT signalling inhibition. On the other hand, the culture medium for the three mCRC PDOs was enriched for LGR5+ stem cell proliferation with additives such as Wnt-3a, a mitogen that activates the Wnt/β-catenin signalling (203). Even with culture

conditions favouring stemness, the presence of a Wnt gradient within the organoid cultures still facilitated the differentiation of some undifferentiated cells into specialised cell types such as enterocytes, Paneth cells, goblet cells, and EC TAC1+ cells in the mCRC PDOs, albeit in limited numbers.

3.5.4 DGE analysis of AKTi-resistant mCRC PDOs

In the concluding phase of the transcriptome analysis, differential gene expression analysis was conducted to compare each AKTi-resistant organoid with the untreated Parental control. To further validate the Smart-seq2 findings, the 10x dataset was employed; however, the insights derived from the 10x dataset did not rely on any projections from the Smart-seq2 data, hence reflecting the intrinsic expression characteristics of the mCRC PDOs.

Exclusively found in the Smart-seq2 dataset, *MUC21* appeared as the highest upregulated gene. A high expression of *MUC21* correlated with a decreased cell-cell and cell-matrix adhesion in a particular variant of lung adenocarcinoma characterised by a scattered arrangement of cells in alveolar spaces (239). Mucins have become important serum biomarkers for monitoring cancer progression, with *MUC16* (*CA125*), for example, being prevalent in over 80% of ovarian cancers but rare in normal tissues (300). Alongside their utility in diagnostics, mucins are emerging as targets for cancer immunotherapy, with ongoing research into vaccines and antibodies aimed at their unique glycoprotein structures for more precise treatment strategies (301).

Also upregulated in the MK1-resistant PDO were genes encoding cell surface proteins frequently overexpressed in various cancers, including CRC, and associated with clinically aggressive tumours with poor prognosis. Among these, *TACSTD2* (240) and *BSG* (241) were noted for their upregulation in both AKTi-resistant PDOs. Basigin (*BSG*) in particular, is a transmembrane glycoprotein implicated in promoting the secretion of extracellular matrix (ECM)-degrading matrix metalloproteinases (MMPs), cytokines, and angiogenic factors, making it a significant prognostic marker in cancer (241). Previous studies have highlighted the effectiveness of cancer treatments targeting basigin with monoclonal antibodies (302).

The collective upregulation of *MUC21*, *TACSTD2*, *and BSG*, each known to increase metastatic potential in colorectal adenocarcinomas (303-305), suggests a synergistic effect that not only enhances cell motility and drives morphological changes, but also reshapes the tumour microenvironment. This synergistic action could lead to more efficient metastases by altering interactions with other cells and their response to external stimuli, ultimately enhancing the tumour's ability to evade or resist therapeutic interventions, such as the MK-2206 AKT inhibitor in this case.

The most interesting discovery within the MK1-resistant PDO was the upregulation of genes involved in glycolysis, the pentose phosphate pathway (PPP) and related biosynthetic pathways, including *SLC2A1/GLUT1, SLC6A14*, *ENO1 and PGD* (the last two were common to both scRNA-seq datasets). *GLUT1* encodes a major glucose transporter dysregulated in various

cancers (306). In ovarian cancer cell lines, the constitutive activation of a mutant PI3KC1-AKT pathway triggered the translocation of *GLUT1* from the Golgi area to the plasma membrane, where it became permanently expressed, thereby enhancing glucose uptake (306). On the other hand, *SLC6A14* is an amino acid transporter scarcely detected in normal colonic tissue but significantly overexpressed in poorly differentiated CRCs, where its upregulation promoted tumour progression by activating the AKT/mTOR signalling pathway (307).

On the other hand, *ENO1* catalyses the conversion of 2-phosphoglycerate (2PG) to phosphoenolpyruvate (PEP) in the glycolytic pathway, ultimately leading to the production of pyruvate, the primary end product of glycolysis (308). The upregulation of two SLC membrane transporters and *ENO1* in the 3994-117/F16 mCRC PDOs, which had a missense mutation in *ABCB1* (also known as multidrug resistance protein 1, *MDR1*), which is known to mediate drug efflux, suggests a role of membrane transporters in mediating drug resistance.

Also upregulated was *PDK3*, which inhibits the pyruvate dehydrogenase (PDH) complex responsible for converting pyruvate to acetyl-CoA before it enters the Krebs cycle (243). By inhibiting PDH, PDK3 effectively blocks the conversion of pyruvate to acetyl-CoA, decreasing the flow of carbon from glycolysis into the Krebs cycle and leading to an increased conversion of pyruvate into lactate (243). While glycolysis generates less ATP per molecule of glucose than oxidative phosphorylation, it offers several advantages: ATP is produced more quickly, and unlike oxidative phosphorylation, glycolysis can occur under hypoxic conditions (243). Additionally, glycolysis provides intermediates for the PPP, which are essential for the synthesis of nucleotides, amino acids, and lipids, all crucial for tumour metabolism.

Conversely, *PGD* encodes 6-phosphogluconate dehydrogenase (6PGD), a key enzyme essential in the PPP (229). 6PGD specifically acts on 6-phosphogluconate, derived from glucose-6-phosphate (G6P), the first intermediate product of glycolysis (309). The function of 6PGD in the PPP is crucial for producing ribulose-5-phosphate, which is then converted into ribose-5-phosphate (R5P), essential for nucleotide synthesis in nucleic acids (309). Additionally, the activity of 6PGD in the PPP contributes to the production of NADPH, a critical cofactor in reductive biosynthetic reactions such as fatty acid synthesis (309). NADPH also plays a crucial role in detoxification, scavenging reactive oxygen and nitrogen species (ROS and RNS, respectively) and free radicals, thereby reducing intracellular oxidative stress and inhibiting the activation of apoptotic and necrotic signalling pathways (309-311). The upregulation of detoxification enzymes like *GPX1*, *GSTP1* and *PARK7* (229) further indicates a mechanism for protection against oxidative damage.

3.5. Discussion

Although the genes predominantly upregulated in the MK1-resistant PDO are related to glycolysis, there was also moderate upregulation of enzymes encoding components of the electron transport chain, such as Complex I and Complex III (e.g., *MT-ND1*, *MT-ND4*, *MT-ND2*, *MT-ND3*, *MT-ND5*, and *MT-ND6*) (229). This expression pattern suggests active electron transport chain activity, indicative of ongoing ATP synthesis through oxidative phosphorylation alongside glycolysis. The simultaneous expression of these metabolic pathways in MK1-resistant cells underscores their metabolic flexibility.

DEFA5, a marker for Paneth cells, was the most downregulated gene in MK1-resistant PDOs (250). An aberrant expression of *DEFA5* has been documented in CRC. For instance, Qiao et *al*. found a decreased expression of *DEFA5* at the protein level in colon adenocarcinomas compared to adjacent tissues (250). This trend was consistent across different colon cancer subtypes, with *DEFA5* showing greater downregulation in primary tumours than in normal mucosa. Experiments where CRC cell lines overexpressing *DEFA5* were subcutaneously injected into nude mice resulted in smaller tumour sizes, indicating a possible tumour suppressor function for *DEFA5* in CRC (250). Mechanistically, *DEFA5* was proposed to interact with the p85 protein subunit of the PI3K complex, attenuating downstream signalling leading to delayed cell growth and metastasis. This hypothesis is supported by observations that *DEFA5* overexpression leads to hypophosphorylation and inactivation of AKT, whereas silencing *DEFA5* increases AKT phosphorylation. Furthermore, activating AKT in colon cancer cells transfected with exogenous *DEFA5* reversed the growth-suppressive effects of *DEFA5* overexpression and enhanced their migratory potential (250). These findings strongly suggest that *DEFA5* interacts with the PI3K complex, inhibiting signal transduction pathways involved in cell proliferation and migration, reinforcing its role as a potential tumour suppressor in CRC.

While the MK1-resistant organoid exhibited dysregulation of several genes potentially contributing to acquired resistance to the MK-2206 inhibitor, the most interesting finding was the upregulation of genes involved in glycolysis and other biosynthetic pathways. This indicates a significant metabolic shift from oxidative phosphorylation to aerobic glycolysis—a hallmark of cancer cell metabolism known as the Warburg effect that is observed even in precancerous colorectal lesions (312, 313).

Incidentally, among the upregulated metabolic enzymes, *ENO1*, *PGD*, and *PARK7* are in close proximity on chromosome 1p36. A similar pattern of gene upregulation was observed with *CFL1* and *GSTP1*, both situated around 11q13, as well as with *TACSTD2* at 1p32.1 and *SLC2A1*/*GLUT1* at 1p34.2. The co-expression and physical proximity of these genes, which are implicated in metabolic processes and cellular morphological changes, suggest a coordinated

regulation. This coordination could result from chromosomal structural changes or shared regulatory elements, indicating a complex level of genetic regulation underlying these metabolic adaptations.

Based on the dysregulation of metabolic genes and the upregulation of ECM-remodelling components observed at the gene expression level in both AKTi-resistant PDOs derived from 3994-117/F-016, these organoids exhibit characteristics of both CMS3 (metabolic) and CMS4 (mesenchymal) CRC subtypes (15). This highlights the heterogeneity of these organoids as detected at single-cell resolution. Complementing these transcriptomic findings, Vlachogiannis *et al*., reported that the molecular landscape of the Parental mCRC organoid was characterised by a high number and diversity of somatic mutations, including amplifications and deletions in *TP53*, *KRAS*, *NRAS*, *SMAD4*, *AKT1*, and *MYC* (109) (**Figure 2.1**A). This profile, along with the absence of mutations in DNA mismatch repair (MMR) genes, suggests that this tumour likely belongs to the chromosomal instability-high (CIN-high) CRC subtype (9). Additional mutations in genes such as *ABCB1* (multidrug resistance) and *CDKN2A* (tumour suppressor), which have been linked to poor prognosis in CRCs (314, 315), may have also contributed to the tumour's behaviour and resistance mechanisms developed to counteract AKT inhibition hereby reported in this chapter.

The next chapter will explore the copy number profiles of MK1- and AZD1-resistant single-cell genomes to characterise the clonal composition of AKTi-resistant PDOs, aiming to further elucidate the genetic alterations associated with resistance mechanisms.

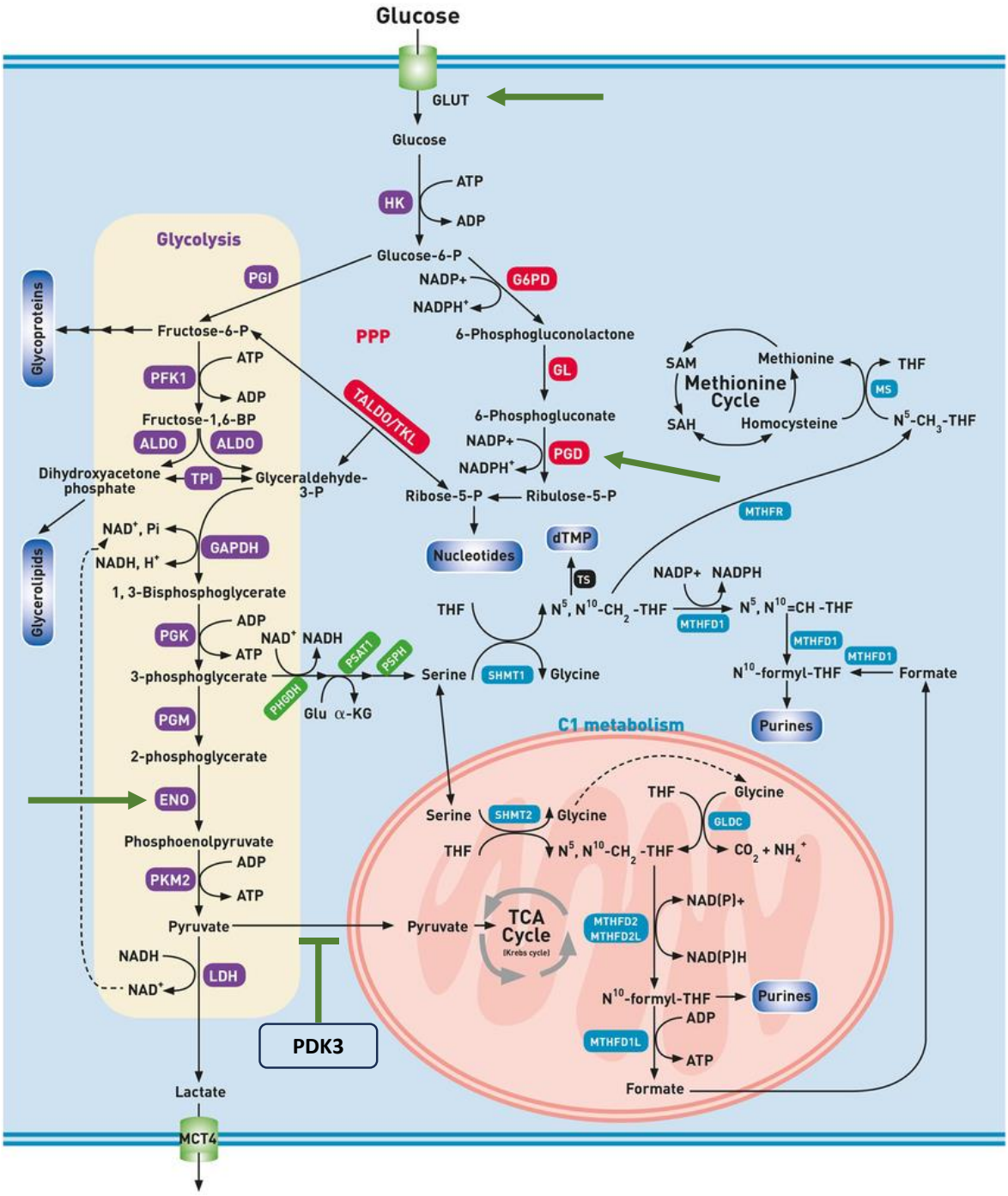


Figure 3.22. Glycolysis and related biosynthetic pathways.

Diagram depicts glycolysis in purple, the pentose phosphate pathway in red, and one-carbon metabolism (which includes the Krebs cycle) in blue. Green arrows highlight genes that were upregulated in the MK1-resistant PDO. PDK3 was included to indicate the step inhibited by this protein, i.e., the conversion of pyruvate to acetyl-CoA. Figure adapted from (316)

.

3.5. Discussion

In the AZD1-resistant PDO, *C6orf15* emerged as the most significantly upregulated gene, identified in both the Smart-seq2 and 10x datasets. The role of *C6orf15* in solid tumours is not yet fully understood. However, it is thought to play a role in various functions, such as extracellular matrix organisation and collagen V and fibronectin binding activities (229). A bioinformatics study by Xiong et al., utilising The Cancer Genome Atlas (TCGA) database—which included data from colorectal adenocarcinoma patients and paired tumour-normal samples—found that elevated mRNA expression of *C6orf15* in colon cancers was associated with shorter overall survival (OS) and progression-free intervals (PFI) compared to patients with lower gene expression (317). Immunohistochemistry staining of clinical samples further demonstrated a correlation between *C6orf15* protein expression and the depth of tumour invasion in colon cancer tissues.

In their research, Xiong *et al*. proposed that *C6orf15* may play a role in CRC progression by activating key pathways involved in extracellular matrix remodelling, including ECM-receptor interactions, as well as Hedgehog and Wnt signalling pathways (317). Furthermore, *C6orf15* appeared to suppress several immune-related pathways, such as those involved in antigen processing and presentation, natural killer cell-mediated cytotoxicity, and Immunoglobulin A (IgA) production, indicating a potential role in promoting immune evasion.

While the MK1-resistant PDO displayed an increased expression of metabolic enzymes, the AZD1-resistant PDO showed upregulation of genes implicated in CRC progression, chemoresistance and ECM-remodelling, e.g., *FNDC3A*, *NEAT1*, *ECM1*, *TACSTD2*, *TMSB4X and PPP1CB* (229), and genes involved in immune regulation, e.g., *ENTPD1/CD39* and *CXCL14* (the latter of which was also upregulated in the MK1-resistant PDO):

Fibronectin Type III Domain-Containing (FNDC) proteins, including Fndc3a, are involved in cell adhesion, migration, and proliferation (263). Although the impact of increased *FNDC3A* expression has not been defined, its significance has been highlighted. Shivakumar et *al*. analysed the expression of genes frequently mutated in CRC, such as *TP53*, *CCND1*, *EGFR*, *C-MYC*, and *FNDC3A* (318). They found that *FNDC3A*, located within the 13q chromosomal copy number variation (CNV) gain—a region associated with CRC (319)—was highly overexpressed in tissue samples from sporadic CRC (318).

Notably, an increased expression of *NEAT1* enhanced resistance to 5-FU in CRC cells by activating autophagy-related pathways by suppressing miR-34a, a small noncoding regulatory RNA with a tumour-suppressing role (320). On the other hand, the extracellular matrix protein 1 (*ECM1*) has been linked to drug resistance in CRC. Studies have shown that CRC patients

resistant to 5-fluorouracil (5-FU)-based chemotherapy, such as capecitabine, exhibited elevated *ECM1* expression, which was associated with shorter overall and disease-free survival rates compared to patients responsive to these treatments (265). Furthermore, knocking down *ECM1* in colonic cells resistant to 5-FU reduced their resistance to the drug, indicating a potential role of this gene in chemoresistance (265). Knockdown of *ECM1* also led to decreased phosphorylation of PI3K, AKT, and GSK3β kinases, whereas overexpression of *ECM1* had the opposite effect. These results indicate that *ECM1* influences the PI3K/AKT/GSK3β pathway, making *ECM1* a viable target for improving the efficacy of 5-FU in CRC treatment (265). This study is particularly interesting considering that the patient from whom the mCRC PDOs were derived had previously been treated with capecitabine, as well as the role of GSK3β in destabilising β-catenin in the Wnt/β-catenin signalling pathway.

PPP1CB is another gene involved in the Wnt/β-catenin signalling pathway, that was also upregulated in the AZD1-resistant PDO. The protein encoded by this gene is part of the PP1 subunit, a protein that inactivates AXIN in the destruction complex resulting in the dephosphorylation of β-catenin, and subsequent activation of Wnt-target genes (321). Thus, the upregulation of *ECM1* and *PPP1CB* may be indicative of overactive Wnt-signalling in this organoid.

Another upregulated gene involved in signalling pathways is TGFA, a ligand of the epidermal growth factor receptor (EGFR). Its overexpression has been shown to play a significant role in resistance to the anti-EGFR drug cetuximab by inducing interactions between EGFR and MET (322). This suggests overactive EGFR signalling, which may have contributed to the resistant phenotype in the AZD1-resistant organoid.

Conversely, *ENTPD1/CD39* is recognised for its role in creating an immunosuppressive tumour microenvironment (252). *CD39* hydrolyses the sequential conversion of ATP to AMP, a process particularly relevant under cellular stress conditions like hypoxia or exposure to anticancer therapies, which cause cells to release ATP into the surrounding environment. Elevated ATP levels in the tumour microenvironment (TME) act as a danger-associated molecular pattern (DAMP), triggering innate and adaptive immune responses against tumour cells. However, the enzymatic activity of CD39, along with CD73—which further metabolises AMP into adenosine—transforms the TME from a pro-inflammatory state to an immunosuppressive one, thereby aiding in tumour evasion of immune detection and response. Indeed, adenosine exerts regulatory effects on various immune cells. It suppresses effector immune cells like T cells, B cells, and NK cells, which are crucial for anti-tumour immunity while promoting regulatory

immune cells like regulatory T cells (Tregs) and myeloid-derived suppressor cells (MDSCs), known for their immunosuppressive functions (252).

In addition, *CXCL14*, which was upregulated in both organoids, plays a significant role in upregulating the expression of major histocompatibility complex class I (MHC-I) on tumour cells (253). The loss of *CXCL14* has been linked to impaired anti-tumour immune regulation and is associated with poor patient prognosis. A recent phase I clinical trial identified a set of genes known as the "Adenosine Gene Signature" (AdenoSig), which includes *CXCL1*, *CXCL2*, *CXCL3*, *CXCL5*, *CXCL6*, and *CXCL8* as biomarkers for monitoring the response to A2AR inhibitors like ciforadenant in patients with renal cell cancer (323). These genes, which correlated with adenosine expression levels in tumours, predicted patient responses to A2AR antagonist treatments. Although *CXCL14* is not included in the AdenoSig, this study highlights the importance of chemokines as potential markers for monitoring adenosine-regulated gene expression signatures in cancer.

Also observed in both mCRC PDOs was a decreased expression of *RACK1* and *IGFBP2*. The downregulation of these genes, both of which are involved in IGF-signalling suggests that the resistance to AKT inhibition likely did not arise from compensatory activation through IGFR-mediated signalling, a pathway known to trigger oncogenic processes akin to those driven by AKT/PI3K signalling (324-326).

Finally, the downregulation of ribosomal genes essential for RNA processing and translation was observed in both AKTi-resistant PDOs. Specifically, in the MK1-resistant PDO, this stands in stark contrast due to the concurrent overexpression of genes associated with glycolysis, the PPP, and protein-folding genes. Intuitively, one might expect an upregulation of ribosomal genes to accompany the enhanced activity of these metabolic pathways.

Although upregulation of components involved in ribosome biosynthesis and protein translation in colorectal cancers has been documented, linked to changes in Wnt, RAS/MAPK, and PI3K/AKT signalling pathways (327), the downregulation of these genes is not very well recorded in the literature. This decrease in protein synthesis efficiency might underlie the increased reliance on glycolysis and the PPP, as cells sought alternative or supplementary means to fulfil their energetic and biosynthetic needs. Additionally, given that the Parental PDO started the analysis with an unusually high ribosomal content (observed in the Smartseq2 and 10x scRNA-seq datasets), the observed downregulation in ribosomal-related genes could simply reflect a relative decrease compared to this initial state.

3.5. Discussion

Undoubtedly, single-cell RNA-seq offers unparalleled resolution for analysing cellular heterogeneity, enabling the identification of distinct cell types and states within complex tissues. Despite its significant advantages, scRNA-seq also presents many challenges. This chapter has illustrated the impact of cell cycle effects on clustering analysis. While regressing these effects can help clarify other biological signals, it requires careful consideration of the biological context to avoid obscuring relevant signals. Biological processes within an organism are interdependent, and thus, adjustments for one process might inadvertently conceal signals of another (261). This approach is particularly critical in organoid models, where the interplay of various cellular processes is essential to understanding the underlying biology.

To validate the Smart-seq2 findings, high-throughput scRNA-seq 10x data from matched mCRC PDO samples was employed. By mapping Smart-seq2 information onto the 10x dataset, the analysis benefitted from the higher sensitivity of Smart-seq2 and the high cell numbers provided by the 10x technology. However, this projection approach meant that unique insights from the 10x dataset might have been overlooked. An alternative strategy could have involved integrating both datasets. However, given the thesis's aim to evaluate the efficacy of the G&T-seq methodology in understanding colorectal cancer resistance, this projection strategy was chosen to avoid potential data overcorrection due to the sensitivity and resolution differences between the methods (Supplementary Figure 12A-D), which would have decreased the power of Smart-seq2.

In summary, the scRNA-seq analysis uncovered potential strategies that mCRC PDOs might employ to overcome AKT inhibition. The MK1-resistant PDO adapted by enhancing its metabolic requirements, whereas the resistance mechanisms of the AZD1-resistant PDO primarily focused on altering cell-cell and cell-matrix interactions and creating an immunosuppressive tumour microenvironment. The next chapter evaluates the matched DNA from these transcriptomes to determine if there are genetic causes behind these resistance mechanisms.

Chapter 4

scDNA copy number analysis of mCRC PDOs

Chapter disclosures

PicoPlex Gold amplification of G&T-seq genomic DNA, as well as the bioinformatics analysis of the scWGS data were performed by Silvia Ogbeide. Bulk WGS of mCRC PDOs and Sequenza DNA copy number analysis of bulk WGS data were performed by the Sottoriva group (Institute of Cancer Research (Sutton, UK). Bioinformatics analysis of raw bulk WGS data was performed by Silvia Ogbeide.

4.1 Introduction

Copy number alterations (CNA) are somatic changes that involve the gain or loss of DNA segments larger than 1 kb (58). CNAs play a significant role in genetic diversity and are critically involved in the development and progression of various diseases, including CRC. In CRC carcinogenesis, both early and advanced adenomas exhibit CNAs levels comparable to those found in carcinomas, indicating that CNAs play a role from the initiation to the progression of CRC. The early development of CRC is associated with the loss of chromosomes 17p, 18, and 22q and the gain of chromosomes 8q, 13q, and 20. Later stages of CNAs in primary CRCs include deletions at chromosomes 4p and 8p and gains at 7p and 17q (59). In contrast, metastatic CRCs exhibit an increase in CNAs compared to invasive CRCs, suggesting that CNAs contribute to the dissemination of malignant cells to distant sites (58, 59). Notably, established liver metastases are characterised by losses at 14q and 17q and gains at 1q, 9p, 11, 12p, 19, and X (59).

Oncogenes are often located within regions of DNA amplification, whereas tumour suppressor genes reside in areas of deletion (58). High-level amplifications or deletions are often required to cause oncogene activation or tumour suppressor inactivation. Consequently, CNAs can disrupt gene function, affect gene expression, and alter cellular pathways contributing to disease phenotypes. Thus, the study of CNAs can reveal patterns such as co-occurrence, where genes are amplified together, suggesting synergistic roles in cancer, or mutual exclusivity, i.e., situations where certain genes are rarely altered together (e.g., *KRAS* and *BRAF*), indicating that these genes function within the same pathway (33, 58). This is particularly important in the context of acquired drug resistance, where DNA copy number analysis can shed light on the evolutionary dynamics driving the transition from drug sensitivity to drug tolerance and highlight potential vulnerabilities within these evolved tumour populations that could be targeted therapeutically.

Building on the foundation laid in the previous chapter, which focused on gene expression profiling of single-cell transcriptomes, this chapter investigates genome-wide CNAs in AKTi-resistant mCRC PDOs using the matched single-cell genomes of these cells. By tracking changes in CNAs over time, i.e., from AKTi-sensitive to AKTi-resistant mCRCs PDOs, this chapter aims to understand the adaptive changes in tumour populations that contributed to the development of drug resistance.

4.1. Introduction

For this purpose, the genomic DNA (gDNA) of mCRC PDOs was amplified using the PicoPlex Gold whole-genome amplification (WGA) kit (Takara). Aside from PicoPlex Gold WGA, several other WGA methods are available for detecting CNAs and SNVs in single-cell genomes, including DOP-PCR (328-332), MDA (333), MALBAC (334) and PTA (335). Each protocol presents unique limitations, including artefact formation, preferential amplification of specific genomic regions, allelic dropouts, and variable depth of coverage across the genome (336). These factors critically impact their suitability for CNA detection using read depth-based approaches, which rely on the principle that the number of reads covering a genomic region is proportional to the copy number of that region in the genome (337). Consequently, assessing the quality of single-cell libraries was a critical preliminary step in the CNA analysis to ensure that the depth of coverage was sufficient for identifying CNAs.

4.2 Aims

The general aims of this chapter are summarised as follows:

1. Amplify matched single-cell genomes with PicoPlex Gold whole genome amplification and assess the quality of libraries for DNA copy number analysis.
2. Characterise the clonal composition of mCRC PDOs based on genome-wide CNA changes.
3. Identify CNAs that are specific to AKTi-resistant mCRC PDOs.
4. Compare the frequency of identified subclones before and after acquiring drug resistance.
5. Compare the copy number changes observed at the single-cell level with those observed in bulk genomes.

The hypothesis behind these aims was that AKT inhibition in mCRC PDOs would lead to distinct genomic adaptations characterised by specific CNAs and changes in clonal composition. These adaptations would be easier to identify by single-cell sequencing than bulk sequencing. By comparing single-cell to bulk genome analyses, this research aimed to uncover the finer details of tumour evolution and resistance development.

4.3 Methods: Computational analysis of WGS data

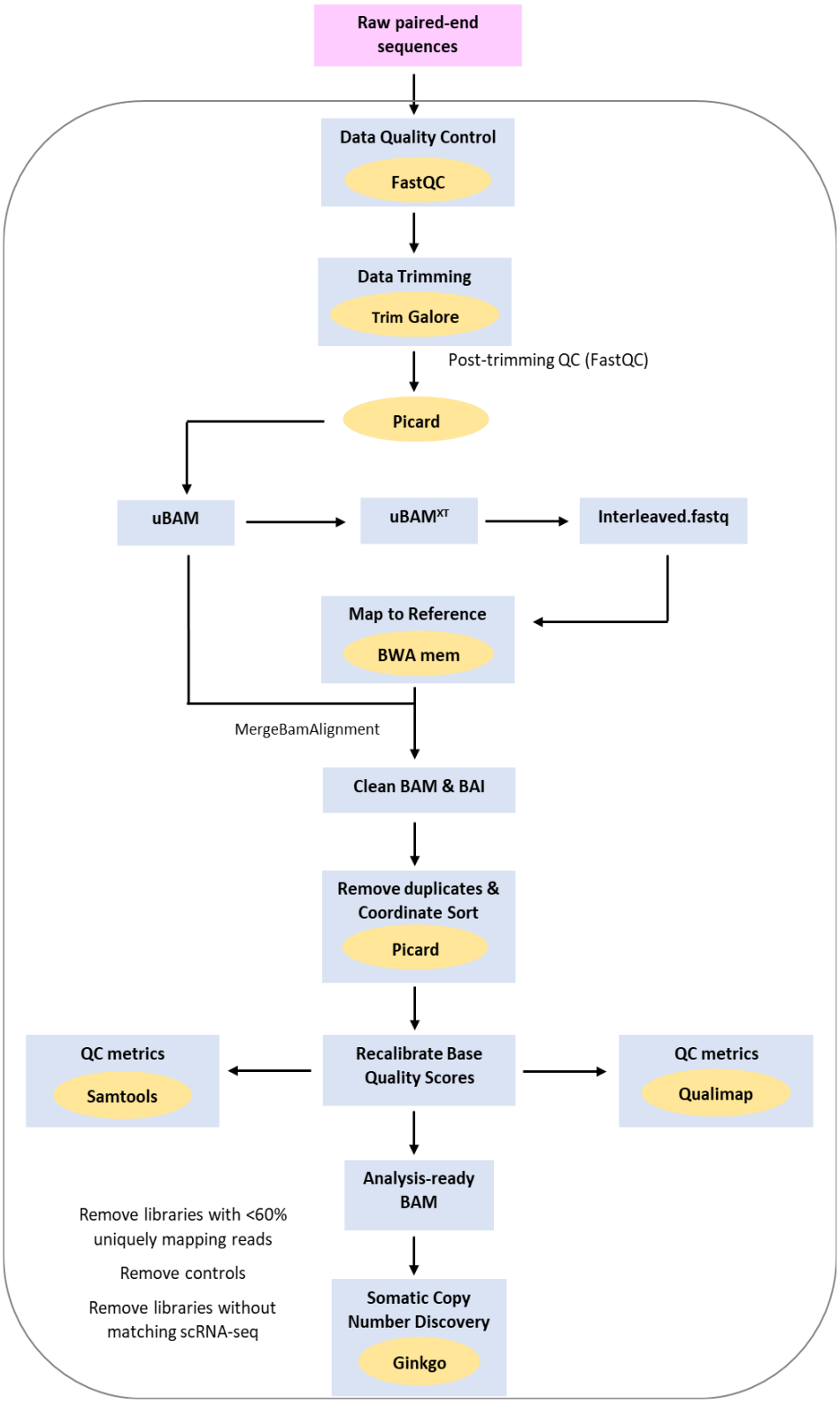


Figure 4.1. Bioinformatics workflow for the analysis of single-cell and bulk WGS data

Data processing includes quality control (QC) of raw sequencing reads (Trim Galore and FastQC), followed by a series of intermediate steps to generate clean and indexed BAM alignment files (Picard and BWA), ready for base quality recalibration (GATK), before copy number alteration analysis (Ginkgo).

4.3. Methods: Computational analysis of WGS data

Table 16. List of software packages employed for the analysis of scWGS data

Software	Access/citation
BEDTools v2.29.2	(338)
BWA v0.7.17	(339)
ComplexHeatmap v2.16.0	(157)
FastQC v0.11.9	(158)
GATK v4.2.0.0	(340)
GenomicRanges v1.52.1	(341)
ggplot v2.3.4.4	(161)
Ginkgo v3.0.0	(342)
Picard v2.25.7	(343)
Qualimap v2.2.1	(169, 170)
R v4.1.2	R Core Team (2022)
RStudio v2023.6.2.561	R Core Team (2022)
Samtools v1.15	(344)
tidyverse v2.0.0	(176)
Trim Galore v0.5.0	(177, 178)
ucsc_utils v333	(345)

4.3. Methods: Computational analysis of WGS data

4.3.1 Primary analysis of PicoPlex Gold-amplified single-cell WGS libraries

For each of the three mCRC PDOs—Parental, MK1-resistant, and AZD1-resistant—a total of 32 single-cell genomes (including positive and negative controls) were amplified using PicoPlex Gold Whole Genome Amplification (WGA), and subsequently submitted for WGS, yielding a total of 96 scWGS libraries.

Similar to the scRNA-seq data, the Genomics Pipelines (GP) team at the Earlham Institute applied their primary analysis pipeline (PAP) on the scWGS libraries. The results from the PAP were compiled into a comprehensive MultiQC report, summarising the quality of each library.

4.3.2 PicoPlex Gold scWGS-seq data processing and filtering of low-quality libraries

All computational steps described in this section were performed on the Norwich Bioscience Institute (NBI) High-Performance Computing (HPC) cluster under the SLURM workload management system version 23.02.7, along with R v4.1.2 and Python v3.10.3. The software packages and tools used to process and analyse scWGS data are detailed in Table 16. Default parameters were employed for all computational tools unless stated otherwise in the text.

Raw FASTQ files were initially trimmed to remove adapters and low-quality bases using Trim Galore v0.5.0. Quality of the trimmed reads was then assessed with FastQC v0.11.9. Further preprocessing was necessary to transform trimmed reads into a high-quality dataset of aligned reads. This involved removing duplicate reads from PCR amplification, which could skew copy number estimates, and ensuring accurate base quality scores for reliable copy number alteration calling. This required the use of Picard v2.25.7 and the Genome Analysis Toolkit (GATK) v4.2.0.0 packages from the Broad Institute. These tools, integral for variant discovery, supported two key workflows. The first involved the GATK workflow for efficient mapping and cleanup of short read sequence data (346). This was followed by the GATK data preprocessing workflow for variant discovery (347, 348) (Figure 4.2).

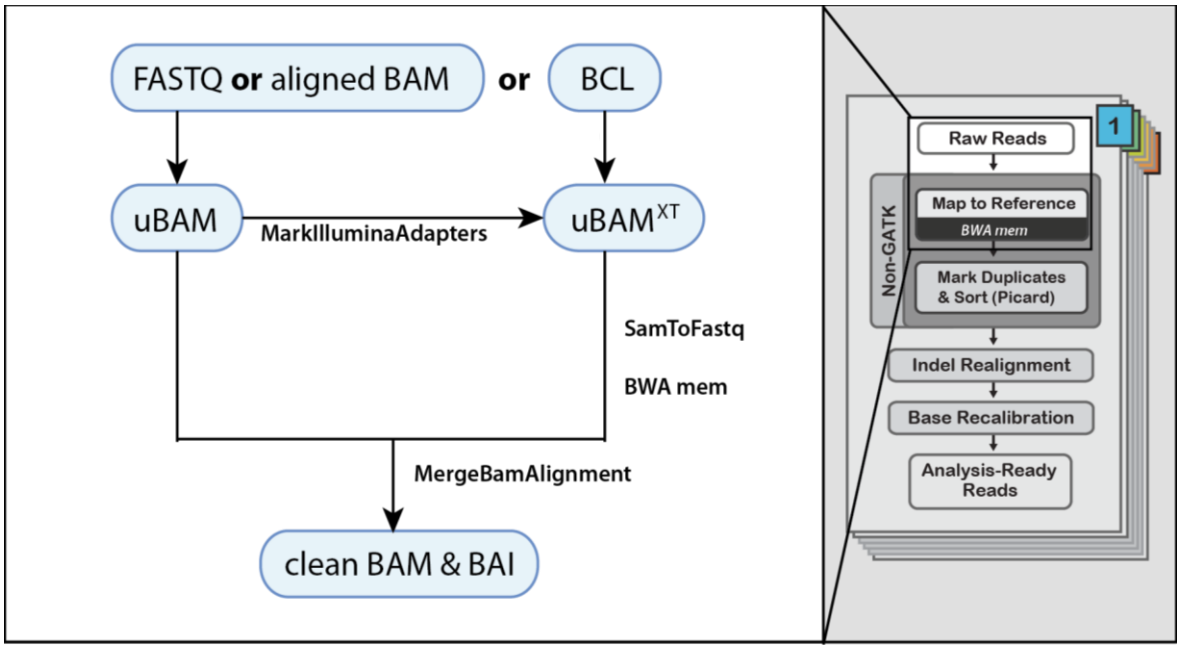


Figure 4.2. GATK workflow to efficiently map and clean up short read sequence data before variant analysis.

The input for this workflow can be files in FASTQ, aligned BAM or BCL formats (left panel). Starting with FASTQ or aligned BAM files, several Picard tools are initially employed to convert these files to unaligned BAM files (uBAM). In the subsequent step, adapter sequences are marked in the uBAM files. These files are then converted to interleaved FASTQ files before being mapped to the reference genome using BWA. In the final step, the uBAM and aligned BAM files from the previous steps are merged to generate clean and indexed BAM files, which are then ready for variant analysis with GATK tools (right panel). Figure adapted from (346).

In the workflow for efficient mapping and cleanup of short-read sequence data, FASTQ files were initially converted into unaligned BAM (uBAM) files using Picard's *FastqToSam*. This conversion was done to retain essential sample metadata and sequencing run information by inputting the following parameters: *--READ_GROUP_NAME, --SAMPLE_NAME, --LIBRARY_NAME, --PLATFORM_UNIT, --PLATFORM, --SEQUENCING_CENTER, --RUN_DATE*. Although adapter sequences had already been removed in previous steps by Trim Galore, the next step in the workflow involved using Picard's *MarkIlluminaAdapters*. This tool identifies and adds XT tags to adapter sequences that might not have been perfectly trimmed, thereby ensuring that any residual adapter sequences are accounted for in subsequent analyses. The third step in the workflow entailed converting the uBAM files back to FASTQ format using *SamToFastq*. With the parameters of *SamToFastq* set to *--CLIPPING_ATTRIBUTE XT*, *--CLIPPING_ACTION 2*, and *--INTERLEAVE true*, the tool produced interleaved FASTQ files without adapter sequences.

4.3. Methods: Computational analysis of WGS data

Before aligning reads with the Burrows-Wheeler Aligner (BWA) v0.7.17, the human GRCh38 assembly, its index and dictionary files were downloaded from the GATK resource bundle (349).

FASTQ reads were next aligned to the reference genome using BWA with the Maximal Exact Matches (MEM) algorithm. The *-p* flag indicated that the input files were interleaved FASTQ files, while the *-M* flag was used to mark shorter split hits as secondary (non-primary) alignments in the SAM file, a recommendation by GATK for generating data comparable to that produced by the Broad Genomics Platform. The output from this step was piped into Samtools (v1.15) *view*, with the *-h* flag to include the header in the output and *-b* to write the output in BAM format.

The final step in this first workflow involved merging the uBAM and aligned BAM files using Picard's *MergeBamAlignment*. This process ensured the preservation of metadata from the original files in the aligned files. It is important to note that "if the alignment file was missing reads present in the unaligned file, then these were retained as unmapped records" (348). *MergeBamAlignment* was employed with selected parameters:

i) *--PRIMARY_ALIGNMENT_STRATEGY MostDistant* to designate primary alignments as those giving the largest insert size relative to their mate pair. This selection prioritises alignments where the read and its mate are farthest apart, assuming that such alignments are more likely to occur in less repetitive and more unique regions of the genome.
ii) *--MAX_INSERTIONS_OR_DELETIONS -1* allows any number of insertions or deletions in the final alignment.
iii) *--ATTRIBUTES_TO_RETAIN XS* to retain secondary alignments marked with the XS attribute due to having suboptimal alignment scores.
iv) *--CLIP_ADAPTERS false* to avoid clipping adapter sequences from the reads.

The results of this first workflow were clean and indexed BAM alignment files.

In the data preprocessing workflow for variant discovery, duplicate reads in the BAM alignments were marked, but not removed, by setting Picard's *MarkDuplicates* *--REMOVE_DUPLICATES* option to *false*. The duplication percentage was then inspected using Qualimap (v2.2.1) *bamqc*. To reduce biases in variant calling due to the overrepresentation of duplicated sequences, PCR and optical duplicates were subsequently removed by setting *--REMOVE_DUPLICATES* to *true*. Afterwards, Qualimap *bamqc* was rerun to inspect the deduplicated BAM files.

4.3. Methods: Computational analysis of WGS data

The final step in this second workflow involved running GATK's Base Quality Score Recalibration (BQSR) tool, which recalibrates base quality scores to correct for systematic errors in the sequencing data. This enhancement of the base quality scores significantly improves the reliability of variant detection. The base recalibration process involves two main steps. Firstly, *BaseRecalibrator* creates a recalibration table using various covariates known to influence the accuracy of the base calls made by the sequencing machine. Examples of these covariates include the position of a base within the read and the machine cycle during sequencing. To create the recalibration table, *BaseRecalibrator* requires three types of input: the clean and deduplicated BAM alignment file, the reference genome used for alignment, and a set of known sites of variation in VCF format. These known sites are essential for masking genomic regions where true variation is expected. The files containing known sites were sourced from the GATK resource bundle (349), and comprised the following:

i) Homo_sapiens_assembly38.dbsnp138.vcf.gz,
ii) Homo_sapiens_assembly38.known_indels.vcf.gz,
iii) Mills_and_1000G_gold_standard.indels.hg38.vcf.gz.

Subsequently, *ApplyBQSR* utilised the recalibration table generated in the previous step to adjust the base quality scores in the original BAM file. This final step created deduplicated, recalibrated, and indexed BAM files ready for variant calling. For each of the 96 recalibrated BAM files, key metrics, including the total number of reads, duplication percentage, overall mapping rate, mean depth of coverage, and the percentage of the reference genome covered at different depths of coverage (i.e., breadth of coverage), were extracted from the "genome_results.txt" file generated by Qualimap *bamqc*. Additionally, the number of uniquely mapping reads was counted using Samtools *view -c* (v1.15), with the mapping quality (*-q*) parameter set only to count reads with a mapping quality score of *20* or higher. This Phred-scaled score implies that there is a 1% chance ($P = 10^{\frac{-Quality\ value}{10}}$) that the mapping position is incorrect. These metrics were compiled into a table and used as input to create box plots illustrating the quality of libraries with ggplot v2.3.4.4. Low-quality PicoPlex Gold-amplified libraries were then filtered out based on the following criteria:

i) Libraries where the percentage of uniquely mapped reads fell below the set threshold of 60% mapping.
ii) Genomic libraries with matching scRNA-seq data that did not meet Seurat's initial quality control standards.
iii) Positive and negative control libraries.

4.3.3 Ginkgo genome-wide copy number analysis of single cells

Ginkgo was used for single-cell copy number alteration (CNA) analysis (342). For this purpose, BAM files were first converted to BED format using BEDTools (v2.29.2) *bamtobed* with default options. While the PicoPlex Gold libraries were aligned to the hg38 (GRCh38) reference genome, the web version of Ginkgo is only compatible with the hg19 (GRCh37) genome build. Therefore, before performing the CNA analysis, it was essential to first convert genomic coordinates from hg38 to hg19. This conversion was achieved using the UCSC *liftOver* (v333) utility. The chain file required for translating genomic coordinates from hg38 to hg19 (hg38ToHg19.over.chain.gz) was downloaded from the UCSC Golden Path liftOver site (350).

Subsequently, the following Ginkgo parameters were employed for the CNA analysis of mCRC single-cell genomes:

i) In line with the recommendations in the Ginkgo publication, the hg19 reference genome was divided into variable-length bins averaging 500 kb (342).
ii) The binning simulation option was configured for 150 bp reads mapped with BWA to establish the variable-length bin boundaries.
iii) The binned data was segmented using independent (normalised) read counts.
iv) "Bad bins", such as those around the centromeres of certain chromosomes, were masked due to their tendency to attain very high read counts compared to other genomic locations (342). Similarly, pseudoautosomal regions of the Y-chromosome were also masked.
v) Both sex chromosomes were included in the analysis, although only the X chromosome will be reported in the results.

To identify noisy DNA libraries after CNA analysis, the "**SegNorm.txt**" output, containing normalised bin counts for each cell at every bin position, was used to compute the median absolute deviation (MAD) for each cell. To calculate the MAD, the pairwise differences in read counts between neighbouring bins were first determined for each cell. This step quantified the change in read counts from one bin to the next. Subsequently, the MAD was computed by taking the median of the absolute values of these pairwise differences (342). Additionally, the "**SegStats.txt**" file was employed to plot other quality control metrics, such as the index of dispersion (IOD) (342). A cut-off for MAD scores was established following the "1.5 IQR rule" for outlier detection (351), after which the Ginkgo analysis was rerun without the "bad" libraries.

Ginkgo provides several visualisations for scCNA analysis; however, the "**SegCopy.txt**" output file was utilised for further analysis in R. This file, containing integer copy numbers for each cell at every bin position, enabled the generation of a heatmap of copy number profiles for cells that passed quality control. This heatmap was created using ComplexHeatmap v2.16.0 and enhanced with PDO-specific row annotations. Additionally, the hierarchical clustering of cells, calculated by Ginkgo using Euclidean distance and Ward's linkage and provided in Newick format, was directly adopted without modifications.

The "**SegCopy.txt**" file was also employed to calculate the average ploidy for each sample type (Parental and AKTi-resistant organoids), as Gingko did not provide this. This process involved initially subsetting the copy number matrix by PDO. Subsequently, the average copy number for each cell was calculated across all bins, with this column-wise average representing the ploidy of each individual cell. To determine the mean ploidy for each PDO, the mean of these ploidy values across all cells within a PDO was computed. Other scCNA visualisations presented in this chapter were directly generated by Ginkgo.

Finally, GenomicRanges v1.52.1 was employed to annotate CNAs with chromosome cytoband information. This process involved using *GRanges* to create GRanges objects, and *findOverlaps* on the resulting objects, to identify overlapping regions between the CNA data stored in the "**CNV1.txt**" output and the cytoband data for the hg19 assembly. The cytoband data file ("cytoBand.txt.gz") was downloaded from the UCSC genome annotation database (352).

4.3.4 Data processing and downsampling of bulk genomes for Gingko copy number analysis

The computational approach used to process single-cell genomes was used for processing the blood, Parental, MK1-resistant, and AZD1-resistant bulk WGS data generated by the ICR (Figure 4.1). However, the filtering steps were omitted in this process.

To validate the results obtained from the scCNA analysis, the cleaned and deduplicated bulk BAM files were downsampled to the single-cell average number of reads, just over 3.5 million, using Picard's *DowsampleSam*. The probability (P) parameter that a given read will be retained in the downsampled BAM file must be first calculated to downsample a BAM file to an approximate number of reads. This probability was calculated based on the total number of reads in each input BAM file, using the following formula (353):

$$P = \frac{Desired\ number\ of\ reads}{\text{Total number of reads in the input BAM}}$$

DowsampleSam was next run with the calculated probabilities, employing a *chained* downsampling strategy.

Finally, Ginkgo was used for copy number analysis of the downsampled BED-converted files, applying the same parameters as those employed with the PicoPlex Gold-amplified single-cell libraries.

4.3.5 Sequenza genome-wide copy number analysis of bulk genomes

The ICR processed the bulk data using a Nextflow workflow (354) specifically tailored for variant detection in whole genome and targeted sequencing data, integrates multiple steps. These include read trimming with Trim Galore, alignment to the GRCh38 reference genome via BWA MEM, duplication marking using Picard tools, and base quality score recalibration with GATK.

CNA calling was next performed using Sequenza, a comprehensive package designed for inferring tumour cellularity, ploidy, and allele-specific copy number profiles from WGS data utilising matched normal and tumour samples (355). This analysis was also implemented using Nextflow (356). The Sequenza analysis was performed using 100 kb bins. Other parameters are specified in the "analyse_cn_sequenza.R" script (357). Consequently, the Sequenza figures included in this chapter were sourced from the Accelerator project's shared drive, showcasing the results obtained from this analytical approach.

4.4 Results

4.4.1 Primary analysis of single-cell genomes reveals DNA libraries free from bacterial contamination

To ensure scWGS analyses were performed on high-quality data, the raw reads corresponding to sequencing libraries underwent extensive processing. The initial inspection of libraries focused on identifying the species represented in the sequences. Table 17 presents an extract from the PAP MultiQC report of the PicoPlex Gold-amplified DNA libraries, focusing on the Centrifuge module. This extract reveals the top two most abundant species in 26 representative libraries. Unlike the scRNA-seq libraries, which showed contamination by *Variovorax* species, the matched genomic libraries primarily consisted of human DNA. This suggests that the contamination observed in the scRNA-seq data occurred after separating the genomic DNA and the polyadenylated mRNA, i.e., during the Smart-seq2 arm of the G&T-seq protocol.

4.4. Results

Table 17. PAP report extract showing the top 2 most abundant species in representative single-cell DNA libraries derived from mCRC organoids

Sample Name	1st Name	1st %	2nd Name	2nd %
SOGTseqPPGoldA10	*Homo sapiens*	97.9	eukaryotic synthetic construct	0.7
SOGTseqPPGoldA11	*Homo sapiens*	71.5	*Pinus taeda*	17.3
SOGTseqPPGoldA12	*Homo sapiens*	94.3	eukaryotic synthetic construct	0.9
SOGTseqPPGoldA1	*Homo sapiens*	97.9	eukaryotic synthetic construct	0.7
SOGTseqPPGoldA2	*Homo sapiens*	97.8	eukaryotic synthetic construct	0.8
SOGTseqPPGoldA3	*Homo sapiens*	97.6	eukaryotic synthetic construct	0.7
SOGTseqPPGoldA4	*Homo sapiens*	97.7	eukaryotic synthetic construct	0.8
SOGTseqPPGoldA5	*Homo sapiens*	97.8	eukaryotic synthetic construct	0.7
SOGTseqPPGoldA6	*Homo sapiens*	97.8	eukaryotic synthetic construct	0.8
SOGTseqPPGoldA7	*Homo sapiens*	97.6	eukaryotic synthetic construct	0.8
SOGTseqPPGoldA8	*Homo sapiens*	97.8	eukaryotic synthetic construct	0.7
SOGTseqPPGoldA9	*Homo sapiens*	97.7	eukaryotic synthetic construct	0.7
SOGTseqPPGoldB10	*Homo sapiens*	97.7	eukaryotic synthetic construct	0.8
SOGTseqPPGoldB11	*Homo sapiens*	97.8	eukaryotic synthetic construct	0.8
SOGTseqPPGoldB12	*Homo sapiens*	97.3	eukaryotic synthetic construct	0.9
SOGTseqPPGoldB1	*Homo sapiens*	96.4	*Pan troglodytes*	0.7
SOGTseqPPGoldB2	*Homo sapiens*	97.7	eukaryotic synthetic construct	0.7
SOGTseqPPGoldB3	*Homo sapiens*	97.5	eukaryotic synthetic construct	0.8
SOGTseqPPGoldB4	*Homo sapiens*	97.7	eukaryotic synthetic construct	0.7
SOGTseqPPGoldB5	*Homo sapiens*	97.6	*Pan troglodytes*	0.7
SOGTseqPPGoldB6	*Homo sapiens*	97.7	eukaryotic synthetic construct	0.7
SOGTseqPPGoldB7	*Homo sapiens*	97.7	eukaryotic synthetic construct	0.7
SOGTseqPPGoldB8	*Homo sapiens*	97.9	eukaryotic synthetic construct	0.7
SOGTseqPPGoldB9	*Homo sapiens*	97.8	eukaryotic synthetic construct	0.7
SOGTseqPPGoldC10	*Homo sapiens*	97	eukaryotic synthetic construct	0.8
SOGTseqPPGoldC11	*Homo sapiens*	97.7	eukaryotic synthetic construct	0.8

4.4.2 Quality control and filtering of single-cell genomes

Figure 4.3A-F displays a series of box plots illustrating the distribution of 96 scWGS libraries across various sequencing metrics before excluding low-quality libraries and controls. The read count distribution in the scWGS libraries exhibited a right-skewed pattern, as evidenced by a longer upper whisker representing libraries with an unusually high number of reads, and a mean of 3.9 million reads (standard deviation=±3.3 million) that exceeded the median of 2.8 million reads (Figure 4.3A). The high standard deviation suggests significant inconsistency and a wide dispersion of read counts across the libraries.

Despite the variability in read counts, the duplication percentage remained generally low across the DNA libraries, averaging 13.4% (±3.8%) (Figure 4.3B). Furthermore, the overall mapping percentage was also satisfactory, with an average of 97.7% (±5.5%) (Figure 4.3C) and uniquely mapping read percentages averaging 79.3% (±6.7%) (Figure 4.3D).

Additional metrics for assessing sequencing quality included the mean depth of coverage across libraries (Figure 4.3E), which averaged 0.29× (±0.24×). The large standard deviation again suggests considerable variation in the sequencing coverage across libraries. Lastly, the relationship between depth and breadth of coverage was examined (Figure 4.3F), which revealed that the average percentage of the genome covered by the libraries at the lowest depth of coverage (1×) was 13.6% (±7.8%), while at the highest depth of coverage (10×) it was 0.25% (±0.46%).

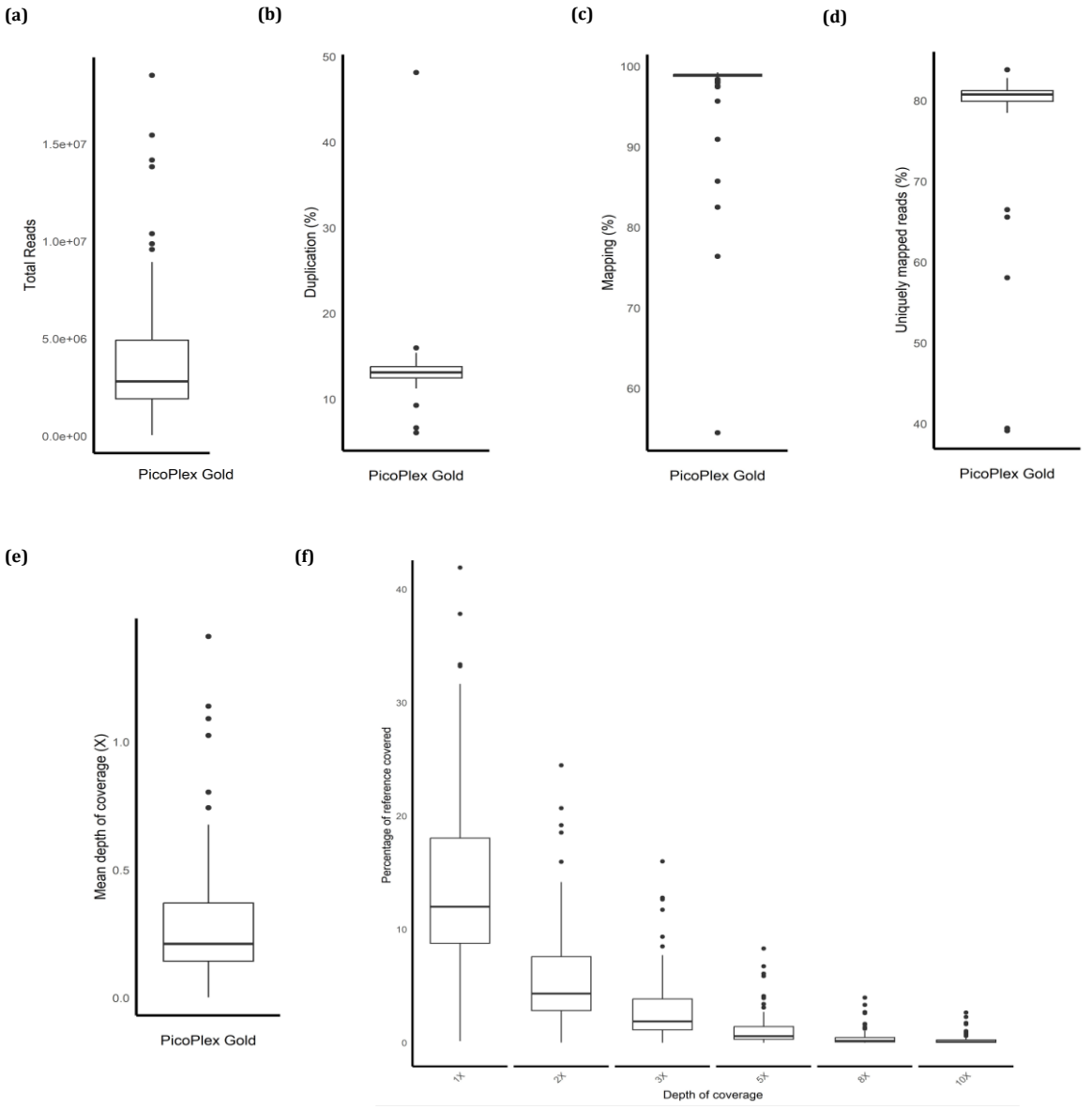


Figure 4.3. Distribution of PicoPlex Gold-amplified WGS libraries derived from mCRC organoid across various quality control metrics prior to filtering low-quality libraries.

Pre-filtering quality control metrics of single-cell genomes derived from Parental, MK1- and AZD1-resistant mCRC PDOs, amplified using PicoPlex Gold, and subsequently subjected to Whole Genome Sequencing. Metrics include: ***(a)*** *total number of reads,* ***(b)*** *percentage of duplicated reads,* ***(c)*** *overall mapping rate,* ***(d)*** *percentage of uniquely mapping reads,* ***(e)*** *mean depth of coverage, and* ***(f)*** *breadth of coverage, defined as the fraction of the reference genome covered at various depths of coverage. Each box plot displays the median value (central line), the interquartile range (25th and 75th percentiles as the box boundaries), and 1.5× the interquartile range (whiskers). 96 libraries in total, encompassing genomes from multi-cell and negative controls and single cells.*

Filtering was next performed based on these metrics. Out of the 96 gDNA samples chosen for PicoPlex Gold WGA and subsequent WGS—which included 30 libraries for each PDO, along with 3 positive (matched genomes from 50 cells obtained after the physical separation of gDNA and mRNA) and 3 negative (lysis buffer) control libraries—a total of 85 libraries were retained. The reduction was due to the exclusion of both positive and negative control libraries and those with matched scRNA-seq libraries that failed to pass Seurat's quality control because of high mitochondrial read content. The remaining DNA libraries comprised 30 Parental, 27 MK1-resistant, and 28 AZD1-resistant libraries. Examining the same sequencing metrics post-filtering revealed that both the average read counts (3.4 million reads) and standard deviation values (±2.3 million reads) were closer in the filtered dataset (Figure 4.4A) than in the unfiltered dataset. This suggests that the filtered samples, especially the positive controls, were outliers that affected the overall statistics. Despite removing "bad" libraries, there was still a right-skewness in the data, as the mean remained higher than the median (2.6 million reads), although the difference between them was reduced.

Other metrics, such as the duplication rate (Figure 4.4B) and mapping percentages (Figure 4.4C-D), remained favourable. On the other hand, the average depth of coverage was slightly reduced after filtering (0.26×, ±0.17×) (Figure 4.4E). The decrease in the standard deviation indicates that the libraries remaining after filtering had less variability in their coverage depth, which is desirable for further analysis. Nevertheless, the average depth of coverage for these libraries exceeded the 0.036× (±0.022×) depth reported for the 96 gDNA libraries amplified with the standard PicoPlex scWGA Kit (Takara) in the original G&T-seq publication (129). Consequently, the coverage level of the mCRC single-cell genomes was sufficient for low-pass CNA analysis using Ginkgo (342).

Regarding the breadth of coverage, the filtered dataset maintained the same trend of decreasing genome coverage with increasing coverage depth (Figure 4.4F). This last plot is useful for understanding how the depth of sequencing affects the proportion of the genome that can be analysed. It exemplifies that at higher depths of coverage, sequencing is not uniformly distributed across the entire genome. Instead, these higher depths are concentrated in specific, limited regions, while lower depths cover the majority of the genome. This is particularly relevant for CNA analysis at low depths, where a more significant portion of the genome is typically covered, providing a broader genomic context for variant detection. However, for accurate CNA detection, it is essential to achieve not only adequate depth but also uniform coverage across the genome.

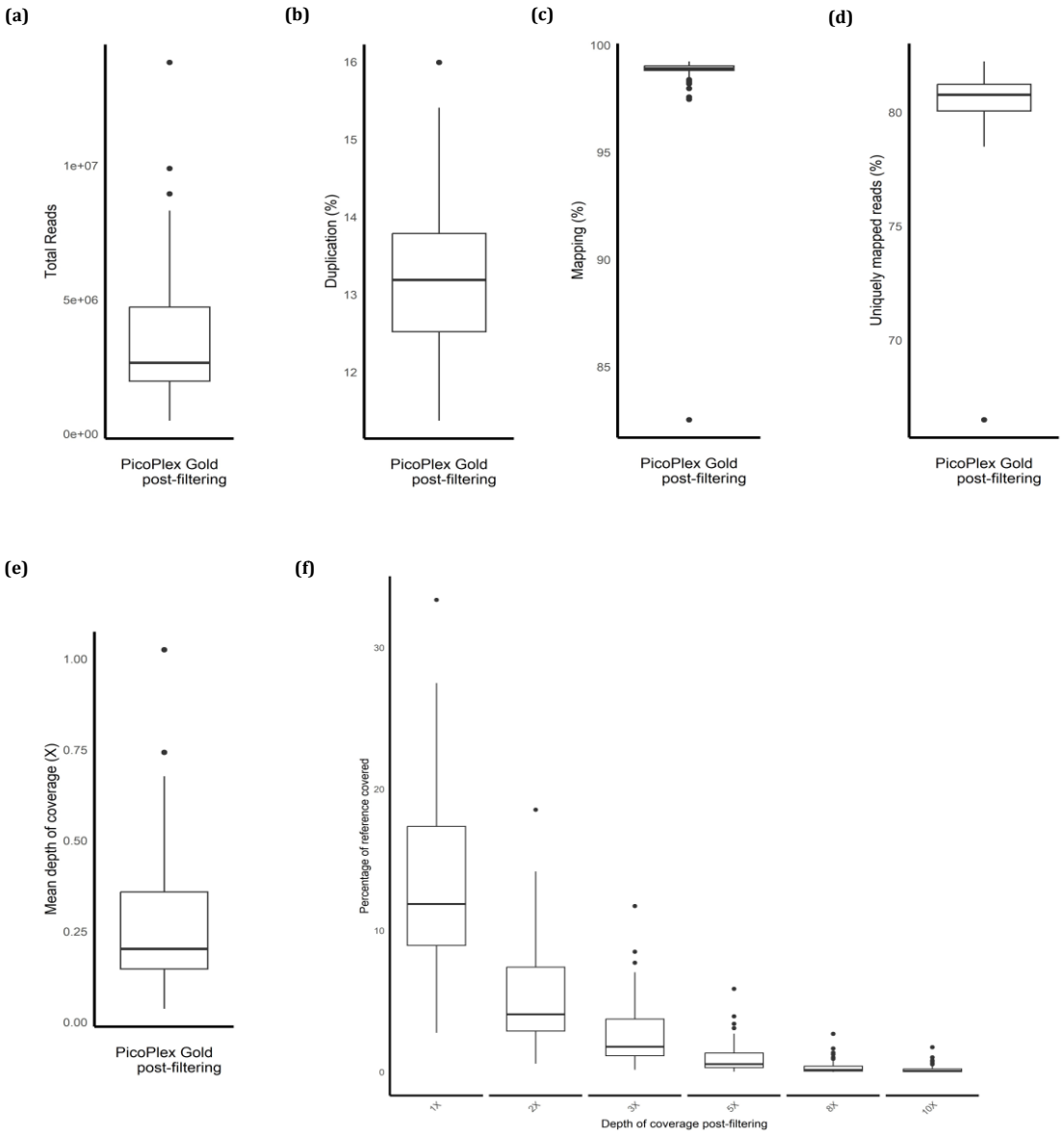


Figure 4.4. Distribution of PicoPlex Gold-amplified WGS libraries derived from mCRC tumoroids across various quality control metrics after filtering low-quality libraries.

Metrics include: ***(a)*** *total number of reads,* ***(b)*** *percentage of duplicated reads,* ***(c)*** *overall mapping rate,* ***(d)*** *percentage of uniquely mapping reads,* ***(e)*** *mean depth of coverage, and* ***(f)*** *breadth of coverage at various coverage depths. Each box plot displays the median value (central line), the interquartile range (25th and 75th percentiles as the box boundaries), and 1.5× the interquartile range (whiskers). Outliers are shown as individual points beyond the whiskers. 85 libraries in total, including 30 Parental, 27 MK1-resistant and 28 AZD1-resistant single-cell genomes passing quality control filters, and with available matching scRNA-seq data.*

4.4.3 PicoPlex Gold WGA provides reliable data for accurate copy number analysis in single-cell genomes

Following a preliminary scCNA analysis with Ginkgo (342), the variability in read depth across segmented genomic bins was evaluated. For CNA detection in sequencing data analysis, it is generally assumed that read counts across the genome follow a Poisson distribution in regions without CNAs (358, 359). This assumption implies a uniform coverage model, where each genomic region is equally likely to be sequenced, although perfect uniformity is not expected due to inherent random variations. Deviations from this Poisson distribution model indicate CNAs, as they lead to unexpected alterations in read count distribution.

The index of dispersion (IOD), also known as the Poisson or Fisher dispersion index, is defined as the variance-to-mean ratio of read counts (360). Ginkgo utilises this metric to assess how read counts across bins deviate from their average, serving as an indicator of coverage dispersion (342) (Figure 4.5A). The higher the dispersion index, the greater the variability between the read counts in each bin, which can indicate more "noise" or unevenness in the coverage distribution across the genome.

Under a Poisson model, an IOD of 1 is expected when the variance of read counts equals the mean (359, 361), indicating that any observed variability is attributable to the inherent randomness of the Poisson process (362). However, the average IOD of 0.53 (±0.08) observed in PicoPlex Gold libraries points to underdispersion relative to the Poisson expectation, a scenario where the variance is less than the mean (363). This underdispersion suggests a more even distribution of read counts across the genome than what would be expected by chance alone. Nevertheless, due to its reliance on the mean, the IOD is sensitive to outliers, meaning regions with exceptionally high or low read depths can significantly skew the metric.

In contrast, the median absolute deviation (MAD) of all pairwise differences in read counts between neighbouring bins is a robust measure of variability, particularly suited for datasets where a normal distribution cannot be assumed (364). Unlike the IOD, MAD focuses on variability around the median of a dataset, making it less susceptible to the influence of outliers (364). Given the positive skewness identified in the PicoPlex Gold dataset, employing MAD provided a more accurate reflection of sequencing depth variability.

The average MAD score for PicoPlex Gold libraries was 0.27 (±0.08) (Figure 4.5B), which was lower than the MAD scores reported for MALBAC and MDA amplified single-cell genomes in the Ginkgo publication (342). This reflects minimal coverage dispersion attributable to technical noise in PicoPlex libraries. Nevertheless, there were a few outliers in the dataset. For this

reason, a maximum threshold for MAD was established to filter out mCRC genomes with elevated noise levels, which carry an increased risk of generating false-positive results. This cut-off was set at 1.5 times the IQR above the third quartile, following the "1.5 IQR Rule" for outlier detection (351), a practice previously employed to filter scWGS libraries based on a similar metric, the median absolute pairwise difference (MAPD) (365). Consequently, genomes with a MAD score greater than 0.42 were excluded, resulting in the removal of 5 MK1-resistant genomes. Indeed, libraries with higher MAD scores exhibited a significant deviation from the line denoting perfect coverage uniformity in the Lorenz curves (Figure 4.5C-D) and displayed noisier CNA profiles compared to libraries with lower MAD scores (Figure 4.6A-D). While more stringent cut-offs for MAD, such as exclusions above 0.3-0.35, have been reported (333, 366), higher MAD and even MAPD thresholds have also been considered for CNA analyses or flagged for review (129, 367).

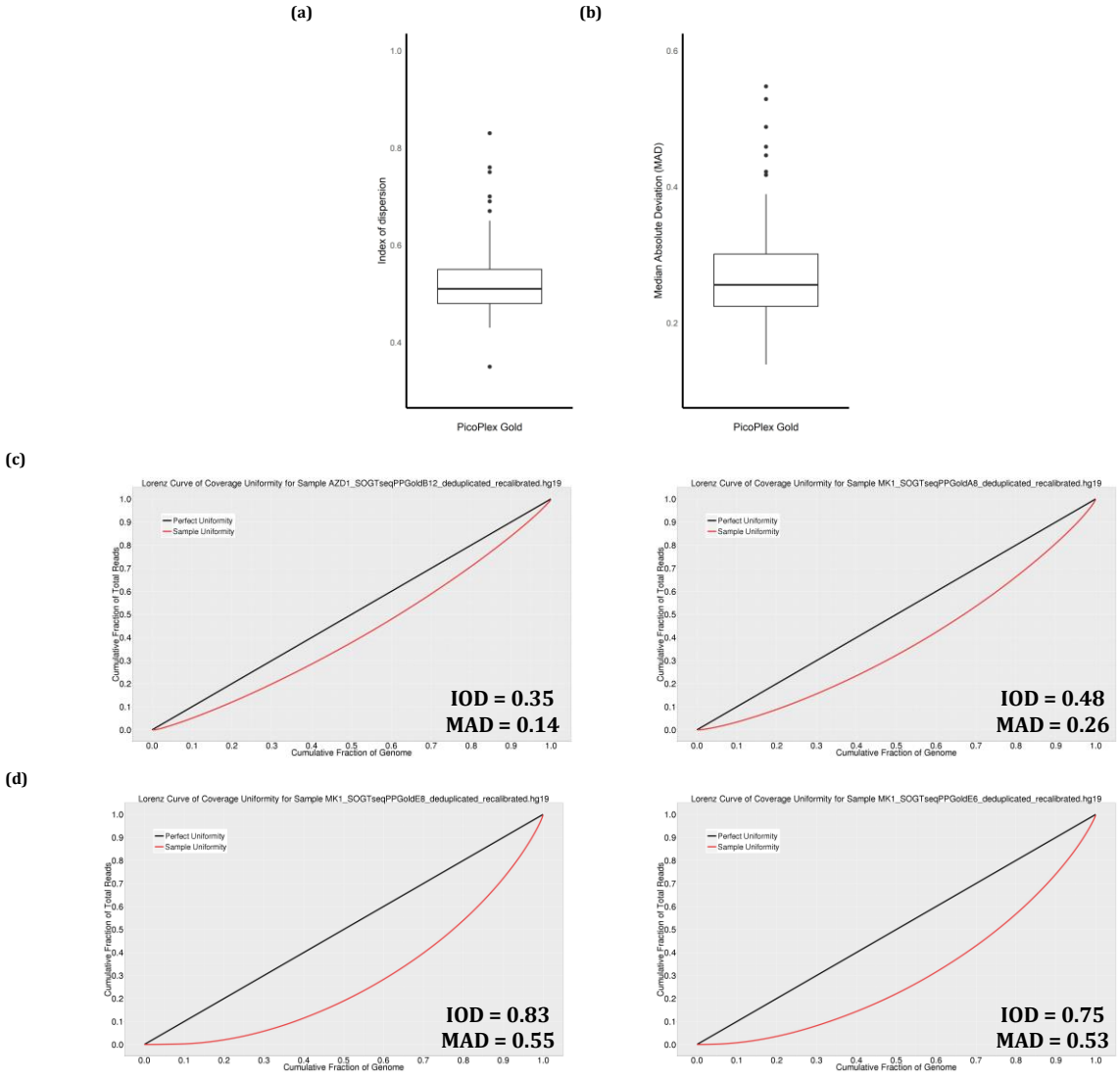


Figure 4.5. Assessment of coverage uniformity in PicoPlex Gold-amplified libraries for accurate CNA detection.

Plots display coverage uniformity metrics for 85 single-cell genomes derived from mCRC PDOs, as generated by Ginkgo following CNA analysis using variable-length 500 kb genomic bins. These metrics include: ***(a)*** *index of dispersion (IOD) and* ***(b)*** *median absolute deviation (MAD) of read counts between neighbouring bins. Each box plot displays the median value (central line), the interquartile range (25th and 75th percentiles as the box boundaries), and 1.5× the interquartile range (whiskers). Outliers are shown as individual points beyond the whiskers. Additionally, Lorenz curves for two single-cell genomes, each with low* ***(c)*** *and high* ***(d)*** *IOD scores, illustrate the deviation from a perfectly uniform genome. Collectively, these visualisations underline the variability in scWGS data for accurate CNA detection.*

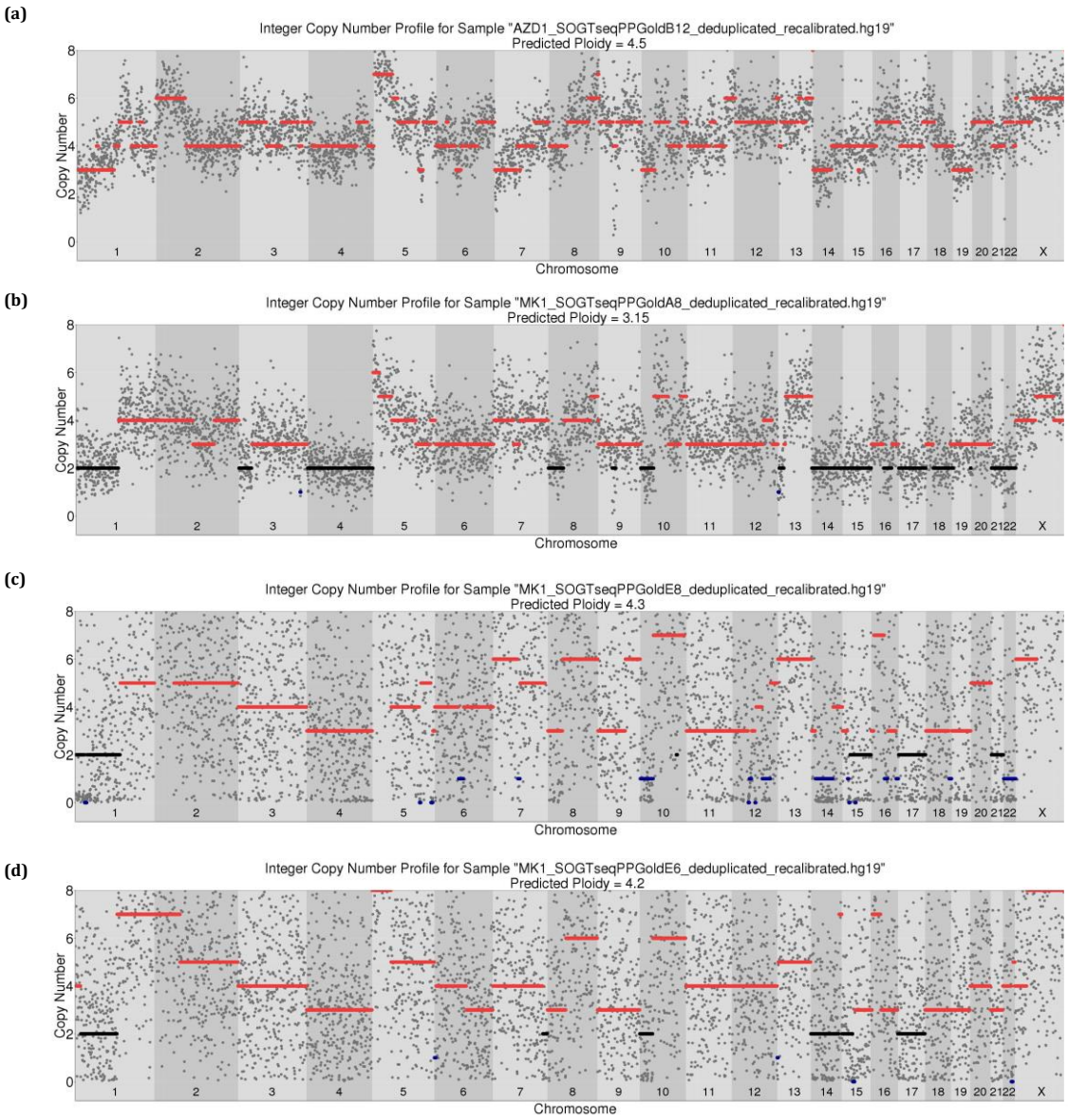


Figure 4.6. Coverage distribution in good and noisy single-cell genomes shows significant differences in the quality and reliability of CNAs detected.

Visual representation of genome-wide CNA profiles of selected single-cell genomes from mCRC PDOs with low ***(a-b)*** *and high* ***(c-d)*** *MAD scores. Grey dots represent the inferred copy number at specific genomic loci. Coloured lines indicate the smoothed median copy number state across genomic segments: black lines denote diploid segments, red lines are segments with amplifications, and blue lines represent segments with deletions. The dispersion of these dots provides insights into the confidence level at which integer copy number states can be accurately determined.*

4.4. Results

After excluding libraries with high MAD scores and subsequently rerunning the Ginkgo analysis, a moderate decrease in the average values of both IOD (0.52 ± 0.06) and MAD (0.26 ± 0.06) was observed (Figure 4.7). This decrease in the average scores, along with a reduction in their standard deviation, indicates reduced variability among the remaining scWGS libraries. Following this filtering process, the refined set of PicoPlex Gold libraries consisted of 30 Parental, 22 MK1-resistant, and 28 AZD1-resistant libraries, representing a total of 80 scWGS libraries.

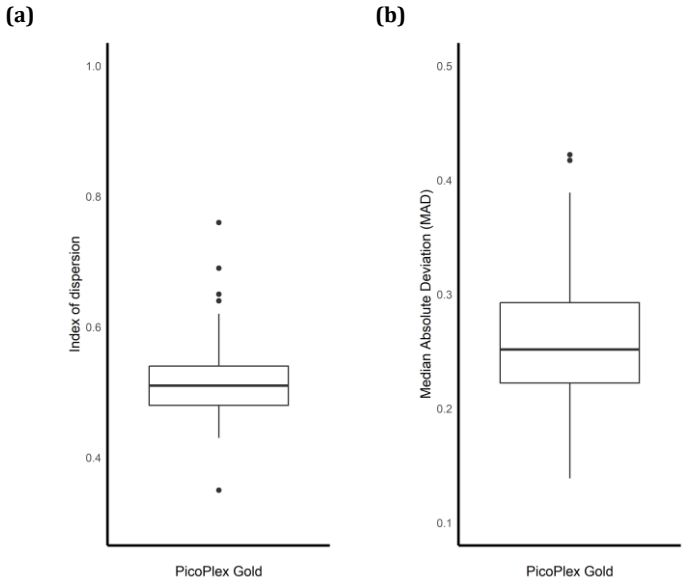


Figure 4.7. Distribution of PicoPlex Gold-amplified libraries after filtering noisy libraries

Box plots display the distribution of coverage uniformity metrics for 80 single-cell genomes from mCRC PDOs after excluding noisy libraries. Metrics presented include: ***(a)*** *index of dispersion (IOD) and* ***(b)*** *median absolute deviation (MAD) of read counts between neighbouring bins. Each box plot displays the median value (central line), the interquartile range (25th and 75th percentiles as the box boundaries), and 1.5× the interquartile range (whiskers). Outliers are shown as individual points beyond the whiskers.*

4.4.4 Comparative ploidy analysis of mCRC organoids at single-cell and bulk resolutions

The CNA analysis of single-cell genomes derived from mCRC PDOs performed using 500 kb genomic bins revealed consistent aneuploidy across the cells (Figure 4.8). This was evidenced by the average ploidy values of PDOs: 3.11 (±1.06) for Parental, 3.38 (±1.17) for MK1-resistant, and 3.21 (±1.12) for AZD1-resistant cells.

To assess the reliability of the single-cell findings, bulk WGS data, including Parental and both AKTi-resistant lines, along with a blood control, were downsampled to match the average number of reads observed in the single-cell genomes (approximately 3.5 million reads). Ginkgo analysis of the downsampled bulk data using 500 kb windows revealed a consistent ploidy of 3.2 for all three organoids (Figure 4.9). Additionally, genome-wide copy number estimates from the downsampled data displayed decreased variability compared to those observed in single cells (Figure 4.8). This is evidenced by the clustering of the "grey dots" in Figure 4.9, which represent CNA calls, around the smoothed median copy number state across genomic segments (coloured horizontal lines). This alignment indicates more reliable CNA predictions and is likely due to the lack of genomic amplification during sample processing of bulk genomes, an advantage that remained even when the data was downsampled to single-cell read depths. (See Supplementary Figure 13 and Supplementary Figure 14 for detailed quality control metrics of the bulk sequencing data.)

The primary challenge when using a read-depth-based approach for calling CNAs lies in selecting the optimal window size for dividing the genome (358, 368). This is crucial for accurately identifying CNAs, particularly in low coverage scWGS data, as in PicoPlex Gold-amplified libraries. In this project, the genomic window size used in the Ginkgo publication (500 kb) was employed for copy number analysis (342). Therefore, the effects of using a smaller window size to detect CNAs were not evaluated. Nevertheless, the Sequenza (355) CNA analysis of the complete bulk dataset using 100 kb bins revealed slightly higher ploidy estimates compared to the previous single-cell and downsampled analyses, with ploidy values of 3.45 for the Parental, 3.35 for the MK1-resistant, and 3.35 for the AZD1-resistant PDO (Supplementary Figure 15). Despite this, the CNA profiles were broadly similar across the single-cell, downsampled bulk, and complete bulk datasets (Figure 4.10, Figure 4.11, Figure 4.12).

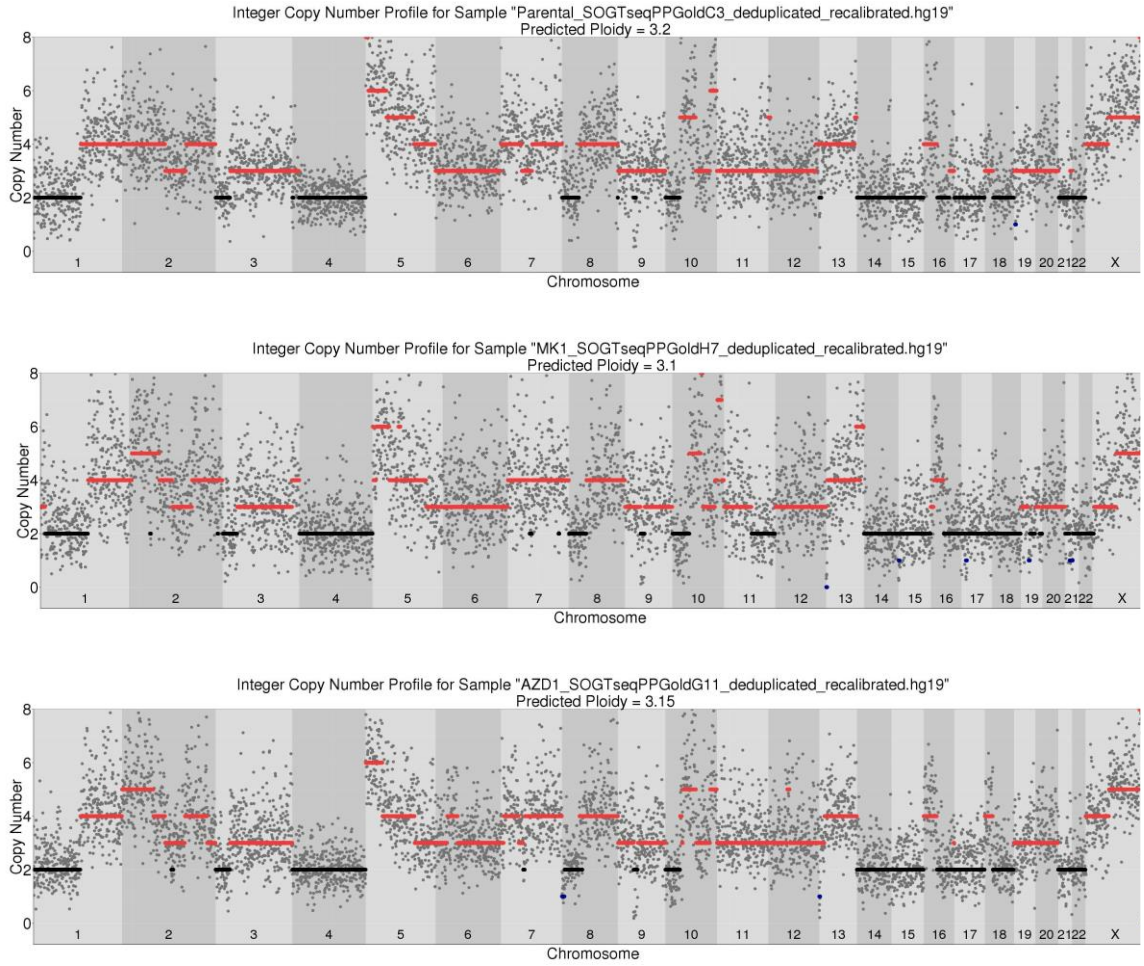


Figure 4.8. Example single-cell CNA profiles from the Parental, MK1- and AZD1-resistant mCRC PDOs

Ginkgo-generated CNA profiles of single-cell genomes derived from the Parental (top), MK1-resistant (middle), and AZD1-resistant (bottom) organoids. The analysis was performed using variable-length bins averaging 500 kb. In these figures, grey dots represent the inferred copy number at specific genomic loci. Coloured lines indicate the smoothed median copy number state across genomic segments: black lines denote diploid segments, red lines are segments with amplifications, and blue lines represent segments with deletions.

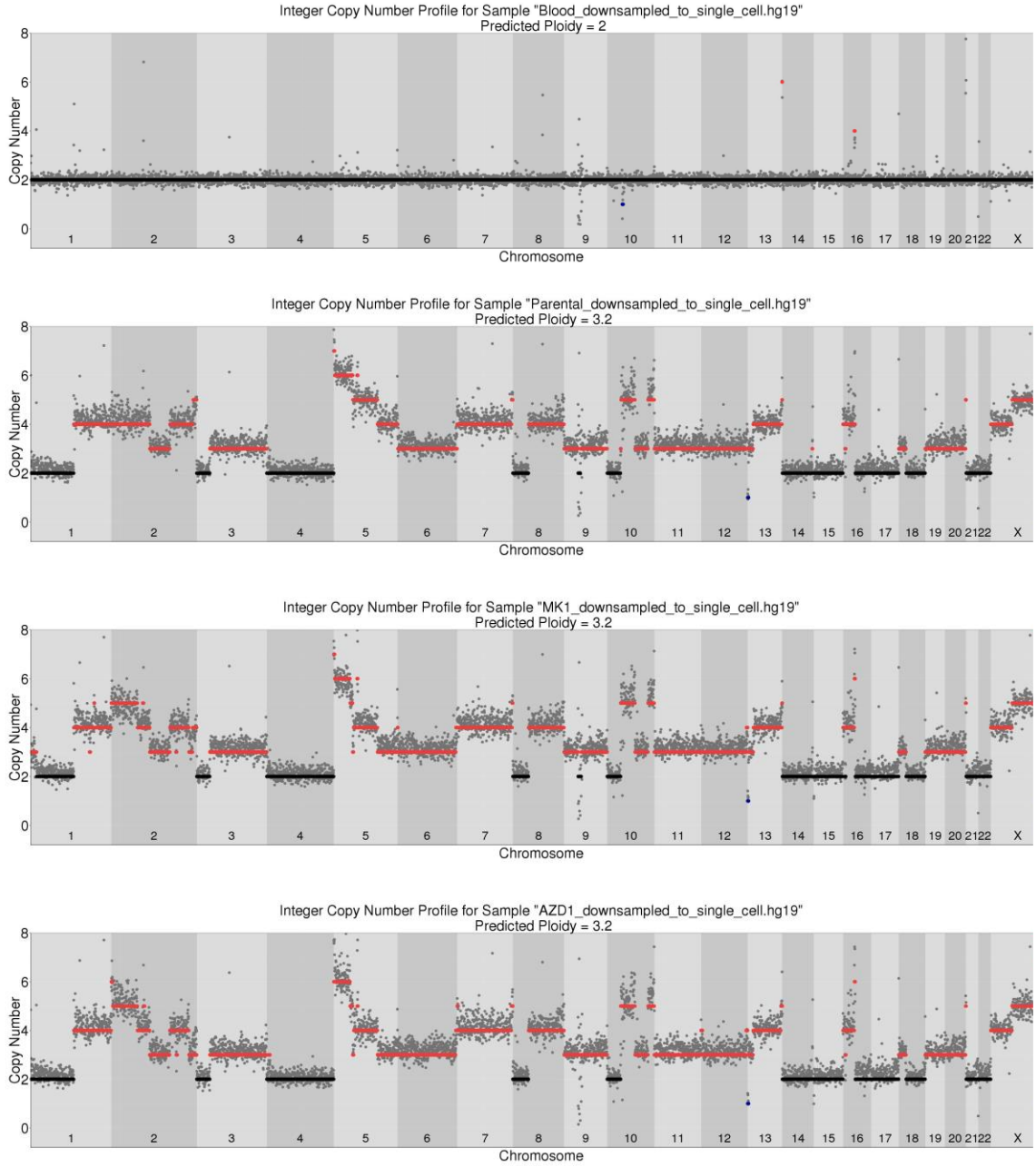


Figure 4.9. Downsampled bulk WGS CNA profiles for mCRC PDOs and matched blood control.

Ginkgo-generated CNA profiles of bulk WGS data downsampled to match the read depth observed at the single-cell level. Samples include a matched blood control (top figure), the Parental (second figure), MK1-resistant (third figure), and AZD1-resistant (bottom figure) PDOs. The analysis was performed using 500 kb bins. Grey dots represent the inferred copy number at specific genomic loci. Coloured lines indicate the smoothed median copy number state across genomic segments: black lines denote diploid segments, red lines indicate segments with amplifications, and blue lines represent segments with deletions.

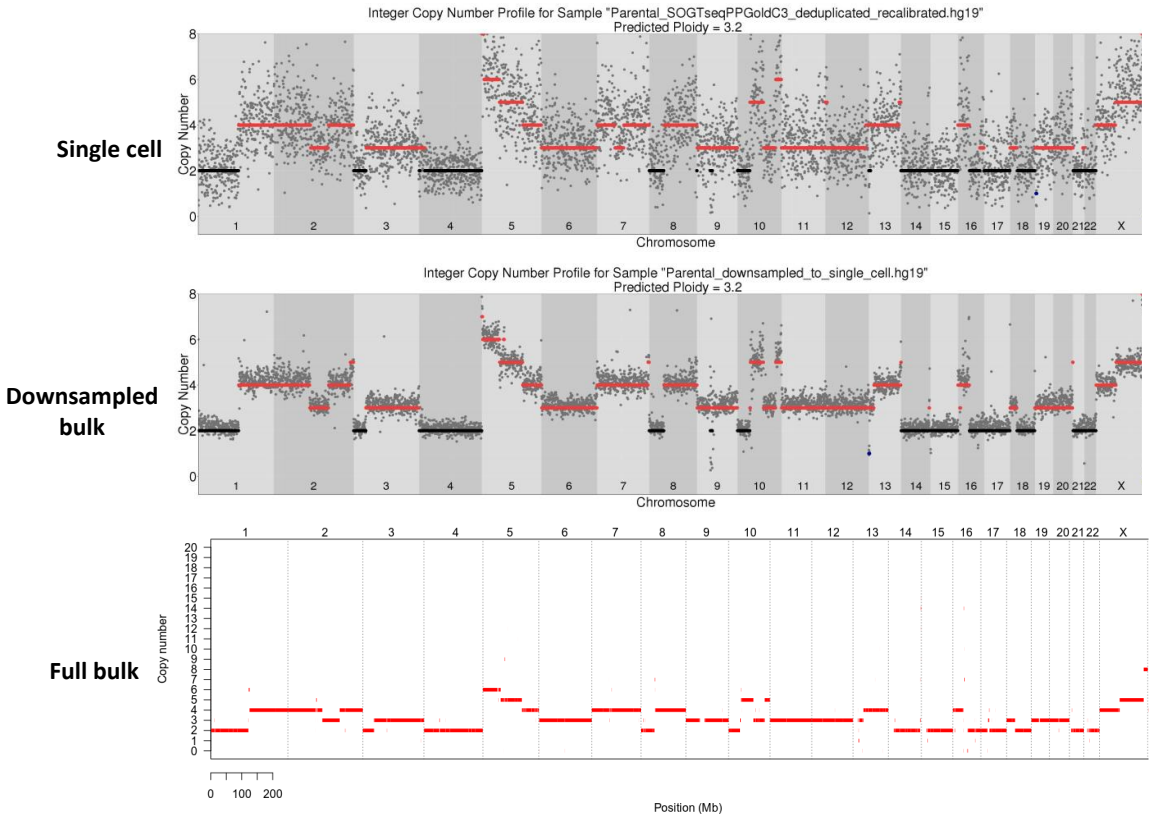


Figure 4.10. Comparative CNA analysis of the Parental PDO via single-cell, downsampled bulk, and full bulk WGS.

The panel figure comprises three distinct CNA profiles. The top and middle figures, derived from Ginkgo analyses, represent the genome-wide CNA states of a single cell and a downsampled bulk dataset, respectively, from the Parental organoid. Binning of the genome for these analyses was performed at 500 kb intervals. Grey dots in these figures depict the inferred copy number at specific loci across the genome. Coloured lines indicate the smoothed median copy number state across genomic segments: black lines denote diploid segments, red lines indicate segments with amplifications, and blue lines mark segments with deletions. The bottom figure, generated by Sequenza, illustrates the CNA profile for the complete bulk WGS data from the Parental organoid with binning performed at 100 kb intervals. In this figure, red segments indicate diploid or aneuploidy copy number regions.

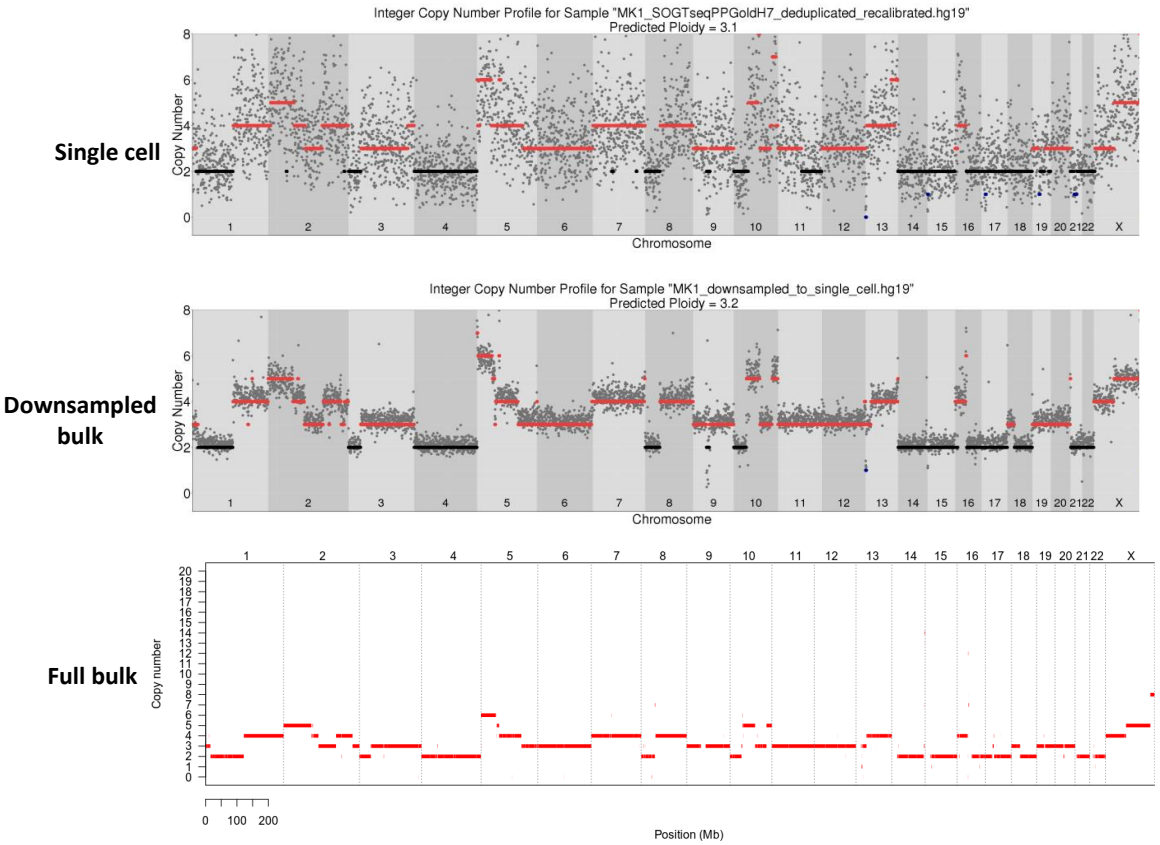


Figure 4.11. Comparative CNA analysis of the MK1-resistant PDO via single-cell, downsampled bulk, and full bulk WGS.

The top and middle figures, derived from Ginkgo analyses, represent the genome-wide CNA states of a single cell and a downsampled bulk dataset, respectively, from the MK1-resistant organoid. Binning of the genome for these analyses was performed at 500 kb intervals. Grey dots in these figures depict the inferred copy number at specific loci across the genome. Coloured lines indicate the smoothed median copy number state across genomic segments: black lines denote diploid segments, red lines indicate segments with amplifications, and blue lines mark segments with deletions. The bottom figure, generated by Sequenza, illustrates the CNA profile for the complete bulk WGS data from the Parental organoid with binning performed at 100 kb intervals. In this figure, red segments indicate diploid or aneuploidy copy number regions.

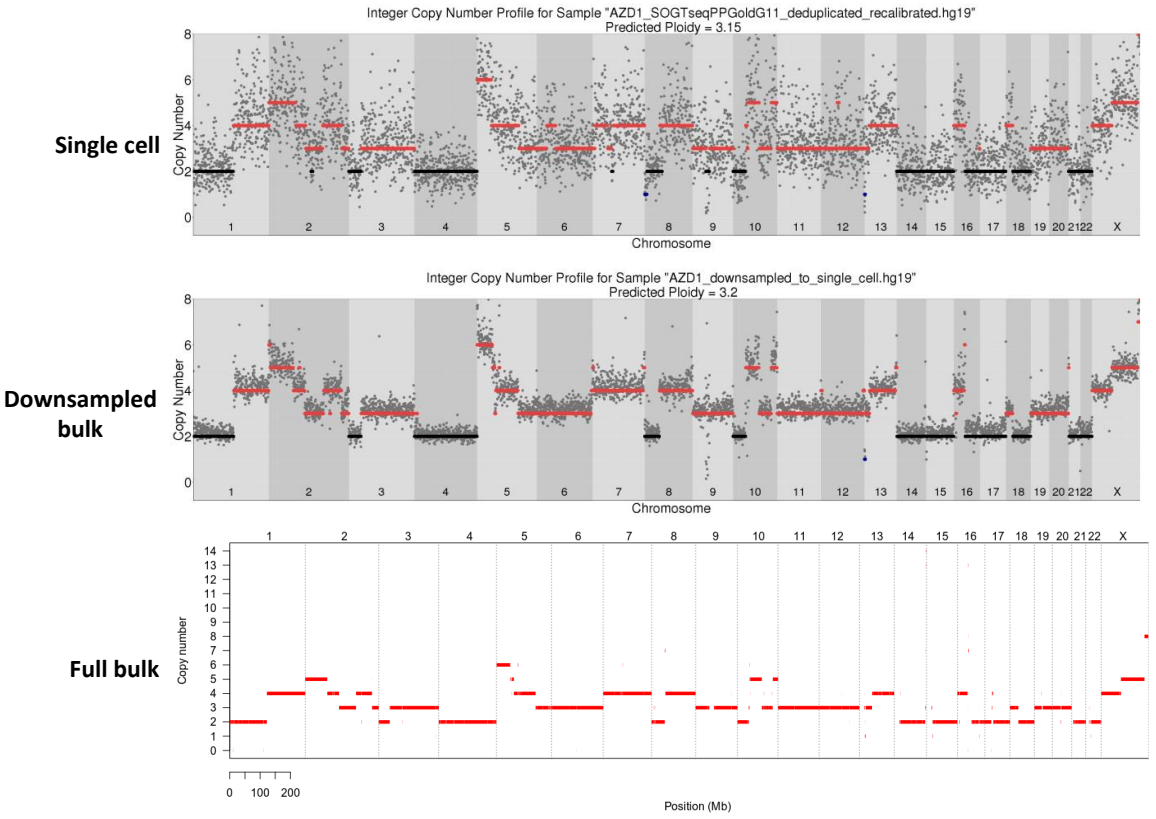


Figure 4.12. Comparative CNA analysis of the AZD1-resistant PDO via single-cell, downsampled bulk, and full bulk WGS.

The top and middle figures, derived from Ginkgo analyses, represent the genome-wide CNA states of a single cell and a downsampled bulk dataset, respectively, from the AZD1-resistant organoid. Binning of the genome for these analyses was performed at 500 kb intervals. Grey dots in these figures depict the inferred copy number at specific loci across the genome. Coloured lines indicate the smoothed median copy number state across genomic segments: black lines denote diploid segments, red lines indicate segments with amplifications, and blue lines mark segments with deletions. The bottom figure, generated by Sequenza, illustrates the CNA profile for the complete bulk WGS data from the Parental organoid with binning performed at 100 kb intervals. In this figure, red segments indicate diploid, or aneuploidy copy number regions.

4.4.5 Exploring the subclonal diversity of mCRC organoids through single-cell CNA analysis

A. Unravelling the subclonal architecture of the Parental mCRC PDO

To explore the subclonal diversity in mCRC PDOs that developed resistance to AKT inhibitors, an initial hierarchical clustering analysis was conducted on the CNA profiles of the Parental organoid at single-cell resolution. This analysis aimed to characterise the clonal composition of this PDO, which served as an untreated control prior to exposure to the MK2206 and AZD5363 AKT inhibitors. Figure 4.13 shows a CNA heatmap for the Parental organoid. The analysis identified CNAs across all chromosomes (chr). The smallest altered region involved a gain of five copies at Xq28 (spanning approximately 1.03 Mb), and the largest was an additional copy of chr6 (~171 Mb). Conversely, the smallest and largest losses included a one-copy deletion at 6q24.3 (~1.13 Mb) and a deletion spanning 15p13-q14 (~37.35 Mb), respectively. Interestingly, the CNA region at 14q32.33, where *AKT1* is located and which was previously reported by (109) to be amplified in the biopsy from which the organoids were derived, exhibited between six and thirteen copies of the gene (Supplementary Figure 16).

The clustered heatmap not only illustrates genomic alterations but also highlights groups of cells based on CNA similarity patterns. Among the Parental cells, there was generally a high degree of similarity in their CNA profiles, with only minor deviations observed. This similarity indicates a close genetic relationship, potentially underscoring their shared lineage origin. At the top level, the root of the dendrogram splits into two branches, each representing a distinct cell population or clone. The predominant clone encompasses 29 cells, while the second, much smaller group consists of just one cell (G3) (Figure 4.13).

Although G3 met the QC standards for mappability and coverage distribution, it exhibited a unique CNA profile compared to other cells (Figure 4.14). With a ploidy of 1.7, lower than the average ploidy previously reported for Parental cells (3.11, ±1.06), G3 was characterised by allelic deletions spanning 4p16.3-q35.2, 7p11.2-q11.21, 16p11.1-q24.3, 17p13.3-q25.3, 18p11.32-q23 and 19p13.3-q13.43. Among the genes impacted by these deletions include *EIF4E* (located at 4q23), a key component in the eukaryotic translation initiation complex (327), genes encoding enzymes involved in drug metabolism such as *UGT2B28* (4q13.2) (369), the *EGFR* (7p11.2) tyrosine kinase receptor, several tumour suppressors such as *WWOX* (16q23.1-q23.2), *AXIN1* (16p13.3), *TP53*, (17p13.1), the Deleted in Colorectal Cancer gene (*DCC*, 18q21.2) and *SMAD4* (18q21), as well as the pro-apoptotic gene *BAX* (19q13.33) (58). On the other hand, while G3 did not exhibit many chromosomal gains, certain noteworthy gains were identified. These included triploid regions at 5q11.2-q23.2 and 10q25.3-q26.2, which

encompass *APC* (5q22.2) and the apoptotic executioner gene *CASP7* (10q25.3), respectively (58). A hexaploid region was also identified at 14q32.33-q32.33, containing *AKT1* (14q32.33).

Focusing on the major clone, further stratification revealed two distinct subpopulations. The smallest subclone, comprising 8 cells (D4 to E3 on the heatmap), accounted for 26.7% of the total Parental genomes sequenced. In contrast, the largest and consequently dominant subclone in the Parental organoid, which includes 21 cells (F1-H1), accounted for 70%. The two subclones contrasted most notably at chromosomes (chr) 5 and X (Figure 4.15). At 5p15.33-q11.2, the minor subclone mostly had five copies, while the major subclone had six. At chrX, the smaller subclone primarily showed tetraploidy at Xq11.2-q27.2, while the larger subclone exhibited pentaploidy or hexaploidy in the same region.

Given the inherent sampling limitations of plate-based single-cell sequencing, there is a possibility that the observed percentages for these subclones might not accurately represent the actual cell populations in the original sample, which consisted of millions of cells. However, bulk WGS analysis, which reflects the genomic profiles of mixed cell populations, allows for the emergence of predominant features. The similarity of the copy number profiles at 5p15.33-q11.2 and Xq11.2-q27.2 in the complete bulk (and downsampled) datasets with those observed in single-cell genomes of the dominant subclone (e.g., C3) confirms its prevalence in the Parental PDO (Figure 4.10).

Upon further inspection of the dominant subclone, it became apparent that a subpopulation within it, consisting of three cells (F1, A3, C2, accounting for 14.3% of the dominant subclone), had a distinct copy number profile not observed in other cells of the same group. These “secondary” subclones, while retaining the six copies characteristic of the dominant group at 5p15.33-q11.2, had four copies of 5q11.2-q23.2 and three copies of 5q23.2-q35.3 (Figure 4.16). In contrast, the “primary” subclone, representing 85.7% of the dominant group, displayed an extra copy in these regions. Additional intra-subclonal heterogeneity among these three cells was also noted, with pentaploid regions identified at 7p22.3-p11.2 (F1), 8q24.22-q24.3 (A3), and 2p25.3-p12 (C2). These cell-specific events highlight the evolving nature of the genomic landscape in the Parental organoid, even before the introduction of AKT inhibitors.

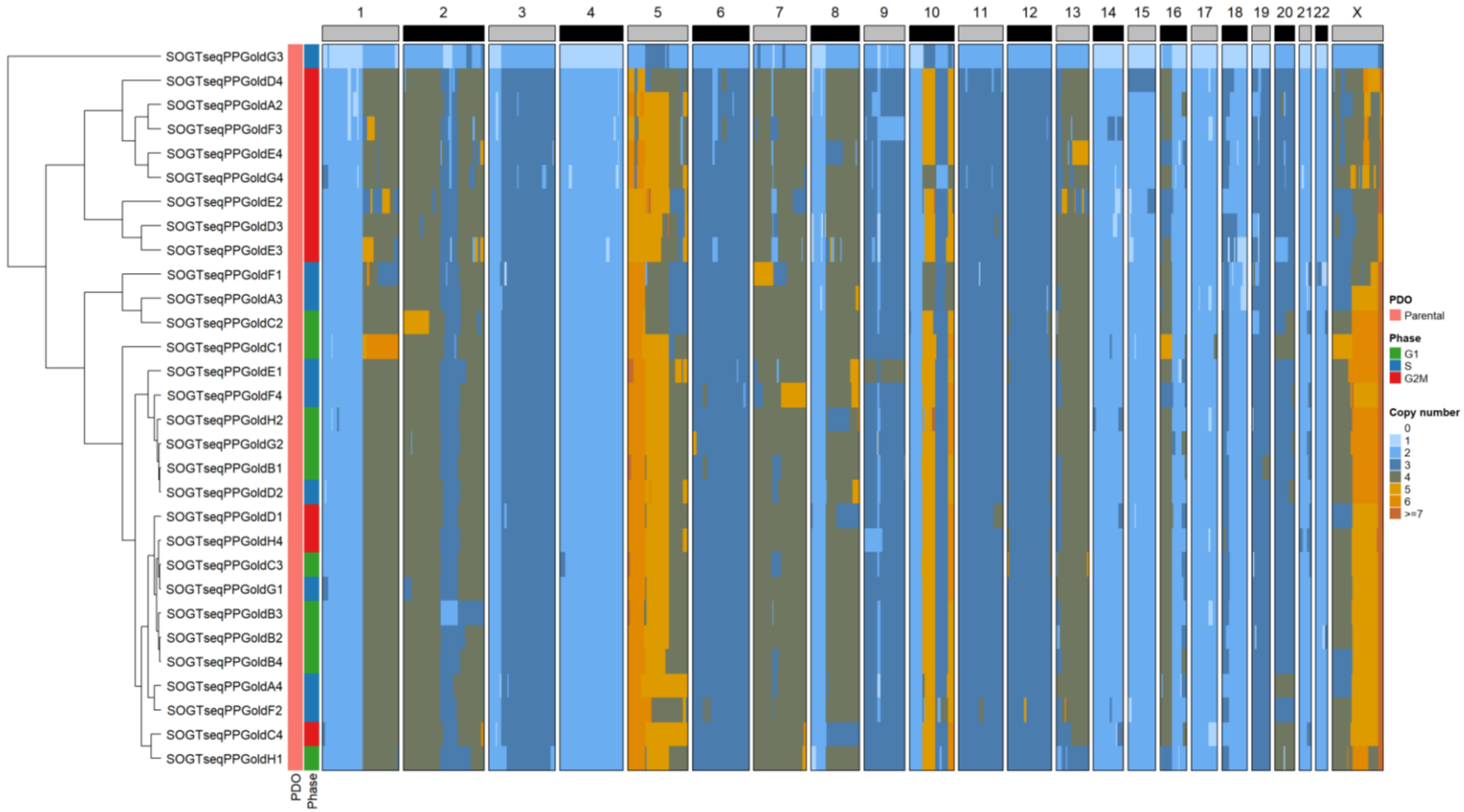


Figure 4.13. Genome-wide CNA heatmap of the genomic profile of Parental mCRC PDO cells

The heatmap provides a clustered, genome-wide representation of copy number profiles from 30 single-cell genomes from the Parental mCRC organoid. Columns represent chromosomes divided into 500 kb bins, and rows correspond to single-cell genomes. Hierarchical clustering of single-cell genomes was performed using Euclidean distance and Ward's linkage method. The colour scale on the heatmap denotes integer copy number states, with additional annotations indicating the cell cycle phase of cells.

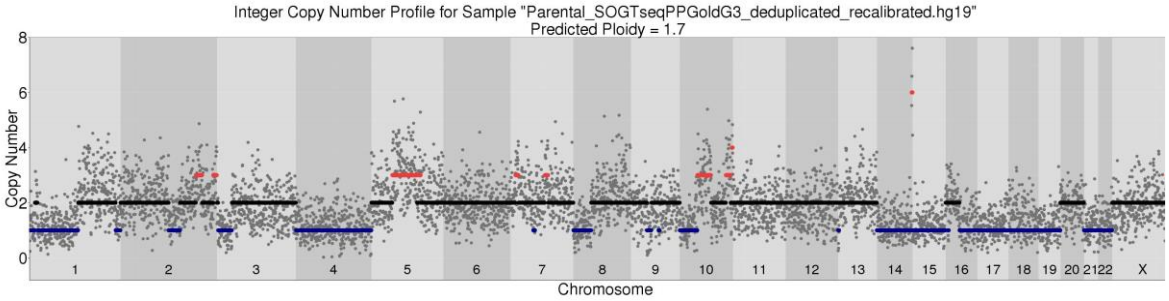


Figure 4.14. Copy number profile of the minor clone identified in the Parental tumoroid.

Genome-wide copy number profile of the minor clone (G3) identified in the Parental mCRC organoid. Coloured horizontal lines represent median copy number states: black for diploid segments, red for amplifications, and blue for deletions.

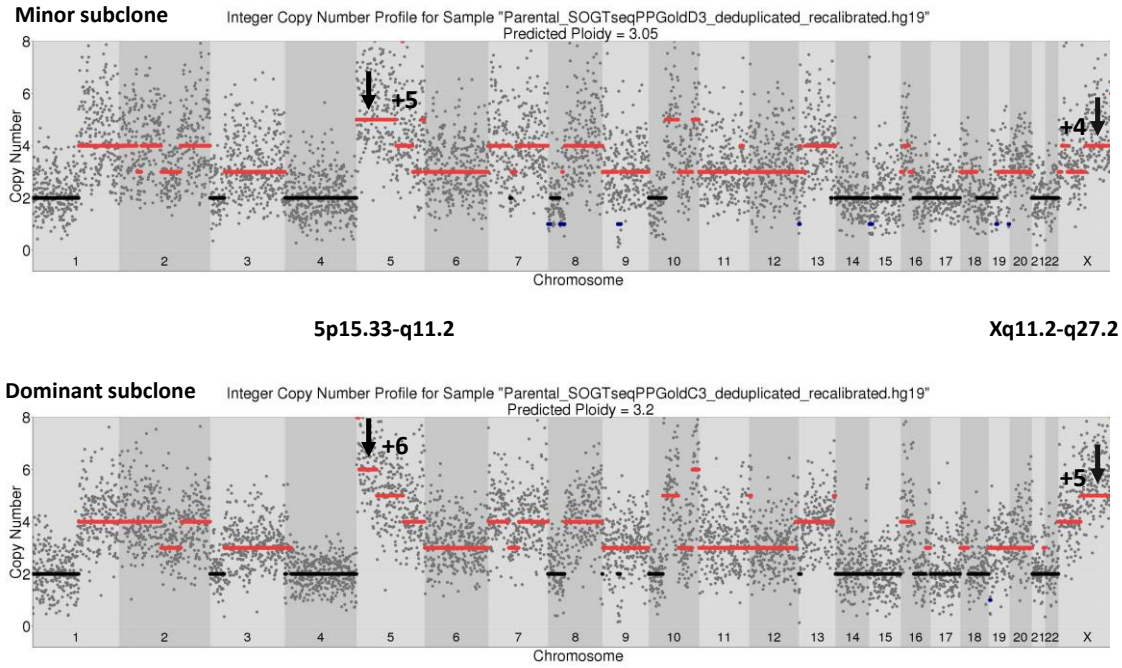


Figure 4.15. Copy number profiles of two representative single-cell genomes from the minor and dominant subclones in the Parental tumoroid.

Plots display genome-wide copy number profiles of representatives single-cell genomes from the minor (D3, top) and dominant (C3, bottom) subclones identified in the Parental mCRC organoid. Coloured arrows indicate regions of CNA gains where the subclones differ at chr5 (p15.33-q11.2) and chrX (q11.2-q27.2), along with the number of copies at each region. Coloured horizontal lines represent median copy number states: black for diploid segments, red for amplifications, and blue for deletions.

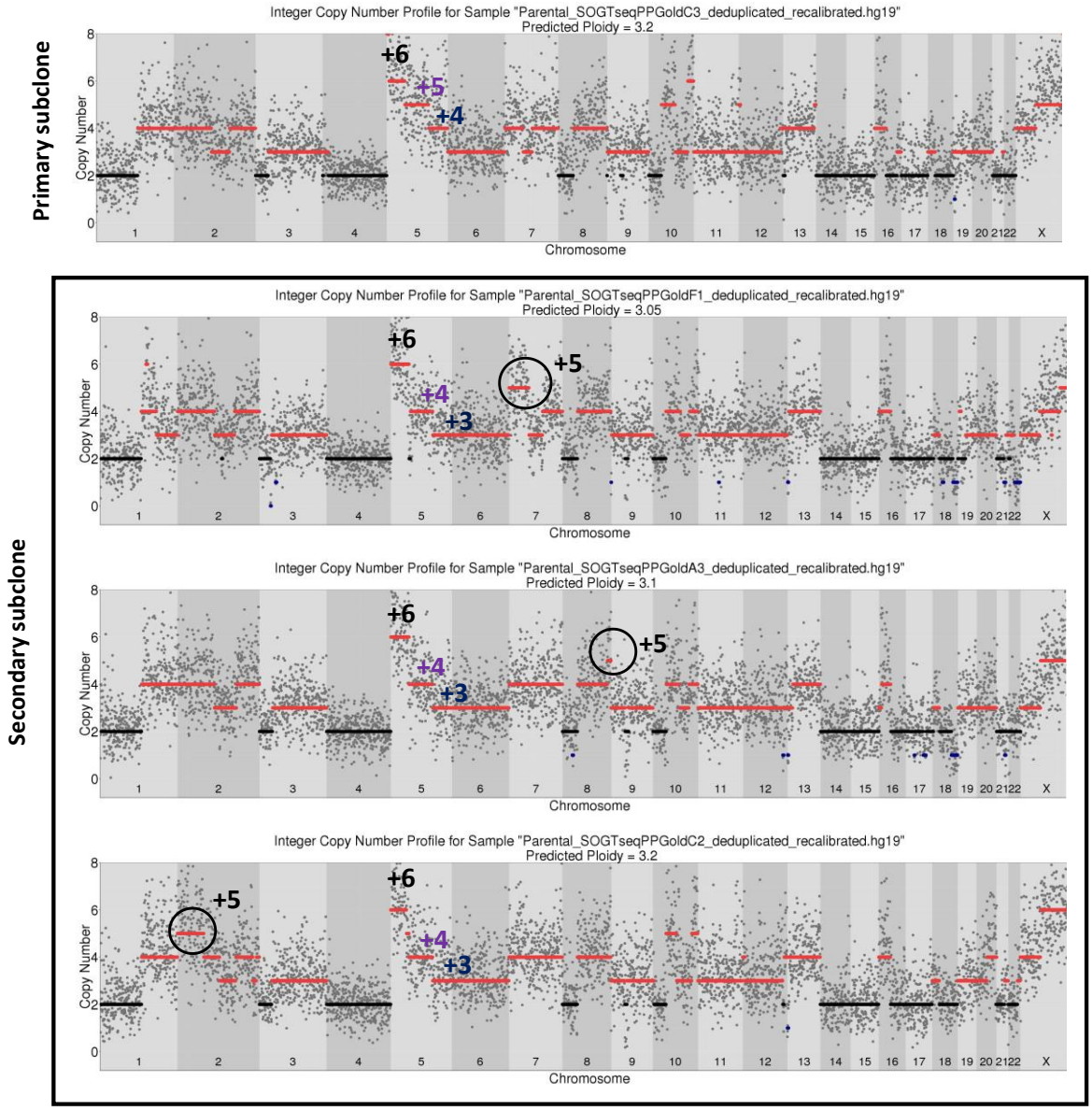


Figure 4.16. Copy number profiles of representative single-cell genomes from the primary and secondary subclones within the dominant group in the Parental tumoroid. *Plots display copy number profiles of representative single-cell genomes from the dominant subclone in the Parental mCRC organoid. Within this dominant subclone, two distinct subclones are identified: the more prevalent "primary" subclone (C3, first plot) and the less prevalent "secondary" subclone (F1, A3, and C2, remaining plots). Coloured annotations on chr5 highlight regions where the primary and secondary subclones within the dominant group converge (black) or diverge (blue and purple). Circled segments indicate CNAs unique to each cell in the secondary subclone. Coloured horizontal lines represent median copy number states: black for diploid segments, red for amplifications, and blue for deletions.*

B. Copy number evolution and subclonal expansion in mCRC organoids resistant to AKT inhibitors

Figure 4.17 presents a genome-wide clustered heatmap showing the copy number states of all MK1- and AZD1-resistant cells alongside the Parental control. In the heatmap, the dendrogram originates from a single root, indicating the collective genomic similarity of the entire dataset. This root splits into two primary branches, representing the first major division in the data. The smaller branch is comprised of three MK1-resistant cells in the G2M phase (B7, B8, and F6), each with noisy copy number profiles that diverge significantly from the main clonal group, as highlighted by the large distance between the clusters in the dendrogram. As a result, these outlier cells were excluded from subsequent analysis.

The predominant clonal group, comprising 77 cells (30 Parental, 19 MK1-resistant, and 28 AZD1-resistant), further subdivides into two branches: a smaller one with one Parental cell (G3) and two MK1-resistant cells (B6, A5) and a larger branch of 74 cells. The wide bifurcation between these subpopulations reflects their distinct copy number profiles. Notably, the smaller subclonal group had the lowest ploidy values in the dataset. In the Parental organoid, G3 was previously characterised as a "nearly diploid" cell with a ploidy of 1.7. B6 and A5 exhibited similar trends with ploidies of 2.2 and 1.85, respectively (Supplementary Figure 17). While these MK1-resistant cells share several similarities with G3, such as one-copy losses at 1p36.13-12, chr4, chr14, and chr22, as well as various diploid chromosomes (e.g., chr6), they also displayed multiple amplifications compared to the Parental cell. These include three copies at 2p25.3-q13 (compared to the two copies in G3), four copies at 5p14.1-q11.2 (two copies in G3), three copies of chr7 (two copies in G3), and three copies at 16p13.3-p11.2 (two copies in G3).

Regarding the larger clonal group, it is further split into two branches. The smaller branch, consisting of three cells with noisy copy number profiles (F10, B12, G5), was excluded from further analyses. The larger branch then bifurcates into two subclusters: the top cluster above the midline on the heatmap consists of 27 cells, while the bottom cluster comprises 44 cells. The similarities in copy number profiles observed across most chromosomes highlight the resemblance between these two subclones. However, a significant proportion of the Parental cells (25 cells, or 83.3% of the total Parental cells sequenced) are grouped in the top cluster. The Parental cells in this top cluster encompass the minor and dominant subclones previously identified in the Parental organoid, which diverged primarily at 5p15.33-q11.2 and Xq11.2-q27.2 (Figure 4.15). This close clustering within the collective heatmap suggests that despite their internal differences, the Parental cells share more similarities with each other than with the majority of AKTi-resistant cells. The latter cells predominantly comprise the bottom cluster,

with the exception of two MK1-resistant cells also exhibiting the characteristic CNAs at 5p15.33-q11.2 and/or Xq11.2-q27.2. These two MK1-resistant cells likely originated from the Parental cells within this top cluster.

Although the two subclusters share many similarities, their genetic differences become apparent at chr5 and chr2. Previously, within the dominant subclone of the Parental PDO, a secondary subclonal population consisting of cells F1, A3, and C2 was distinguishable from the primary, dominant subclone due to their unique genomic characteristics: four copies of 5q11.2-q23.2 and three copies of 5q23.2-q35.3, instead of the extra copy that the primary subclone had in both regions (Figure 4.16). These three cells are clustered at the bottom of the heatmap, interspersed among the MK1- and AZD1-resistant cells (63.6% and 92.9% of cells sequenced, respectively), which also exhibit these CNAs. Furthermore, each of these three cells has specific CNAs unique to them. Notably, C2 is the only cell with five copies at 2p25.3-p12, a CNA shared by all resistant cells in the bottom subcluster (Figure 4.18). The specific genomic signature of C2 mirrored in the resistant lines suggests that the subclone represented by C2 may be a resistant population that expanded to become the predominant subclone in the resistant organoids.

Furthermore, both AKTi-resistant lines exhibit PDO-specific CNAs. These CNAs are highlighted within black boxes in the clustered heatmap (Figure 4.17). For instance, only 2 Parental cells (6.7%, corresponding to cells G1, G4) in the entire dataset have three copies at 1p36.33-p36.13. However, this specific CNA was predominantly observed in 8 MK1-resistant cells (36.4%, corresponding to cells A7, H5, F7, G6, G8, D8, B8, C5) and 1 AZD1-resistant cell (3.6%, A10). This CNA is absent in C2. Instead, it is exclusively present in cells from the primary subclone within the Parental organoid. Similarly, the bottom subcluster further splits into two branches. The most notable one consisted of a subset of 10 AZD1-resistant cells (35.7%, C11, D11, E10, G11, G12, H8, C10, E11, H10, B10) with four copies at 18p11.32-q12.1. In contrast, the remaining cells in the bottom and top clusters exhibit three copies in this region. These CNAs not only reflect the genomic differences that result from different types of AKT inhibition but also hint at the emergence of distinct subclonal populations in the resistant organoids. This is likely a response to the selective pressures exerted by the AKT inhibitors, highlighting the dynamic adaptation and evolution of tumours under these anti-cancer treatments.

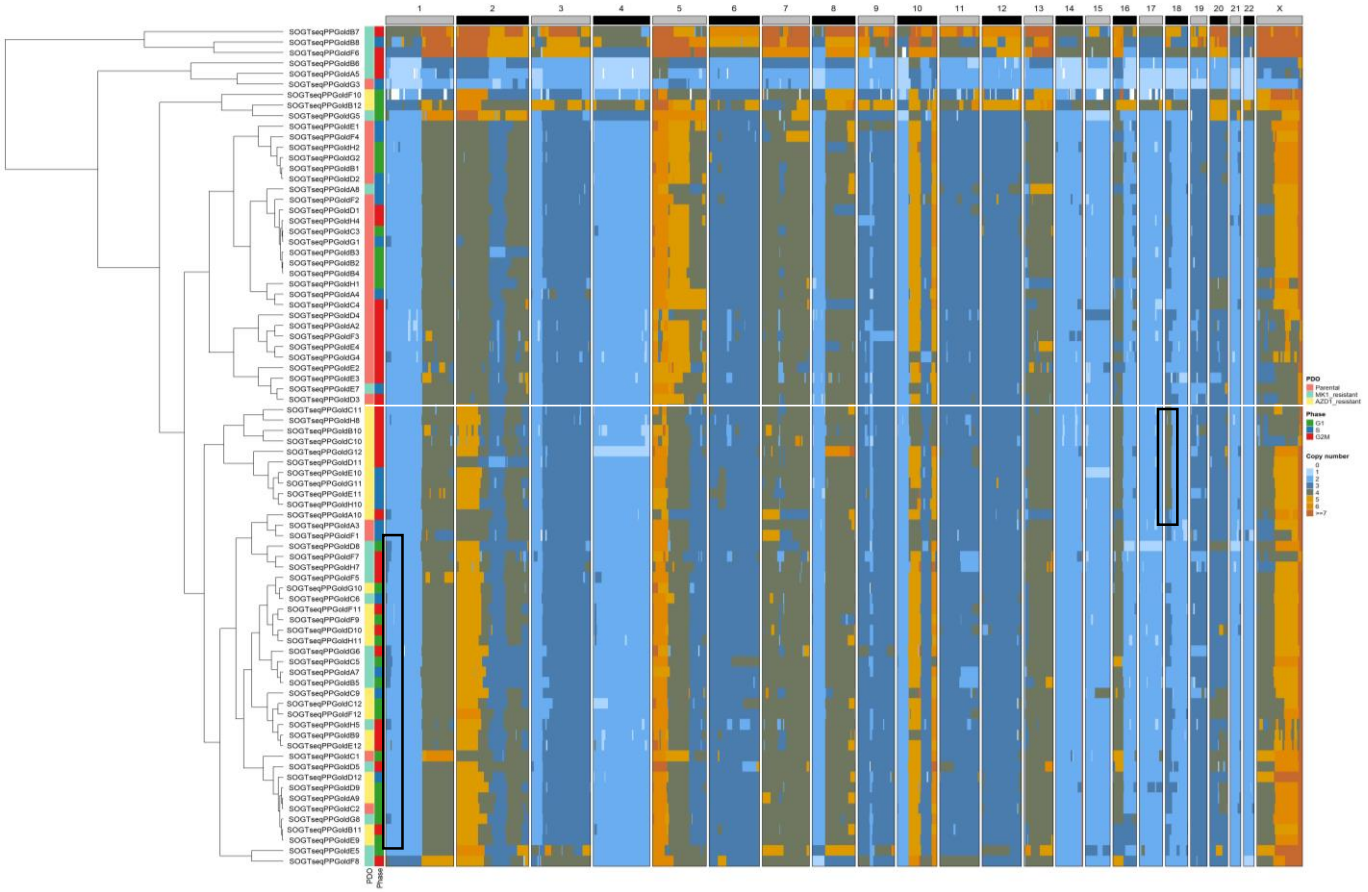


Figure 4.17. Genome-wide CNA heatmaps of mCRC PDOs

Clustered heatmap depicts genome-wide copy number states of 80 single-cell genomes. Columns represent chromosomes segmented into 500 kb bins, rows correspond to single-cell genomes from Parental (n=30), MK1-resistant (n=22), and AZD1-resistant (n=28) organoids. Black boxes represent CNAs predominant in AKTi-resistant organoids. Hierarchical clustering of the genomes was performed using Euclidean distance and Ward's linkage method. The colour scale indicates integer copy number states.

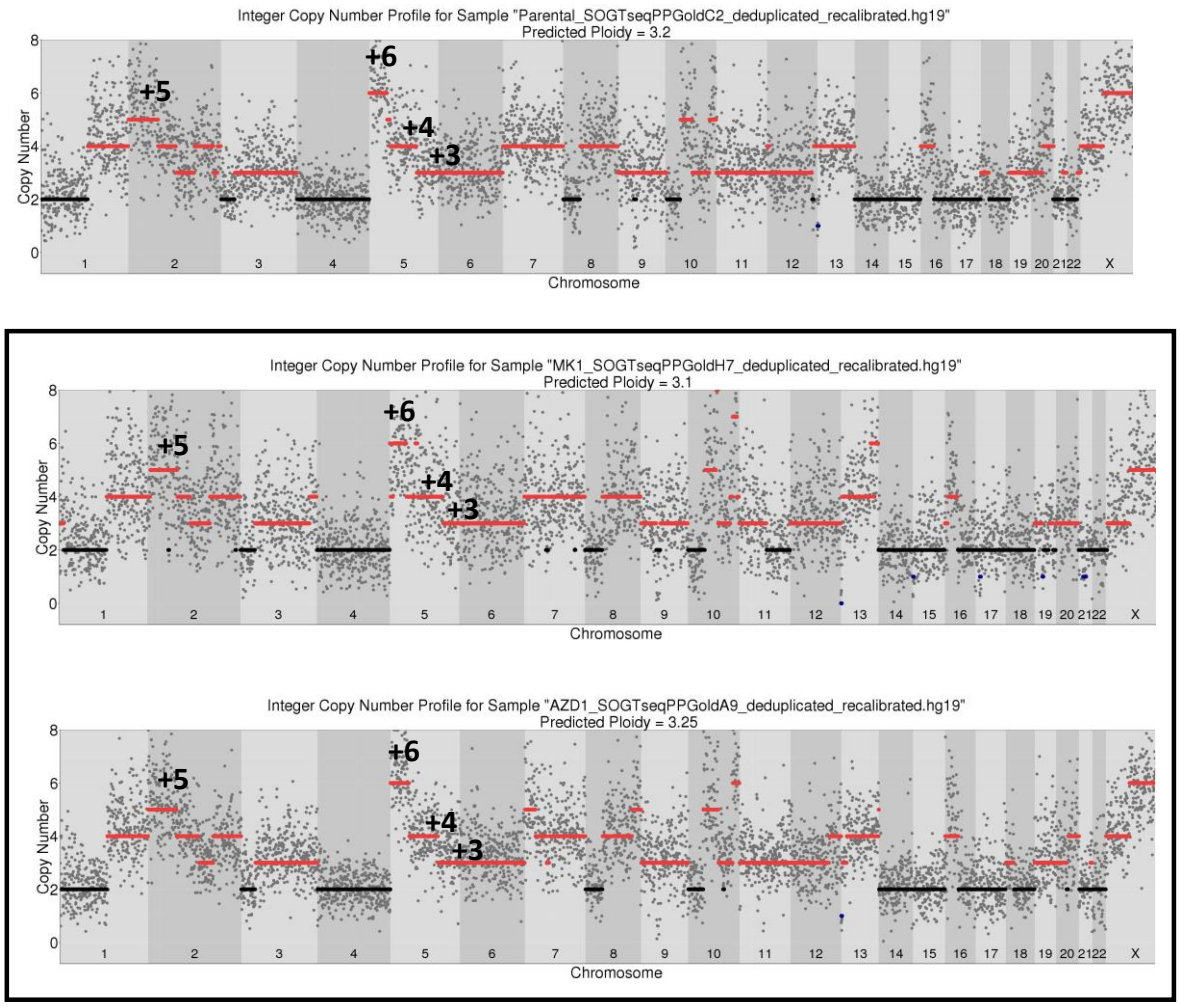


Figure 4.18. Genome-wide copy number profiles of Parental cell C2 and representative AKTi-resistant Cells.

Plots display genome-wide CNAs for cell C2 from the Parental organoid (top) and two representative cells from the MK1- and AZD1-resistant PDOs (plots inside black box). The shared CNAs highlighted in the profiles indicate a potential lineage relationship, suggesting that these resistant cells may have evolved from C2 or a similar subclone. Coloured horizontal lines represent median copy number states: black for diploid segments, red for amplifications, and blue for deletions.

4.5 Discussion

Single-cell whole-genome sequencing (scWGS) provides a more detailed view of the genomic landscape within tumours at the individual cell level, revealing subtleties and variations that traditional bulk sequencing approaches often miss. However, the limited amount of starting DNA material in a single cell (approximately 6 pg) makes whole genome amplification (WGA) a necessary step during sample processing (370), a step that can introduce biochemical biases. For instance, preferential amplification of certain genomic regions may lead to an uneven distribution of sequencing reads (370). This non-uniform read distribution can significantly impact the accuracy of copy number alteration (CNA) detection and ploidy estimation, as these analyses largely depend on measuring read depth of coverage to identify genomic alterations (358, 370). Therefore, selecting the most appropriate WGA method is crucial when the goal is to conduct CNA analysis in single-cell genomes.

The initial sections of this chapter were dedicated to evaluating the suitability of 96 PicoPlex Gold-amplified single-cell genomic libraries derived from the Parental, MK1-, and AZD1-resistant mCRC organoids for CNA analysis. Single-cell genomic libraries exhibited considerable variability in total read counts, which affected the depth of read coverage and inevitably introduced sample-to-sample variability. The observed variability in the data might be due to differences in gDNA extraction efficiency, variations in gDNA content among cells, or, more likely, the lack of a molarity (concentration) normalisation step in the PicoPlex Gold protocol. In this protocol, newly synthesised DNA libraries are pooled by volume before purification instead of individually purified and then pooled based on molarity.

Despite the considerations stated above, the PicoPlex Gold libraries demonstrated high mapping rates and low levels of duplicate reads. These metrics reflect an efficient use of the sequencing data, as most of the reads were retained for downstream analyses. Additionally, while the average depth of coverage was low (<0.3×)—as is typical in scWGS (342)—, it remained adequate for CNA detection using Ginkgo (342). In fact, Ginkgo has demonstrated its ability to detect CNAs in gDNA libraries preamplified with MALBAC (334) and DOP-PCR (328-332), even at shallow depths of coverage (<0.15×). Theoretically, Ginkgo could also infer CNAs with data downsampled to as low as 0.01× coverage (342). This highlights the robustness of Ginkgo in handling low-coverage single-cell genomic data for CNA analysis.

4.5. Discussion

For assessing read distribution across genomic segments or bins, the index of dispersion (IOD) and median absolute deviation (MAD) proved to be very useful metrics. PicoPlex Gold-amplified libraries were found to be underdispersed relative to the Poisson expectation (363). On the other hand, the low average MAD score observed aligns with several studies that have shown standard PicoPlex to outperform other WGA methods, including MALBAC (334), PTA (335), and droplet MDA (333), in terms of amplification evenness. These findings indicate a higher degree of coverage uniformity across bins, which is beneficial for CNA analysis as it implies reduced technical variability, thereby enhancing the reliability of detecting true biological signals.

The scCNA analysis using 500 kb windows revealed genome-wide aneuploidies in the Parental, MK1-, and AZD1-resistant organoids, with average ploidy values ranging from 3.11 to 3.38. Validation of these ploidies was performed in two ways: First, the WGS data from the bulk organoids (including a matched blood control from the donor) were downsampled to single-cell read depths, employing the same binning approach used for scCNA analysis. The second form of validation involved CNA analysis of the complete bulk dataset using 100 kb windows. In particular, the ploidy estimates reported by the full bulk dataset were slightly higher than those derived from the single-cell average ploidies (3.35-3.45). This discrepancy might stem from the inherent characteristics of bulk sequencing, which averages the signal across thousands to millions of cells, thereby masking individual cellular variations (371). Nevertheless, the copy number profiles across bulk datasets closely matched those observed in the single-cell genomes. Moreover, they were generally less noisy, owing to the absence of DNA preamplification. The consistent triploidy in all organoids, evident at both single-cell and bulk levels, suggests a potential whole-genome doubling event (WGD). This phenomenon is frequently observed in human cancers and is often associated with poor prognosis in CRC (372).

The 500 kb window size used in the scCNA analysis of mCRC organoids, compared to the smaller window (100 kb) used in the analysis of the complete bulk data, leads to some implications: On one hand, employing a narrow window when the coverage is low results in many genomic windows having no sequencing reads at all (368), which increases the noise level in the data analysis. Consequently, it becomes difficult to distinguish between real CNAs and random fluctuations in the sequencing data, leading to decreased accuracy and reliability in copy number calls. On the other hand, using a wider window size, which effectively encompasses more reads, helps mitigate the noise and inter-bin variability typical of low-coverage data. However, this approach might lead to missing smaller or focal CNAs, as these get smoothed out in larger windows (368). This “smoothing” effect occurs because smaller

alterations get lost within the average read depth of the larger window, thus potentially lowering the sensitivity needed to detect small but significant CNAs. Although these small-scale variations might occur in only a subset of cells, they could offer critical insights into the cellular heterogeneity within a tumour, which could be important for understanding the underlying mechanisms of disease progression and resistance. In essence, the smoothing effect from larger window sizes could lead to an underestimation of the genomic complexity in tumours, undermining one of the key benefits of single-cell sequencing, i.e., the ability to reveal detailed genomic variations at the individual cell level.

This situation highlights the existing trade-off between selecting the most optimum window size and the capacity to detect smaller CNAs in single-cell sequencing. It underscores the need for scWGA methods to improve coverage uniformity while maintaining the ability to detect smaller-scale genomic alterations. Nevertheless, despite these differences in window size, the fact that similar CNA patterns were observed in both single-cell and bulk WGS analyses suggests that although window size impacts the resolution and sensitivity of CNA detection, the broader patterns of genomic alterations can still be discerned across different sequencing methodologies. This finding underscores the potential for scWGS, even with its current limitations in resolution, to match the findings obtained from the more traditional bulk WGS approach.

A preliminary characterisation of the clonal architecture in the Parental organoid, conducted before deconvoluting the clonal composition of the AKTi-resistant lines, uncovered the presence of two distinct clones. The major clone, comprising 29 out of 30 cells in the Parental PDO, presented genome-wide copy number gains of chromosome arms or large portions of chromosomes. This finding aligns with the chromosomal instability (CIN) characterising colorectal cancers (373). This clone was further stratified into two subclones, diverging due to copy number gains at 5p15.33-q11.2 and Xq11.2-q27.2. The more prevalent or dominant of these subclones, comprising 21/30 cells, included a primary subclone (19 out of 21 dominant cells) characterised by the presence of five copies at 5q11.2-q23.2 and four copies at 5q23.2-q35.3. This contrasted with the secondary subclone (3/21 cells), which exhibited four and three copies in these respective regions, suggesting a trend toward more extensive genomic amplification within this dominant group. Furthermore, the presence of cells such as C2 in the secondary subclone, which was the only cell that exhibited five copies at 2p25.3-p12, not only indicates ongoing genomic diversification within the dominant subclone itself but also highlights the inherent heterogeneity in the genetic makeup of mCRC cells even before treatment.

4.5. Discussion

The scarcity of the subclone represented by C2 made the 2p25.3-p12 event undetectable at the bulk level in the Parental organoid (Supplementary Figure 18). However, C2 emerged as the primary subclone in the MK1- (14/22 cells) and AZD1-resistant (26/28 cells) lines. This observation suggests that prolonged exposure of mCRC organoids to the MK-2206 and AZD5363 inhibitors led to the near extinction of cells sensitive to the treatments, which were predominant before introducing the AKT inhibitors. Concurrently, this environment facilitated the expansion of pre-existing, albeit initially minor, treatment-tolerant cells like C2. In this model of intrinsic (primary) drug resistance (374), the CNA at 2p25.3-p12 was pre-existing in the Parental organoid; thus, it is likely that this chromosomal alteration conferred a selective advantage under drug pressure, contributing to the expansion of C2 in both resistant lines despite the use of different AKT inhibitors. Moreover, while the exact clone expanded in both AKTi-resistant organoids, it also developed treatment-specific CNAs. Notably, there were three copies at 1p36.33-p36.13 in 11 out of 22 MK1-resistant cells and four copies at 18p11.32-q12 in 10/28 cells in the AZD1-resistant line. These specific CNAs further underscore the subclone's adaptability to different treatment environments.

Resistance in AKTi-resistant PDOs likely originated from a pre-existing Parental cell carrying specific CNAs at chromosomes 2 and 5 that expanded under MK-2206 and AZD5363 treatments. Additional CNAs, such as those on chromosomes 1 and 18, likely provided further adaptive advantages under the different AKT inhibitors. This is supported by the observation that the CNAs on chromosomes 1 and 18 do not appear together in either of the AKTi-resistant organoids, suggesting that these CNAs likely arose to adapt to the specific treatments. Furthermore, the presence of the chromosome 1 CNA in the majority of MK1-resistant cells suggests that this alteration, although observed in only a small minority of Parental cells, was crucial for resistance to MK-2206. This indicates that this CNA on chromosome 1 was very important, even if insufficient alone, for resistance. In contrast, the presence of the chromosome 18 CNA in a smaller number of AZD1-resistant cells suggests that it was not as necessary for resistance to AZD5363, and might be more of a passenger CNA rather than a functional one.

The expansion of inherently resistant, pre-existing clonal populations with specific CNAs in resistant cancers, as observed in the current research, has been documented before in the literature. To investigate whether resistance to radiation therapy in rectal cancer was driven by pre-existing genetic factors or acquired during treatment, Andel *et al*., established PDOs from rectal cancer samples and exposed them to radiation (375). Using scWGS and targeted genotyping before and after radiation, they tracked subclonal evolution in response to radiation. The authors observed three distinct patterns: subclonal persistence, extinction, or

expansion, but rarely the emergence of new, shared genomic aberrations. Interestingly, radiosensitive subclones exhibited copy-number patterns indicative of mitotic segregation errors, suggesting that chromosomal instability may be a marker of sensitivity. Furthermore, they found that specific copy-number alterations, such as amplifications of certain oncogenes (*JAK2*, *FANCG*, *FANCF*, *DDB2*, *AR*) and deletions of tumour suppressors (*PTPN13, AFF1)*, were associated with radioresistance. These findings suggest that resistance to radiation therapy is largely determined by pre-existing radioresistant subclones that persist or expand rather than being newly created. The authors propose that this inherent resistance could potentially be predicted by analysing pre-treatment biopsies.

The presence of inherent resistance has also been observed in blood cancers such as acute myeloid leukaemia (AML). Ding *et al*., used bulk WGS and deep sequencing to comprehensively analyse the mutational landscape of AML relapse in eight patients (376). Two main patterns of clonal evolution were observed at relapse in these patients: either the founding clone present at diagnosis acquired additional mutations, or a minor subclone survived initial therapy and expanded. In one patient, they observed the founding clone harbouring mutations commonly found in AML, e.g., *DNMT3A* (a gene involved in DNA methylation) and *NPM1* (a protein involved in various cellular processes) (377), giving rise to a subclone that gained further mutations in other genes that affect the progression of AML, such as *ETV6* (a transcription factor involved in haematopoiesis) (378), ultimately leading to relapse. A key observation in this study was the increase in specific types of mutations, particularly transversions, in relapsed AML compared to the primary tumours. This suggests that chemotherapy, while crucial for initial remission, might contribute to relapse by inducing DNA damage and generating new mutations that drive clonal evolution (376).

These studies emphasise the critical role of clonal evolution and pre-existing genetic heterogeneity in driving resistance to therapy. The presence of specific genetic alterations, whether CNAs in CRC or mutations in AML, can confer a survival advantage to certain subclones, allowing them to persist or expand in the face of treatment. This shared principle underscores the importance of developing therapeutic approaches that can effectively target these resistant subpopulations to improve patient outcomes.

Focusing on the current research, the uniform response to the MK-2206 and AZD5363 inhibitors—despite their distinct mechanisms of action but shared target in the AKT pathway—suggests a resistance linked to the subclone's ability to adapt to or tolerate the effects of any treatment interfering with the AKT pathway, not just a specific drug. The role of the AKT pathway in mediating multidrug resistance (MDR) has been extensively studied and

documented in previous research (379). However, this pathway alone is not solely responsible for MDR. It is often accompanied by the transduction of upstream and downstream targets, which involve the modulation of apoptosis, cell growth, and cellular metabolism, all of which are associated with the mechanisms of MDR (379).

A high frequency of chromosomal gains suggests the presence of oncogenes favouring cell growth and survival (9). Although several gains involving many proto-oncogenes were identified in resistant mCRC cells (e.g., *APC*, *EGFR*, *KRAS*, *WNT2*), the anaplastic lymphoma kinase (*ALK*) gene, located at 2p23.2-p23.1, is an important proto-oncogene that lies within the 2p25.3-p12 amplicon. As a tyrosine kinase receptor, ALK activates several signalling pathways, including PI3K/AKT/mTOR, RAS/ERK and JAK/STAT (380). When dysregulated, these pathways can affect cell proliferation, migration, inhibition of apoptosis, and angiogenesis. In human cancers, constitutive activation of ALK has been identified through various mechanisms, including genomic amplification, chromosome translocations, or point mutations, with the most well-known being chromosomal translocations leading to the ALK-EML4 fusion in non-small cell lung cancer (NSCLC) (381). In a 2013 screening of 756 CRC samples from Saudi Arabia, an increase in ALK copy number gain or amplification was discovered in 3.4% of the samples (381). This genetic alteration was associated with poorly differentiated neoplasms, indicative of more aggressive cancers, and was linked to a worse prognosis, regardless of CRC stage or microsatellite instability. With their findings, Bavi *et al*., suggested that alterations in *ALK* alone were correlated with a poorer prognosis in this subtype of CRC, even when considering other variables. Besides *ALK*, *ENO1* at 1p36.23 and *PGD* at 1p36.22 in the MK1-resistant organoid are also noteworthy. These genes are both located within the 1p36.33-p36.13 CNA specific to the MK1-resistant organoid and were upregulated in this resistant line compared to the untreated control (Figure 3.18). This observation hints at a potential link between genes located within CNA regions and the impact of these variations on their expression levels.

With respect to the minor clone identified in the entire dataset, represented by G3 in the Parental dataset, it was characterised by a "nearly diploid" genome (ploidy = 1.7). Interestingly, this clone was also represented in 2 out of 22 MK1-resistant cells, which displayed similar ploidies and CNAs. Of particular interest was a shared deletion of chr4, encompassing multiple tumour suppressors, including the gene encoding the F-box protein *FBXW7* (4q31), a crucial component of the SCF (SKP1-CUL1-F-box protein) complex. This complex is a part of the really interesting new gene (RING) family of E3 ubiquitin ligases (382). Within the E3 complex, F-box proteins are crucial for recognising protein targets for ubiquitin-mediated proteasomal degradation. In this way, F-box proteins act as tumour suppressors by targeting the degradation

of approximately 20% of proteins involved in cell-cycle regulation, transcription, and apoptosis, including key proteins such as cyclin E, c-Jun, c-MYC, Notch, MCL-1, p53 and mTOR (382). Consequently, the inactivation of F-box proteins through genomic deletions, mutations, or promoter hypermethylation leads to the accumulation of these oncoproteins, disrupting multiple downstream signalling pathways and leading to uncontrolled cell proliferation (382). Indeed, loss-of-function mutations in *FBXW7* are commonly observed in various cancers. These include approximately 30% of cholangiocarcinomas and T-cell acute lymphoblastic leukaemias (T-ALL), as well as 4-15% of cases in pancreatic, gastric, and colon carcinomas, in addition to prostate and endometrial cancers (383). Moreover, in CRC, lower expression of *FBXW7* has been associated with a lower 5-year survival rate compared to cases with higher expression levels (384).

While the copy number profile of the "nearly diploid" MK1-resistant cells derived from G3 matched that of their progenitor, these cells exhibited additional amplifications in regions where G3 maintained a diploid state or had fewer copies. For example, the three copies gained at 2p25.3-q13 affect the *ALK* gene at 2p23.2-p23.1. Besides *ALK*, various other gains identified in the MK1-resistant cells were observed affecting chromosomes that harbour proto-oncogenes (*EGFR* and *MET* on chr7), signalling molecules (*WNT2* on chr7), genes involved in cell adhesion and migration (*PODXL* on chr7), and transcription factors (*MYC* on chr8). All these genes are known to play roles in CRC's development and progression (58).

The distinct copy number profile observed in these "nearly diploid" genomes, as opposed to the average triploidy reported for all PDOs, raises several possibilities. One hypothesis is that these cells might represent contamination from less-transformed cancer cells within the Parental and MK1-resistant mCRC organoid cultures. This theory is supported by the presence of chromosomal changes typically reported in the early stages of primary colorectal carcinomas, such as deletions at 1p and 18p and gains at 8q and 13q, as well as alterations characteristic of later stages, such as the loss of 4p (59) (Figure 1.4). Additionally, CNAs contributing to the progression from primary colorectal carcinomas to liver metastases further support this theory, particularly with early-stage changes such as loss of 8p and gain of 7p, and late-stage modifications like loss of 14q and gain of 1q (59). Furthermore, the persistence of these cells (although scarce) in the untreated Parental and MK1-resistant lines suggests they represent a minor yet inherently treatment-resistant cell population. The plate-based G&T-seq approach may not have sufficiently detected this population, which was also indiscernible in the bulk data analysis, highlighting the potential limitations of these methods in identifying minor clonal populations.

4.5. Discussion

Despite the organoids being treatment naïve to AKT inhibitors, the observation of intrinsic drug resistance within the mCRC organoids is not surprising. This is because the donor had received FOLFIRI at the time of sample donation and, prior to that, had been treated with oxaliplatin and capecitabine. Indeed, every instance of treatment failure acts as a selective event, resulting in tumours progressively becoming more aggressive and accumulating complex genomic aberrations within their resistant phenotype (385).

The primary drug resistance observed in mCRC organoids fits well with Darwin's "survival of the fittest" theory of evolution, as it explains why tumours become more aggressive with continuous drug exposure (386). However, a significant limitation of this study is the challenge in discerning the tumour's inherent suitability to withstand AKT inhibition before organoid development. Without directly comparing the mCRC liver specimen to the patient-derived organoids, it is difficult to ascertain which CNAs were pre-existing in the tumour and which may have developed during the organoid culture process. Undoubtedly, some CNAs may be the result of the patient's previous treatments, while others could have emerged during the culture process. A potential solution to this limitation would be to analyse the bulk sequencing data from the original liver specimen. While this approach might not reveal as much heterogeneity as single-cell sequencing, it would at least provide a comprehensive overview of the tumour's genomic landscape prior to the development of the organoids, which would enhance the interpretation of the changes observed in all three organoids. Despite these challenges, the study does shed light on the changes that emerged due to prolonged drug exposure, likely reflecting the tumour's adaptive evolutionary process under therapeutic pressure.

In addition, while hierarchical clustering was used in the heatmap of copy number profiles to group cells based on their CNA similarities, this method alone is insufficient to determine the lineage relationships between cells. Lineage tracing would be more appropriate to truly understand the evolutionary relationship among subclones. This analysis could offer insights into how the subclones diverged over time, enhancing our understanding of the tumour's adaptive mechanisms under drug treatment from an evolutionary perspective.

Other limitations of this study, particularly regarding the sampling constraints of plate-based single-cell WGS, must be acknowledged. It is unclear if the differences in the proportion of resistant subclones are due to the limited number of cells analysed or if they accurately represent the initial tumour condition. The restricted cell numbers sequenced may not provide a comprehensive view of the tumour's heterogeneity, potentially leading to an incomplete understanding of the subclonal landscape and the extent of drug resistance. These limitations highlight the need for a careful approach to interpreting the data. Nevertheless, the

deconvolution of Parental genomes from a single-cell sequencing perspective revealed crucial insights that were not apparent in the bulk data analysis. Indeed, the identification of the 2p12-p25.3 amplicon, obscured in bulk data due to the rarity of the minor subclone in the Parental organoid, underscores the value of single-cell analysis in understanding subclonal heterogeneity. Understanding subclonal diversity in human cancers can provide critical guidance for developing treatment strategies to prevent the emergence of drug-resistant genotypes.

In summary, the copy number analysis of single-cell genomes discussed in this chapter revealed the subclonal heterogeneity in mCRC organoids, characterised by a diverse landscape of genomic alterations, some of which were not detectable using bulk sequencing approaches. Despite overall similarities among Parental cells, distinct subpopulations with unique genomic signatures were identified. This included a predominant clone, which showed further stratification, and a particular subgroup within that expanded in the MK1- and AZD1-resistant organoids. The presence of subclonal heterogeneity and drug resistance even before the start of treatment highlights the challenges of targeting cancer at the genetic level. Understanding these dynamics is crucial for developing effective cancer treatments, specifically those to predict treatment responses. Such an approach could enable the evolutionary steering of the disease towards more clinically treatable phenotypes.

The next chapter will explore the transcriptional impact exerted by the CNAs identified in this chapter. To achieve this, the study will integrate G&T-seq datasets, aiming to provide a more comprehensive understanding of the genomic landscape and its implications on gene expression.

Chapter 5

Transcriptional impact of CNAs on AKTi-resistant mCRC PDOs

5.1 Introduction

Thus far, this thesis has explored the impact of acquired drug resistance to AKT inhibition in mCRC PDOs through distinct analyses performed initially at the transcriptomic level and subsequently at the genomic level. At the gene expression level, the MK1-resistant PDO demonstrated metabolic reprogramming, mainly characterised by the upregulation of glycolysis and related biosynthetic pathways and genes involved in extracellular matrix (ECM) remodelling. In contrast, the AZD1-resistant PDO, while also exhibiting upregulation of ECM-related genes and metabolic genes, was marked by the upregulation of genes involved in several processes, such as the regulation of immune responses, cell surface receptor signalling, cell detoxification and protein folding, degradation, or cleavage.

At the genomic level, the clonal composition of both control and resistant mCRC PDOs was characterised by inspecting genome-wide CNA changes, which helped identify shifts in clonal composition driven by drug resistance. Single-cell CNA analysis revealed a minor subpopulation within the untreated Parental PDO that expanded in the MK1- and AZD1-resistant PDOs. Notably, chromosomal alterations on chromosomes 2 and 5 were prevalent in these subpopulations, with resistance-specific CNAs observed on chromosome 1 predominantly in the MK1-resistant PDO and on chromosome 18 in the AZD1-resistant PDO. These findings prompted an investigation into the relationship between CNAs and gene expression changes in MK1- and AZD1-resistant PDOs to determine whether these structural variants were directly responsible for the transcriptional adaptations leading to drug resistance. By identifying differentially expressed genes within regions affected by CNAs, this chapter aims to elucidate the chromosomal adaptations that allowed mCRC cells to survive AKT inhibition.

Aneuploidy occurs in up to 90% of solid tumours and 75% of blood cancers (387). Aneuploidy can be observed even in benign polyps, highlighting its role in the early stages of CRC development. Indeed, aneuploidy often acts as a catalyst for genetic instability, driving the transformation of tumour cells. This leads to patterns of genomic instability that may result in cell death when the level of genomic chaos becomes unsustainable. For example, conditions like monosomy create significant instability, usually lethal due to the loss of essential genes required for cell survival and proper functioning (387). However, in some cases, aneuploidy enables cells to surpass critical error thresholds, resulting in malignant cells with stable karyotype configurations that facilitate drug resistance and metastasis. The balance between

the destabilising effects of aneuploidy and the stabilising selection pressures favouring oncogenic functions illustrates the adaptive nature of cancer evolution (387).

The gain or loss of chromosomes alters the dosage of numerous genes, i.e., the amount of gene product, affecting the expression levels of various proteins, including those that regulate other genes across different genomic loci (387). This widespread dosage imbalance plays a significant role in furthering cancer progression. Consequently, regulating gene expression through dosage compensation mechanisms is crucial for restoring and maintaining cellular homeostasis in conditions of genomic instability like aneuploidy (387). Given its importance, this chapter also delves into the potential dosage-compensatory mechanisms operating in the context of differentially expressed genes residing within regions of CNA, exploring how these mechanisms might mitigate the disruptive effects of aneuploidy on cellular function and, at the same time, promote drug resistance.

5.2 Aims

The general aims of this chapter are summarised as follows:

1. Perform copy number analysis using expression data to confirm correlations between CNAs identified through genomic analysis and those detected using transcriptomic data.
2. Integrate single-cell genomic and transcriptomic data from mCRC PDOs to determine if CNAs contributed to gene expression profiles associated with drug resistance in AKTi-resistant mCRC PDOs.

The hypothesis behind these aims was that genes differentially expressed and contributing to resistance would be found within CNA regions, particularly on chromosomes 2 and 5. This would suggest a direct link between structural genomic alterations and the transcriptional reprogramming observed in AKTi-resistant mCRC PDOs.

5.3 Methods: Bioinformatics integration of scRNA and scWGS datasets

A list of the software packages and R-based tools employed for the integration of the scRNA-seq and scWGS datasets is provided in Table 16. Default parameters were employed for all computational tools unless stated otherwise in the text.

Table 18. Software packages used for G&T-seq data integration

Software	Access/citation
ComplexHeatmap v2.16.0	(157)
Ensembl BioMart web-based tool	(388)
GenomicRanges v1.52.1	(341)
ggplot v2.3.4.4	(161)
InferCNV v1.16.0	(389)
R v4.1.2	R Core Team (2022)
RStudio v2023.6.2.561	R Core Team (2022)
tidyverse v2.0.0	(176)
UCSC liftOver web-based tool	(345)

5.3. Methods: Bioinformatics integration of scRNA and scWGS datasets

5.3.1 Transcriptome-based DNA copy number inference

InferCNV v1.16.0 was used to infer DNA copy number states instead of integer copy numbers in mCRC organoids, leveraging the single-cell gene expression data presented in Chapter 3. InferCNV was developed by the Trinity Cancer Transcriptome Analysis Toolkit (CTAT) (390), a project that aims to provide bioinformatics tools to address various challenges in cancer transcriptomics.

To perform the copy number variation (CNV) analysis, inferCNV requires three files to create an inferCNV object. These include a raw expression matrix, an annotation file containing sample information, and a gene ordering file. The raw expression matrix was exported from the Seurat object created during the scRNA-seq analysis. This matrix included cells that passed the quality control standards of the scRNA-seq processing pipeline and met Seurat's filtering criteria. The Parental organoid was selected as the "normal" reference in the annotation file to infer the copy number states of AKTi-resistant PDOs. Lastly, the gene ordering file from GENCODE hg38 (version 27) was sourced from the Trinity's CNV repository (391). Importantly, the Y and mitochondrial chromosomes were excluded from the inferCNV object generation step.

The actual copy number analysis was conducted by invoking *infercnv::run* in the "*subclustering*" mode, with the "*denoise*" parameter set to "*TRUE*", and by enabling a Hidden Markov Model (HMM) alongside the default Bayesian latent mixture modelling approach to predict CNV states. Genes with a mean count below 1 were also filtered out by specifying "*cutoff=1*" (set to 0.1 for 10x scRNA-seq data). Finally, "*cluster_by_groups*" was set to "*FALSE*" to cluster subclones based on their CNV similarities at specific chromosomal regions rather than predefined classifications such as PDO of origin.

While inferCNV generates a variety of plots, ComplexHeatmap v2.16.0 and ggplot v2.3.4.4 were employed for visualising its output files for detailed analysis:

- "**HMM_predHMMi6.rand_trees.hmm_mode-subclusters.observations.txt**":
  For the AKTi-resistant organoids, a heatmap of genome-wide CNV states was generated using this output. InferCNV does not generate a corresponding file for the reference sample. Consequently, only CNV states pertaining to the resistant organoids are documented in the "Results" section of this chapter.
- The "**infercnv.references.txt**" and "**infercnv.observations.txt**" files contain normalised and adjusted gene expression matrices. These files were used to assess the impact of CNV

states on gene expression in Parental and AKTi-resistant PDOs, respectively. Additionally, the "**infercnv.observations_dendrogram.txt**" output was used to group resistant cells using hierarchical clustering based on Euclidean distance and average linkage.

- "**HMM_CNV_predictions.HMMi6.rand_trees.hmm_mode-subclusters.Pnorm_0.5.pred_cnv_regions.dat**": This output specifies genomic coordinates containing CNV states. These regions were subsequently mapped to chromosome arms using the UCSC hg38 cytoband data (392). For this purpose, an adapted version of the "*inferCNV-postprocess.r.txt*" script was employed (393, 394).
- The "**HMM_predHMMi6.rand_trees.hmm_mode-subclusters.cell_groupings**" file was used to calculate the frequency of scRNA-seq CNV subclones in AKTi-resistant PDOs. Along with the previous file, it was used to represent the CNV events characterising the subclones identified.
- The "**infercnv.invert_log_FC.observations.txt**" file contains data obtained by subtracting the average fold change of normal cells from that in tumour cells. Thus, this file provided a relative measure of gene expression changes in AKTi-resistant cells. This file and "**the infercnv.invert_log_FC.references.txt**" were used to compare gene expression levels across chromosomes between the Parental and AKTi-resistant cells.
- The "infercnv.observations_dendrogram.txt" output was used to group AKT-resistant cells using hierarchical clustering based on Euclidean distance and average linkage.

5.3. Methods: Bioinformatics integration of scRNA and scWGS datasets

5.3.2 Mapping differentially expressed genes to CNA regions

To determine if copy number alterations (CNAs) in specific genomic regions drove changes in gene expression in AKTi-resistant organoids, the genomic coordinates of differentially expressed genes (DEGs) identified in single-cell transcriptomes were compared to the genomic coordinates of CNAs presented in the scWGS chapter. For this purpose, the Ensembl BioMart web-based tool (395) was employed to find the hg38 genomic coordinates of DEGs. The attributes retrieved included the *gene stable ID*, *gene name*, *chromosome/scaffold name*, *gene start*, and *gene end*. The output was then formatted in R to create a file compatible with the UCSC LiftOver web-based platform (396). Using *LiftOver*, the hg38 genomic coordinates of the DEGs were converted to hg19 coordinates, ensuring they aligned with the hg19 coordinates of genomic CNAs.

Next, GenomicRanges v1.52.1 was used to find overlapping regions between the genomic coordinates of CNAs documented in the "**CNV1.txt**" created during the Ginkgo CNA of scWGS libraries and the genomic coordinates of DEGs. The intersections were then matched with hg19 cytoband information to provide a chromosomal context for these CNAs. With the cytoband information integrated, further information was extracted from the intersections, such as the CNAs coordinates, the genes located within the CNA region, the sample names presenting these structural variants and the copy number at the specific CNA.

The following step in this integrated analysis involved using tidyverse v2.0.0 and custom functions to summarise the cytoband range of CNAs affecting DEGs. This was achieved by identifying all cytoband regions overlapped by the gene's CNA in AKTi-resistant genomes and determining the broadest contiguous cytoband range encompassing the affected gene. This analysis yielded information about the broadest span of CNA cytobands in which DEGs were found, offering insights into the genomic landscape of the CNAs that could contribute to gene expression changes.

5.4 Results

5.4.1 Transcriptome-based DNA copy number inference can detect large-scale copy number alterations but has limited accuracy

The previous chapter focused on detecting CNAs using single-cell whole genome sequencing data. To maximise the utility of the G&T-seq approach and assess whether the same structural variants were detectable in the matched single-cell RNA sequencing data, DNA copy numbers were further inferred from the Smart-seq2 (and 10x scRNA-seq) expression data using inferCNV (389, 397).

InferCNV identifies copy number variations (CNVs) in tumour cells by applying a moving-average approach to scRNA-seq data (397). Specifically, inferCNV averages the relative expression of genomically adjacent genes (typically across a window of 100 genes), thereby reducing gene-specific expression variability while preserving signals indicative of chromosomal aberrations. This process is further refined by comparing the resulting profiles with a "normal" sample, which serves as a reference to normalise the CNV profiles in test samples. This is performed by subtracting the mean expression signal of the normal cells from that of the treatment cells. The resulting residual expressions highlight regions in the sample that might be over-expressed (potential amplifications) or under-expressed (potential deletions). A Hidden Markov Model (HMM) then translates these residual intensities into CNV predictions by making a probabilistic decision about the most likely CNV state for a genomic window based on the observed residual expression of the window in question and the predicted states of neighbouring windows (389). As a result, inferCNV does not provide exact copy numbers but classifies genomic windows into broader categories such as "normal", "gain", or "loss", which are then mapped to a numerical 6-state CNV model. Under this model, State 1 refers to the loss of two copies; State 2 represents the loss of one copy; State 3 is neutral; State 4 represents the addition of one copy; State 5 represents the addition of two copies; and State 6 represents the addition of more than two copies.

The heatmap presented in Figure 5.1A illustrates genome-wide CNV regions inferred from MK1- and AZD1-resistant single-cell transcriptomes as depicted under the HMM-based approach. InferCNV identified four subclonal populations primarily diverging at chromosomes 2 and 5. The most abundant clusters, 1.1.1 and 1.1.2, exhibit one-copy gains (CNV state = 4) at 2p and one-copy losses at 5q, with these structural variants sometimes co-occurring within the same subclone. Meanwhile, the minor subclusters, 1.2.1 and 1.2.2, exclusively exhibit one-copy losses at 5q (CNV state = 2), alongside one-copy gains at 12p, 15q, and 19q.

5.4. Results

In the MK1-resistant PDO, the most predominant subclones include 1.1.1.1 (comprising approximately 46% or 40 out of 87 cells), 1.1.1.2 (~30%), 1.2.2.1 (10.3%), and 1.1.2.1 (5.8%), as illustrated in Figure 5.1B. These subclones exhibit chromosomal aberrations consistent with those previously observed in single-cell genomes. These include a one-copy gain at 1p, almost exclusive to the MK1-resistant PDO (1.1.1.2), a one-copy gain at 2p (found in 1.1.1.1, 1.1.1.2, and 1.1.2.1), and a one-copy loss at 5q (observed in 1.1.1.2 and 1.2.2.1). Conversely, the AZD1-resistant PDO predominantly comprises the 1.1.2.1 subclone (75.7% or 106/140 cells) and the 1.2.1.2 subclone (10.7%), characterised by a one-copy loss at 5q (Figure 5.1B).

On the other hand, all subclones identified through copy number analysis of the 10x scRNA-seq data consistently exhibit a gain in 2p (Supplementary Figure 19A-B). Similar to the Smart-seq2 findings, the loss of the 5q region does not consistently co-occur with the 2p gain across subclones within both AKTi-resistant organoids. For instance, within the MK1-resistant PDO, the two most prevalent subclones, originating from the 1.1.1 branch, constitute 63.8% (1,727/2,708) of the total cell population. These subclones exhibit a gain at chromosome 2p yet do not show the loss at 5q. Conversely, the second most predominant subclones within this resistant organoid, belonging to the 1.1.2 branch, are characterised by gains at 1p and 2p, along with the 5q loss. This pattern aligns with the genomic profile of the second most abundant subclone identified through Smart-seq2 analysis in the MK1-resistant PDO.

In the case of the AZD1-resistant PDO, the dominant subclones were also part of the 1.1.1 branch, accounting for 60.8% (1,382 out of 2,289) of the total cell population[1] (Supplementary Figure 19B). The remaining, less abundant subclones, specifically the 1.2.1 and 1.2.2 subclones, exhibit gains at 2p and losses at 5q. Meanwhile, the least predominant subclones within the 1.1.2 cluster show gains at 1p.

Unfortunately, due to the nature of inferCNV analysis, which excludes CNVs of the reference sample from the final output, it was not possible to determine the copy number state of the Parental genome using its expression data through inferCNV. This limitation prevented a direct comparison of the CNV landscape between the Parental genome and derived cell populations in the AKTi-resistant from a scRNA-seq perspective.

[1] Please note, although the subclones across the Smart-seq2 and 10x scRNA-seq datasets share identical naming conventions, this does not imply that they correspond to the same subclonal populations.

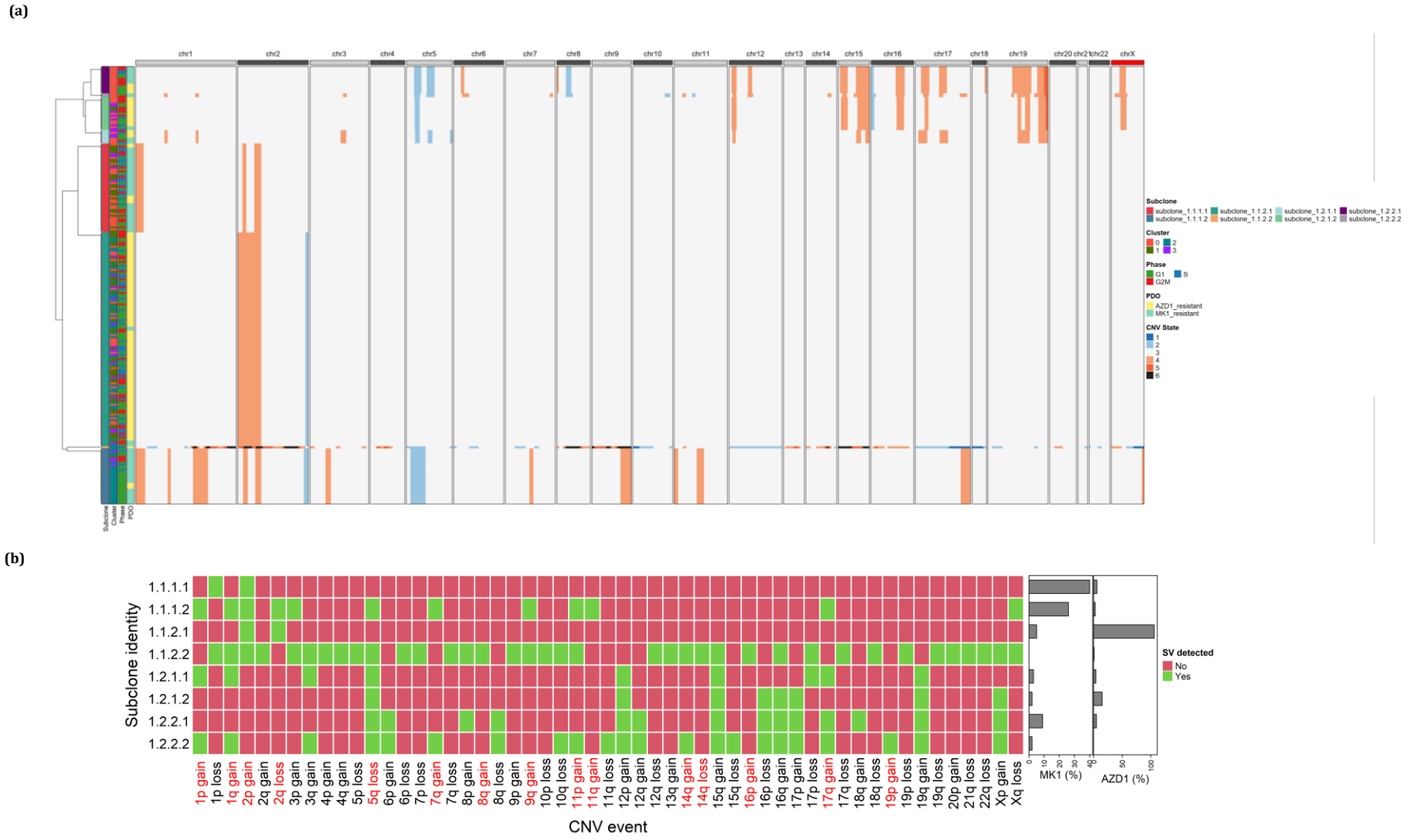


Figure 5.1. Genome-wide copy number states of AKTi-resistant cells inferred from scRNA-seq data.

5.4. Results

(a) *Heatmap displays predicted CNV regions as identified by inferCNV using a six-state Hidden Markov Model (HMM). Heatmap colours correspond to one of the six HMM states, indicating varying degrees of copy number alterations: State 1 represents the loss of two copies; State 2 indicates the loss of one copy; State 3 denotes a neutral state with no change; State 4 represents the addition of one copy; State 5 represents the addition of two copies; and State 6 denotes the addition of more than two copies. Columns correspond to genomic windows covering 100 adjacent genes, providing an averaged view of CNV states. Rows represent 227 single-cell transcriptomes derived from MK1- (n=87) and AZD1-resistant (n=140) mCRC organoids. Row annotations include information about the cells, such as organoid of origin, cell cycle phase, Seurat cluster identity, and inferCNV subclone identity.* ***(b)*** *Large-scale structural variants (SV) characterising inferCNV subclones, with adjacent bar plots denoting their frequency (%) in AKTi-resistant PDOs. CNVs highlighted in red represent CNVs observed in both Smart-seq2 and 10x scRNA-seq datasets.*

5.4. Results

The subsequent part of the inferCNV analysis aimed to evaluate the impact of CNVs identified in resistant organoids on the expression of genes within the affected chromosomal regions.

Figure 5.2 illustrates the average residual expression in AKTi-resistant cells (bottom heatmap), derived by subtracting the expression levels in the Parental reference (top heatmap), which served as a baseline for comparison. Notably, although not all RNA-seq CNV subclones exhibited a loss at 5q, genes located within chromosome 5 were generally downregulated in both resistant PDOs and across all RNA-seq subclones, compared to the Parental organoid. Similarly, AKTi-resistant organoids showed upregulation of genes located within 2p, while MK1-resistant cells (and, to a lesser extent, AZD1-resistant cells) also exhibited upregulation of genes within 1p.

Furthermore, examining the global expression of genes in all chromosomes revealed that in AKTi-resistant organoids, the overall expression of genes located on chromosome 5 was significantly lower than in the Parental control (Figure 5.3 and Supplementary Figure 21). Although the differences were less pronounced, AKTi-resistant PDOs also demonstrated increased overall expression of genes on chromosomes 1 and 2 compared to the control. Notably, the chromosomal aneuploidies hereby reported did not arise de novo in AKTi-resistant lines; instead, these variations were pre-existing, albeit at lower frequencies, in the Parental PDO. This observation was particularly noticeable in the 10x scRNA-seq data, likely due to the analysis of a larger number of cells (Supplementary Figure 20).

Several conclusions can be drawn from the CNV analysis using single-cell expression data. First, the main (sub)chromosomal copy number changes previously identified in single-cell genomes at 1p36.33-p36.13, 2p25.3-p12, and 5q11.2-q35.3 were also detected at the transcriptional level, regardless of the scRNA-seq platform employed. Although the 1p gain was present in both AKT-resistant lines, it was more frequently observed in the PDO line treated with the MK2206 inhibitor, a finding that was also observed in single-cell genomes. Furthermore, transcriptome-based CNV analysis not only established the existence of subclones carrying these chromosomal alterations in the Parental PDO before treatment, but also confirmed their expansion in AKTi-resistant organoids. Lastly, while the 2p gain and 5q loss were simultaneously observed in single-cell genomes, the 5q loss was not consistently identified in RNA-inferred CNV subclones, especially in AZD1-resistant cells. In fact, in AKTi-resistant PDOs, the most abundant subclones primarily exhibited the 2p gain without the 5q loss across both scRNA-seq platforms, even though the majority of subclones, including those not explicitly exhibiting a 5q loss, showed a decreased expression of genes within the 5q region.

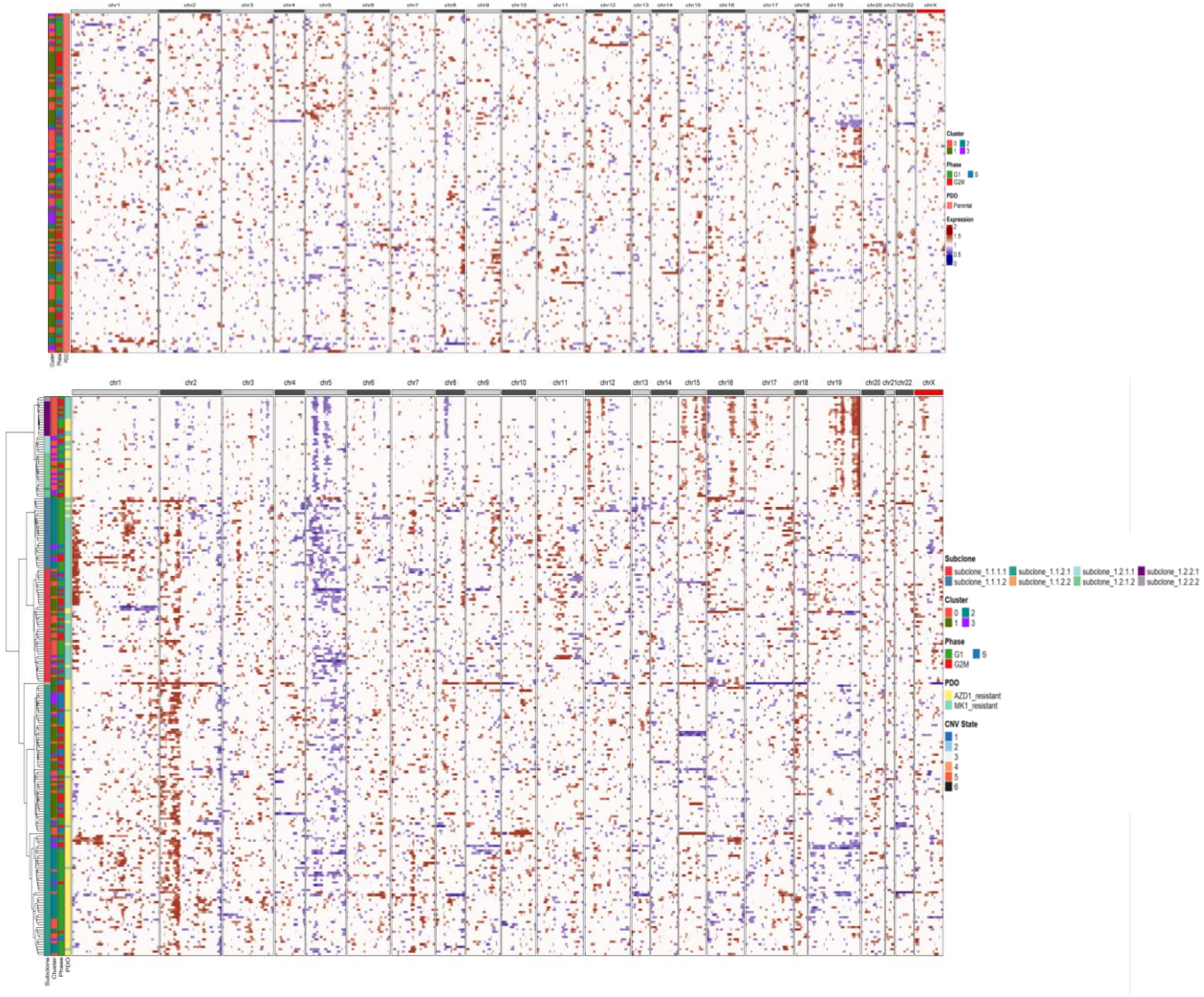


Figure 5.2.Heatmap of relative expression values of all genes across mCRC single-cell transcriptomes.

5.4. Results

In this figure, the top heatmap displays the expression values of Parental cells (n=134), which were used as a "normal" reference. Columns represent genes ordered by their absolute genomic position across chromosomes. This heatmap defines the baseline expression levels for genes in reference cells. The bottom heatmap shows the residual expression values for MK1- and AZD1-resistant cells (n=87, n=140 cells, respectively). These values were calculated by subtracting the baseline expression data of Parental cells from each AKTi-resistant sample. Colour intensities indicate chromosomal regions with significantly higher or lower expression. Specifically, red indicates regions likely containing large, amplified segments, while blue denotes regions with potential deletions. Rows (cells) are organised using hierarchical clustering based on Euclidean distance and average linkage. Row annotations include information about cells, such as organoid of origin, cell cycle phase, Seurat cluster identity, and inferCNV subclone identity.

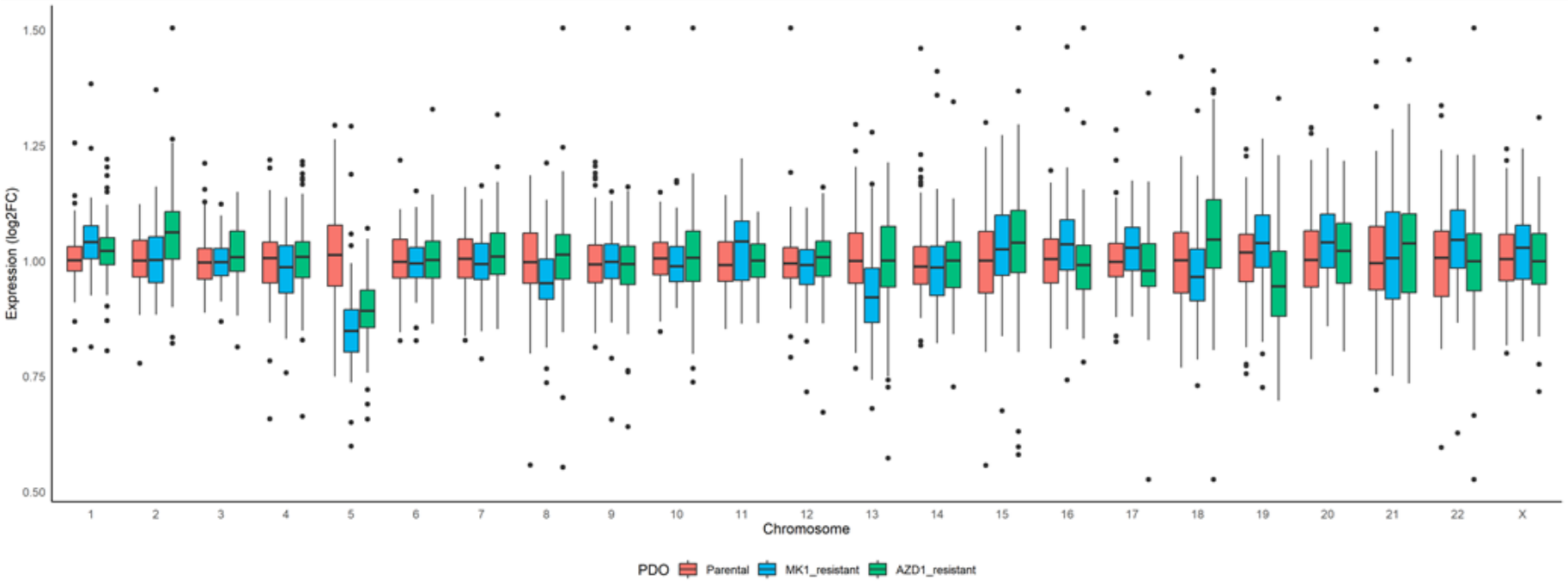


Figure 5.3. Genome-wide gene expression binned per chromosome in mCRC single-cell transcriptomes.

Box plots represent the median expression values across each chromosome for the Parental (n = 134), MK1-resistant (n=87), and AZD1-resistant (n=140) mCRC PDOs. Expression data was obtained by subtracting the average log2 fold change values in Parental cells from those in AKTi-resistant cells. The plot offers a detailed overview of how treatment-induced gene expression variations are distributed across chromosomes, shedding light on the genomic response to therapy.

5.4.2 Unveiling the intricate relationship between copy number alterations and gene expression in AKTi-resistant organoids

The single-cell genome and transcriptome analyses reviewed so far point towards a potential impact of copy number alterations (CNAs) on gene expression in mCRC PDOs. This prompted a re-evaluation of the global differential gene expression (DGE) analyses previously conducted on AKTi-resistant PDOs, aiming to uncover gene expression patterns indicative of CNA influence. This initial investigation, which examined the distribution of differentially expressed genes (DEGs) across chromosomes—specifically on chromosomes 1, 2, and 5, as they harbour the most significant CNAs in resistant organoids—revealed intriguing findings.

In the MK1-resistant PDO, chromosome 5 emerged as the chromosome with the highest concentration of downregulated genes (Figure 5.4A). In contrast, chromosome 1 exhibited the highest number of upregulated genes, with chromosome 2 also showing a modest number of upregulated genes.

On the other hand, the AZD1-resistant PDO presented a distinct expression profile, with chromosome 5 showing fewer downregulated genes than the previous PDO (Figure 5.4B). Instead, the highest number of upregulated genes came from chromosomes 1 and 2, with the latter chromosome standing out for contributing 14 upregulated genes to the pool of DEGs within this organoid.

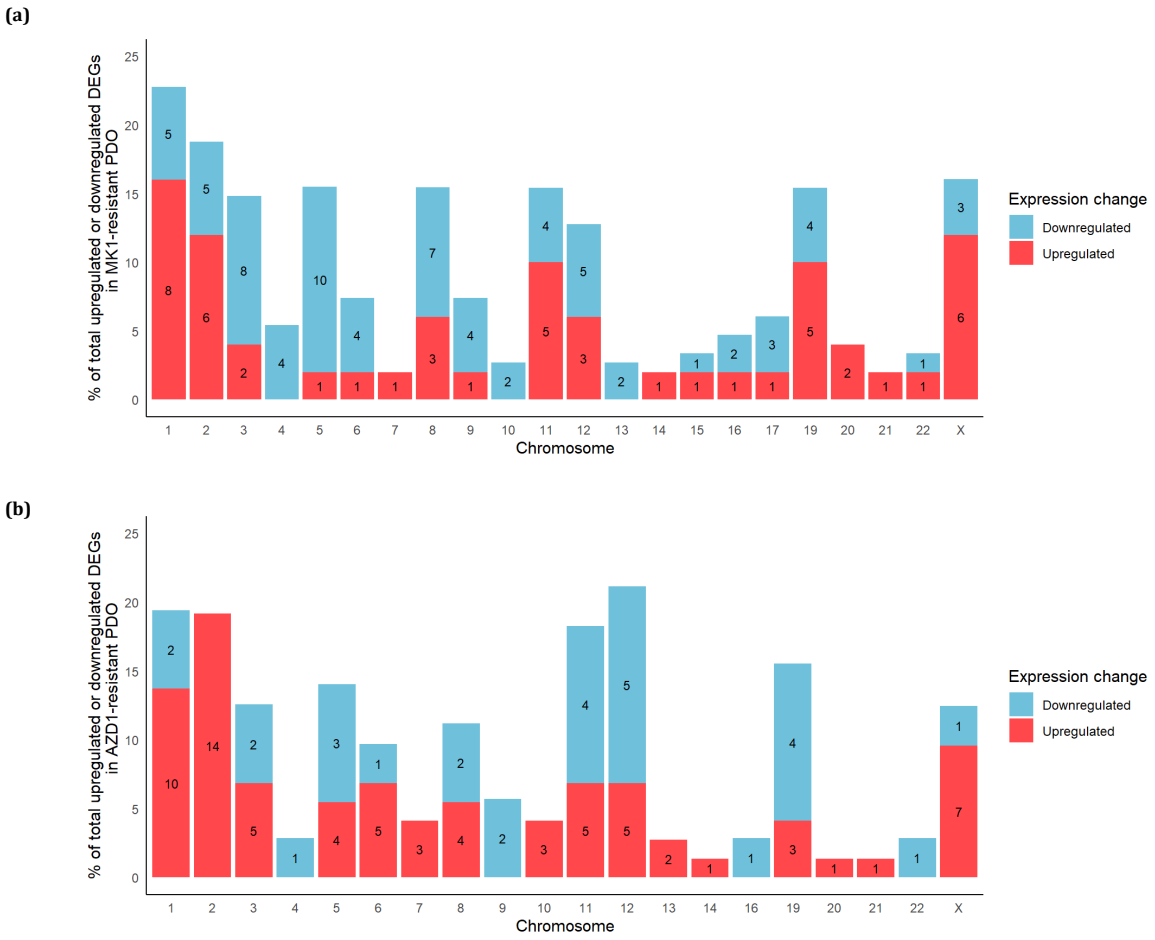


Figure 5.4. Distribution of differentially expressed genes by chromosome in AKTi-resistant PDOs.

Stacked bar plots illustrate the distribution of DEGs across chromosomes, derived from the Smart-seq2 global differential gene expression analysis detailed in Chapter 3. Plots are divided into two panels: ***(a)*** *MK1-resistant PDOs, featuring 131 DEGs, and* ***(b)*** *AZD1-resistant PDOs, with 109 DEGs. Within each bar, DEGs are categorised as either upregulated or downregulated, with the specific count of DEGs displayed inside the corresponding segment. The height of each segment represents the percentage of upregulated or downregulated DEGs on a specific chromosome, calculated relative to the total count of upregulated or downregulated DEGs identified across all chromosomes.*

5.4. Results

Given the marked downregulation of genes on chromosome 5, particularly in the MK1-resistant PDO, alongside the significant upregulation observed on chromosomes 1 and 2 in resistant organoids, the analysis proceeded to explore overlaps between the loci of differentially expressed genes and regions affected by copy number alterations. This exploration aimed to investigate the potential link between CNAs and DEGs on affected chromosomes and to understand their contribution to the observed resistance phenotype. The investigation mainly focused on genomic regions exhibiting one-copy losses, e.g., 5q11.2-q23.2 and 5q23.2-q35.3, and one-copy gains at 1p36.33-p36.13 and 2p25.3-p12. According to single-cell genome-based CNA analysis (and later supported by transcriptome-based CNV findings), these specific chromosomal alterations characterise a minor cell population in the untreated Parental organoid, which later proliferated to become the dominant subclone in the MK1- and AZD1-resistant PDOs.

From the initial 131 and 109 DEGs identified in MK1- and AZD1-resistant PDOs, respectively, 7 and 1 mitochondrial-related genes were lost during the lift-over conversion step, necessary for aligning the hg38 coordinates of DEGs to the hg19 coordinates of CNAs. Nonetheless, the entire set of DEGs for each of the resistant PDOs was found to overlap with CNA-affected regions. Figure 5.5 illustrates the broadest cytoband ranges within which CNAs identified in single-cell genomes, intersect with genes that were differentially expressed in MK1-resistant transcriptomes compared to the Parental control. Among the eight genes upregulated in chromosome 1 of this resistant PDO, three genes—*ENO1*, *PARK7*, and *PGD*—were located within the one-copy gain spanning 1p36.33-p36.13. This particular CNA was observed more frequently in the MK1-resistant line than in either the Parental or AZD1-resistant PDOs.

On the other hand, *YWHAQ*, *PDIA6*, *LRATD1*, *PREB*, *EPAS1*, and *PCYOX1* are located within the one-copy gain at 2p25.3-p12, which encompassed all the upregulated genes on this chromosome, while genes immediately downstream this CNA were all downregulated.

Furthermore, the two CNAs with one-copy losses at 5q11.2-q23.2 and 5q23.2-q35.3 predominantly contained nine of the ten downregulated genes on this chromosome, namely *TBCA*, *RPS23*, *TMED7*, *SLC12A2*, *RPS14*, *MRPL22*, *NPM1*, *HNRNPH1*, and *RACK1*. The remaining downregulated gene, *RPL37*, was situated upstream of the 5q11.2-q23.2 region. Notably, *CXCL14*, located at 5q31.1, stood out as the only gene upregulated within in this chromosome, also lying within the one-copy loss at 5q23.2-q35.3.

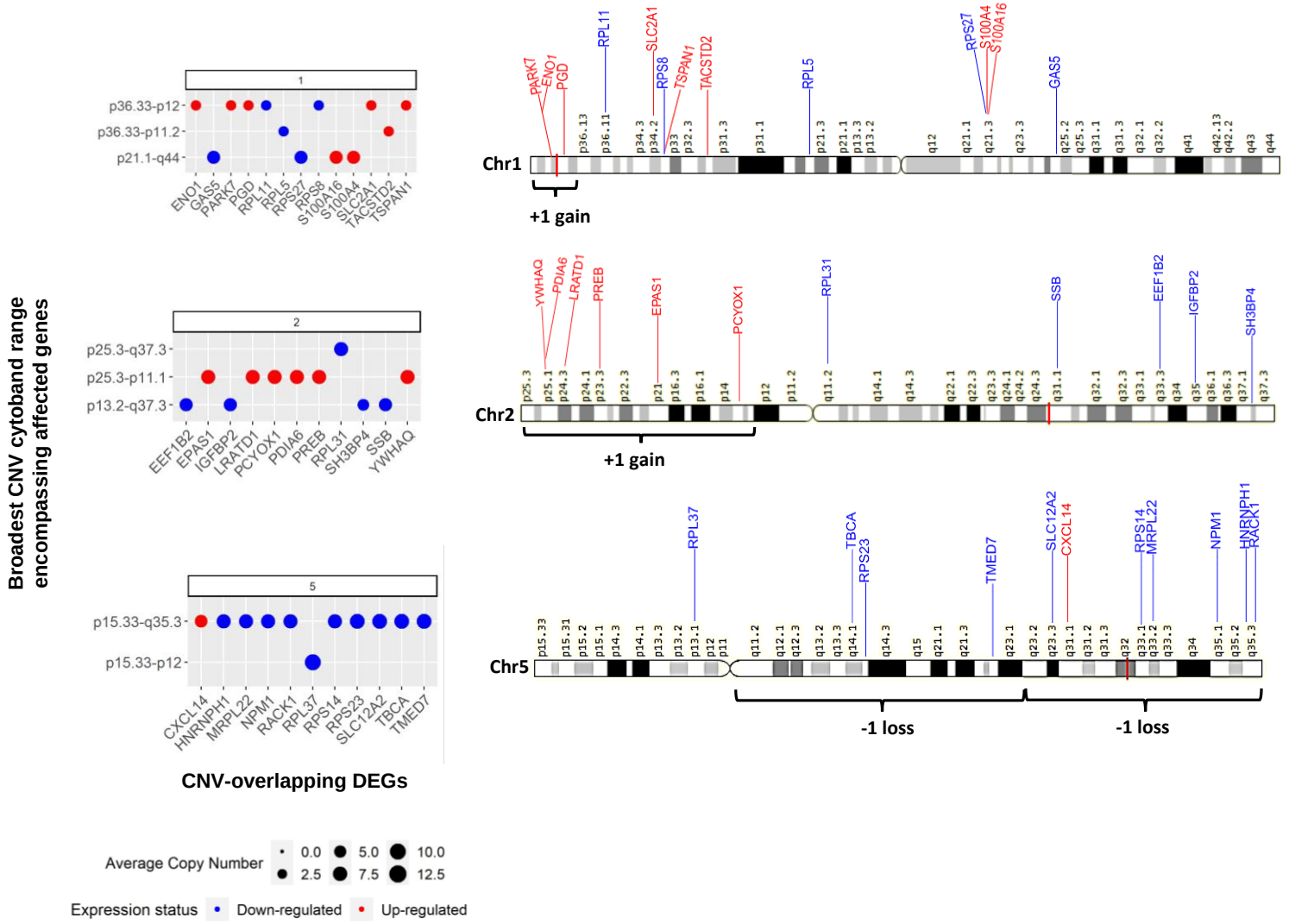


Figure 5.5. Genomic landscape of copy number alterations affecting differentially expressed genes in MK1-resistant cells.

5.4. Results

For chromosomes 1, 2 and 5, dot plots (left panel) illustrate the broadest contiguous cytoband ranges impacted by CNAs that encompass DEGs, providing a summary of key genomic regions affected by these structural variants. Dot sizes indicate the average copy number of single-cell genomes exhibiting the CNAs that make up the range, while the colours represent the direction of gene expression: red for upregulated and blue for downregulated DEGs. The right panel displays chromosome figures annotated with cytoband regions affected by one-copy losses (5q11.2-q23.2 and 5q23.2-q35.3) and gains (1p36.33-p36.13 and 2p25.3-p12) relative to the Parental control, including DEGs in these areas to illustrate the possible influence of CNAs on gene expression.

Figure 5.6 also illustrates the relationship between CNAs and DEGs in the AZD1-resistant PDO. Unlike the resistant organoid previously described, the AZD1-resistant PDO exhibited no upregulated genes within the one-copy gain spanning 1p36.33-p36.13. However, within this region, *MTCO1P12*, located at the start of the CNA at 1p36.33, was notably downregulated.

A similar pattern was observed in chromosome 5, where regions with one-copy losses did not consistently correlate with gene downregulation. Specifically, the 5q11.2-q23.2 region exclusively harboured upregulated genes, whereas 5q23.2-q35.3 featured a mix of upregulated (e.g., *CXCL14*, also upregulated in the MK1-resistant PDO) and downregulated genes, including *RPS14*, *NPM1*, and *RACK1*, all of which were also downregulated in the MK1-resistant PDO.

Lastly, nine of the fourteen genes upregulated in chromosome 2 were located within the gain at 2p25.3-p12. These included *MRPL33*, *PPP1CB*, *BIRC6*, *EPCAM*, *XPO1*, *B3GNT2*, *PCYOX1* (also upregulated in the MK1-resistant PDO), *SNRPG*, and *TGFA*.

The overlap analysis of CNAs and DEGs in MK2206- and AZD5363-resistant mCRC PDOs revealed that similar genetic alterations did not uniformly result in the anticipated patterns of gene expression changes. Specifically, gains in chromosomes 1 and 2 did not invariably lead to upregulation, nor did losses within chromosome 5 always result in downregulation, though in some instances, these expected relationships were observed. This variability points to a multifaceted basis of resistance to AKT inhibition in these two organoids.

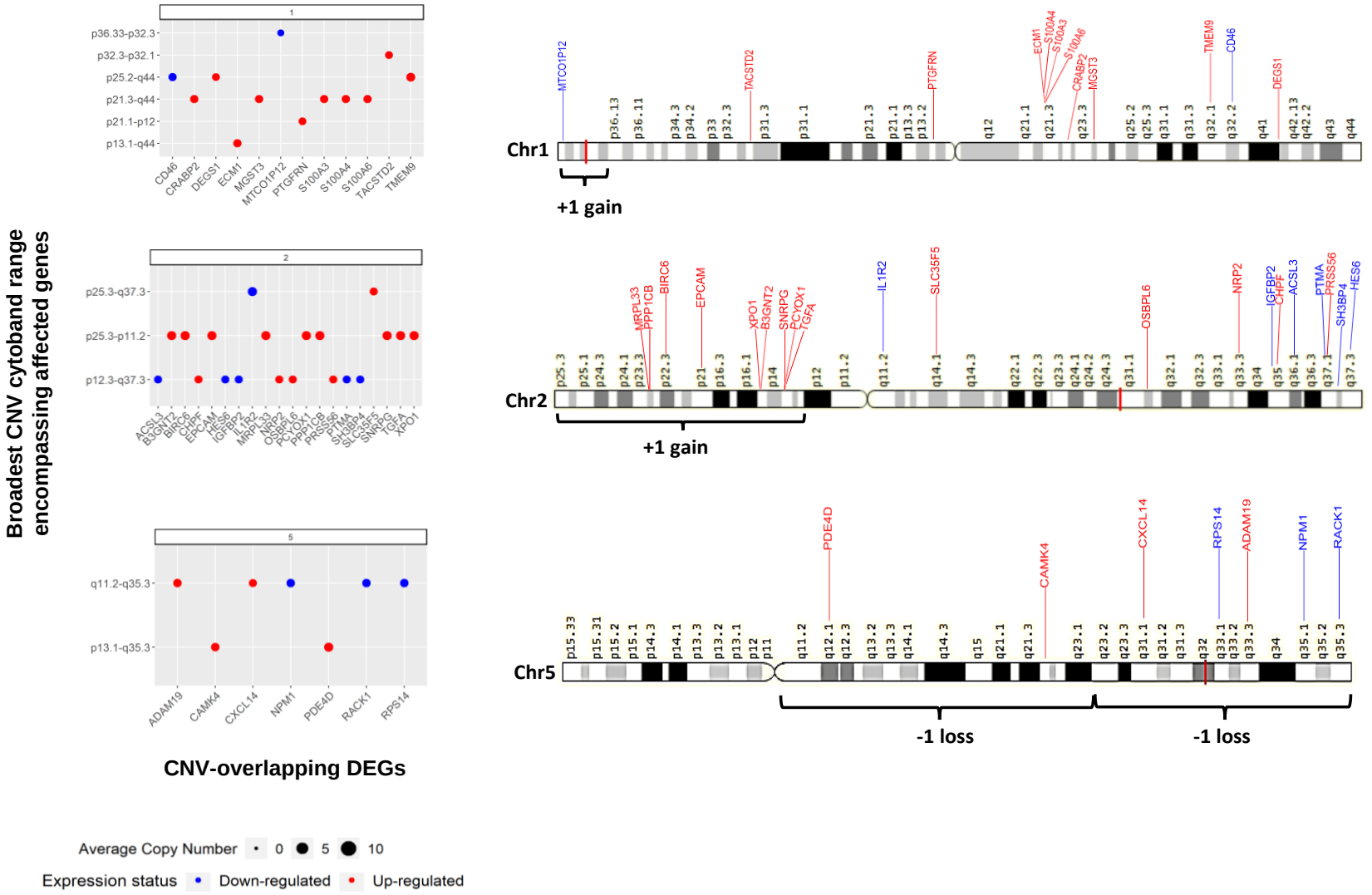


Figure 5.6. Genomic landscape of copy number alterations affecting differentially expressed genes in AZD1-resistant cells.

5.4. Results

For chromosomes 1, 2 and 5, dot plots (left panel) illustrate the broadest contiguous cytoband ranges impacted by CNAs that encompass DEGs, providing a summary of key genomic regions affected by these structural variants. Dot sizes indicate the average copy number of single-cell genomes exhibiting the CNAs that make up the range, while the colours represent the direction of gene expression: red for upregulated and blue for downregulated DEGs. The right panel displays chromosome figures annotated with cytoband regions affected by one-copy losses (5q11.2-q23.2 and 5q23.2-q35.3) and gains (1p36.33-p36.13 and 2p25.3-p12) relative to the Parental control, including DEGs in these areas to illustrate the possible influence of CNAs on gene expression.

5.5 Discussion

In this chapter, the single-cell genome and transcriptome datasets independently analysed in previous chapters were integrated to identify the impact of chromosomal instability on gene expression and consequently, on the development of resistance to AKT inhibition in mCRC PDOs.

The analysis began by employing inferCNV to detect DNA copy number variations (CNVs) from gene expression data, aiming to verify if the copy number alterations (CNAs) previously identified in single-cell genomes were also detectable at the transcriptional level. This investigation focused primarily on CNAs that characterised a minor subpopulation of cells within the Parental PDO that later expanded to dominate AKTi-resistant organoids. These CNAs included a one-copy gain spanning 2p25.3-p12 in single-cell genomes, resulting in a copy number of 5 relative to the Parental PDO, which showed 4 copies at this locus. Another key structural variant observed in resistant PDOs was a one-copy deletion at 5q11.2-q23.2 and 5q23.2-q35.3, with copy numbers of 4 and 3, respectively. Additionally, the one-copy gain at 1p36.33-p36.13 (copy number = 3) was also of interest as it was scarcely observed in Parental genomes but became more frequent in MK1-resistant than in the AZD1-resistant genomes, albeit still at low frequencies, suggesting its role as a passenger CNA (398).

Indeed, inferCNV identified the main CNAs located at 1p, 2p, and 5q across Smart-seq2 and 10x scRNA-seq datasets. Moreover, inferCNV detected the 2p and 5q alterations within the Parental organoid, thereby confirming that these structural variants did not arise de novo, as they were present in a subset of cells within the control PDO prior to treatment with AKT inhibitors. Nevertheless, some ambiguity remained with respect to the CNAs detected. Although the 2p25.3-p12 and 5q11.2-q35.3 CNAs were concurrently identified in the genomic analyses of resistant PDOs, analysis based on RNA expression data for CNA detection revealed a different pattern. The majority of RNA-inferred CNA subclones exhibited the 2p gain without the accompanying 5q loss across both scRNA-seq platforms, with the co-occurrence being more frequently detected in MK1- than in AZD1-resistant transcriptomes. However, when examining genome-wide expression levels in resistant cells, an overall downregulation of genes within chromosome 5 was observed. This suggests that, although the 5q loss may not have been directly identified as a CNA through RNA-inferred methods, the expression data subtly indicated its functional impact.

The underlying reason for these observations was further explored in the second part of the analysis, which assessed the distribution of previously identified differentially expressed genes (DEGs) across all chromosomes in resistant organoids. The genomic coordinates of DEGs were then compared with CNA-affected regions, with results showing that all DEGs—124 for the MK1-resistant and 108 for the AZD1-resistant PDO—fell within areas of chromosomal instability. This observation aligns with expectations considering that all PDOs, including the Parental control, were triploid and exhibited genome-wide aneuploidies.

While some concordance was observed between CNAs and gene dosage changes, with "gains" often associated with upregulation and "losses" with downregulation, this observation did not consistently apply across the main CNAs explored. For instance, a special situation was observed within the CNA spanning 1p36.33-p36.13. On one side, the MK1-resistant PDO exhibited upregulation of *ENO1*, *PARK7*, and *PGD* within the region. In contrast, the AZD1-resistant PDO did not show any upregulated genes. In fact, only *MTCO1P12* was found downregulated within this CNA.

The observation that the metabolic enzymes *ENO1*, *PARK7*, and *PGD*, located within the 1p CNA region were upregulated in the MK1-resistant PDO—despite more cells lacking this specific CNA than having it—raises questions about how genes are found to be differentially expressed between conditions at the single-cell level. This situation prompts an investigation into whether such upregulation is attributable to the relative frequency of CNAs across the conditions being compared, the cumulative effect of an increased gene product resulting from the CNA, or a combination of both factors.

The MK1-resistant PDO also exhibited numerous downregulated genes within the 5q11.2-q35.3 CNA, where 9 of the 10 DEGs on the chromosome were located. These included *TBCA*, *RPS23*, *TMED7*, *SLC12A2*, *RPS14*, *MRPL22*, *NPM1*, *HNRNPH1*, and *RACK1*. In contrast, the AZD1-resistant organoid did not show significant downregulation of genes in the same region, with the exception of *RPS14*, *NPM1*, and *RACK1*. These genes, located within 5q23.2-q35.3, were the only ones downregulated in chromosome 5. Interestingly, *CXCL14*, also located within 5q23.2-q35.3, was upregulated in both organoids. This pattern suggests that inferCNV, which detects CNAs based on expression data, may not have identified the 5q regions as potential CNAs in the AZD1-resistant PDO due to the less pronounced downregulation of genes in this region, especially when compared to the MK1-resistant PDO. The difficulty in detecting the 5q loss among RNA-inferred CNA subclones underscores that certain CNA events, particularly losses, pose challenges for identification through RNA-based methods.

5.5. Discussion

The differential gene expression patterns observed in chromosome 5, with regions of CNAs leading to the upregulation of *CXCL14*, alongside the downregulation of *RPS14*, *NPM1*, and *RACK1* in both resistant organoids present an intriguing situation. In the case of *CXCL14*, the upregulation of this gene could be regulated by factors independent of copy number such as epigenetic modifications or cis/trans-acting elements located beyond the sites of CNA (399, 400). Alternatively, *CXCL14* might be involved in compensatory pathways aimed at counteracting the effects of AKT inhibition, an adaptive strategy often associated with the development of resistance to certain treatments (401). This mechanism could account for the consistent upregulation of *CXCL14* across organoids.

The role of *CXCL14* as a chemokine seems to be context-dependent, with studies suggesting both tumour-suppressive and tumour-promoting activities. When secreted by cancer-associated fibroblasts (CAFs), *CXCL14* promotes tumour growth by inhibiting apoptosis and encouraging glycolysis and metastasis (402). In contrast, its secretion by epithelial cells establishes immune surveillance (253), and inhibits tumour invasion by recruiting B-cells, monocytes, and dendritic cells into tumour tissues (403). The contribution of *CXCL14* to cancer cell metabolic reprogramming through the Warburg effect (312) is particularly interesting as several other glycolytic enzymes like *ENO1* and *PGD* were previously found to be upregulated in the MK1-resistant PDO.

Regarding *RPS14* and *NPM1*, they are interconnected through their roles in protein synthesis and ribosome biogenesis (255). While *RACK1* (receptor for activated C kinase 1) is typically identified as a scaffolding protein involved in various cellular processes, ribosomal *RACK1*, situated on the 40S subunit near the mRNA exit channel, contributes to translation control by recruiting eukaryotic initiation factor 4E (404). The overall downregulation of genes associated with protein synthesis in both AKTi-resistant PDOs was noted in previous chapters (Figure 3.18A And Figure 3.20A). The dysregulation of the translational machinery in CRC often stems from alterations in key signalling pathways, including WNT, RAS/MAPK, and PI3K/AKT/mTOR. These alterations not only deregulate proteins essential for protein synthesis but also impact ribosome biogenesis (327). The consistent expression of these four genes in both samples, despite the CNA differences relative to the Parental control, suggests they may provide a survival advantage under different forms of AKT inhibition.

Regarding chromosome 2, the two AKTi-resistant PDOs demonstrated a considerable number of genes upregulated within the 2p25.3-p12 CNA. In the MK1-resistant PDO, all upregulated genes on this chromosome (i.e., 6 genes) fell within this region, while the AZD1-resistant PDO exhibited 9 out of 14 upregulated genes in the same region, including *PCYOX1*, which was also

upregulated in the MK1-resistant PDO. The significant upregulation of genes in this region supports the hypothesis that only elevated gene expression levels facilitate the detection of large-scale CNAs using transcriptome-based CNA detection methods like inferCNV. Furthermore, the notable upregulation within the 2p region, coupled with the observation that the majority of RNA-inferred CNA subclones in AKTi-resistant PDOs predominantly exhibited the 2p gain without the 5q loss, suggests that this specific genomic alteration may play a pivotal role in providing a survival advantage, and could have promoted the proliferation of cells harbouring this feature under the selective pressure imposed by the AKT inhibitors.

The observation that certain CNAs lead to an increase in the total amount of gene product, whereas others lead to a decrease, supports the fundamental concept that changes in gene dosage resulting from CNAs, can directly influence gene expression levels (405). However, as previously exemplified, the relationship between CNAs and DEGs extends beyond simple gene dosage effects. Multiple levels of gene dosage compensation exist to counteract the imbalances in gene expression resulting from aneuploidy to maintain tumour fitness despite the presence of genomic abnormalities (387). For instance, copy number gains in 6p are commonly observed in melanomas and breast carcinomas (BRCA) (406). Given that changes in copy number frequently affect large chromosomal segments, this often leads to collateral or "passenger" copy number alterations in nearby genes that do not necessarily provide a growth advantage (407). In melanoma and BRCA, the gain in 6p includes passenger genes involved in MHC class I-mediated antigen presentation including *TAP1*, *TAP2*, *TAPBP* and *XPO5* (407, 408). Notably, overexpression of *TAP1* is associated with immune infiltration. The expression of these genes is often disconnected from the 6p gain, partially due to promoter hypermethylation, which results in their silencing and contributes to immune evasion (408).

Mohanty and collaborators investigated the complex regulatory mechanisms that contribute to the uncoupling of gene expression from copy number alterations (UECN) across six different cancers, aiming to identify therapeutic targets within these regulatory factors (407). Their analysis revealed that genes with uncoupled expression from CNAs (nCpD) were less sensitive to changes in copy number compared to genes with coupled expression (CpD). Specifically, they found that amplified nCpD genes were more strongly suppressed by promoter methylation, indicating a complex interplay between genetic alterations and epigenetic regulation of gene expression in cancer. This study highlighted the pivotal role of transcription factors (TFs) in mediating this uncoupling, demonstrating that TFs associated with nCpD genes significantly influence their expression irrespective of copy number changes. This suggests that the

regulatory effect of TFs can override the expected gene dosage effects that would normally result from CNAs.

Building on their findings, Mohanty *et al*., proposed a novel strategy for cancer treatment focused on targeting specific TFs to reverse UECN (407). By identifying TFs that could reestablish expression coupling of "tumour-toxic" gene copy number changes, they suggested that manipulating these TFs could compromise tumour fitness, exploiting gene dosage sensitivity to combat cancer more effectively. The application of this strategy across six cancer types in The Cancer Genome Atlas (TCGA) led to the identification of 21 TFs as potential targets. For example, in lung adenocarcinoma (LUAD), they discovered clusters of amplified nCpD genes associated with anti-tumour phenotypes and identified TFs capable of regulating these clusters in a way that could potentially enhance anti-cancer responses.

The study above highlights the complex interplay between genetic and epigenetic factors in cancer, serving as a prime example of how tumours can fine-tune the effects of aneuploidy to their advantage. Additionally, the relocation of dosage-sensitive genes to different parts of the genome can help maintain the right balance of gene expression (387). These insights can be extrapolated to the current research thesis, offering a plausible explanation for the observed discrepancies between gene copy number and gene expression in chromosomes 1 and 5 within the AZD1-resistant PDO. It suggests that these discrepancies may represent adaptive mechanisms through which cancer cells leverage genetic imbalances, not just for survival, but to actively resist targeted therapies. This adaptive capacity of cancer cells presents both a challenge and an opportunity for therapeutic intervention, where disrupting these compensation mechanisms could offer a novel strategy for cancer treatment (387).

Based on the findings from the integration analysis, several genes emerge as promising targets for follow-up studies to further elucidate their roles in driving resistance to AKT inhibition in mCRC PDOs. The genes of interest include *PARK7*, *PGD*, and *ENO1*, particularly in the MK1-resistant PDO, as they are likely involved in metabolic processes. Additionally, *PPP1CB* and *TGFA* (both located within the 2p25.3-p12 CNA) seem to be interesting genes to explore in the AZD1-resistant PDO, given their roles in inactivating AXIN, a component of the destruction complex in the Wnt-signalling pathway (321), and in driving resistance to cetuximab (322), respectively. Finally, *PCYOX1*, *CXCL14*, *RPS14*, *NPM1*, and *RACK1* should also be studied due to their expression in both resistant organoids.

In the first instance, validation of target genes should start by performing quantitative PCR (qPCR) assays to ascertain their expression levels in resistant versus control organoids before

any functional experiments are conducted. This would confirm whether the observed differences in expression are statistically significant and biologically relevant.

To assess whether resistance to AKT inhibition can be reversed, a combination of genome editing experiments and functional assays could be performed. These experiments would involve overexpressing target genes (*RPS14*, *NPM1*, *RACK1*) using CRISPR-mediated transcriptional activation (CRISPRa) (409). Conversely, to determine the functional role of upregulated genes (*PARK7*, *PGD*, *ENO1*, *PCYOX1*, *CXCL14*), gene knockout experiments using CRISPR-Cas9 or transcription activator-like effector nucleases (TALENs) (410), or gene knockdown experiments using small interfering RNA (siRNA) or short hairpin RNA (shRNA) should be performed (411). These methods target mRNA transcripts for degradation or inhibit their translation. Combined gene manipulation experiments could also be performed to investigate the synergistic effects of these genes on resistance by, for example, reactivating the expression of genes in chromosome 5 (e.g., *RPS14*, *NPM1*, *RACK1*) and silencing the expression of genes in chromosome 2 (e.g., *PCYOX1*). Following these genetic modifications, the modified MK1- and AZD1-resistant PDOs would be treated with the original MK-2206 and AZD5363 inhibitors, and a cell viability assay would assess changes in drug tolerance in these mCRC PDOs. These approaches would provide insights into the underlying mechanisms of resistance and help determine whether the manipulation of these genes could restore sensitivity to AKT inhibition in advanced CRC cases.

In addition, for *CXCL14*, since it was the only gene upregulated in an area of copy number loss, epigenetic studies could be used to investigate the modifications regulating this gene. Techniques such as bisulfite sequencing (412, 413), chromatin immunoprecipitation followed by sequencing (ChIP-seq) (414), and assay for transposase-accessible chromatin using sequencing (ATAC-seq) (415, 416) could be employed. Bisulfite sequencing can analyse DNA methylation patterns, ChIP-seq can identify histone modifications and protein-DNA interactions, and ATAC-seq can assess chromatin accessibility. Alternatively, methods such as single-cell genome and epigenome by transposases sequencing (scGET-seq) can probe both open and closed chromatin (416). These studies at both bulk and single-cell levels could provide comprehensive insights into the epigenetic mechanisms regulating *CXCL14* and contribute to understanding its role in resistance to AKT inhibition in mCRC PDOs.

While single-cell genome and transcriptome data were integrated in this chapter, it is important to establish the strengths and limitations of the chosen approach. Firstly, this integration does not strictly qualify as a pairwise analysis because the transcriptomic data from each cell was not directly correlated with its corresponding genomic data. Instead, the analysis compared

overall patterns of CNAs identified in a subset of 80 genomes with differential gene expression patterns observed in a larger set of 361 transcriptomes. This approach provided broad population-level insights into genetic and transcriptional changes across the entire cell population in MK1- and AZD1-resistant organoids, allowing the identification of common patterns and trends associated with drug resistance that may have been missed in a strictly pairwise analysis. Additionally, this approach reduced costs by focusing on a subset of genomes, which still allowed for the identification of key chromosomal regions where CNAs correlated with differential gene expression, rather than sequencing all 361 genomes. This approach helped identify key chromosomal regions where CNAs correlated with differential gene expression, highlighting important areas for further study. Moreover, by utilising the 361 single-cell transcriptomes for gene expression analysis, this approach increased the statistical power of differentially expressed genes, leading to more robust results.

However, there are notable disadvantages to this approach as well. The lack of single-cell pairwise correlations may limit the full potential of G&T-seq by restricting the ability to understand the direct relationship between genomic changes and transcriptomic responses at the individual cell level. Additionally, population-level analysis might overgeneralise findings, potentially missing unique or rare cell-specific interactions that could be critical for understanding specific resistance mechanisms. Lastly, any observed correlations between CNAs and gene expression changes are indirect and require further validation to confirm causality.

Undoubtedly, establishing the relationship between CNAs and gene expression can be complex, as CNAs can influence gene expression through several mechanisms beyond dosage compensation. A strategy that could be applied to determine if changes in gene expression correlate with their copy number would involve comparing copy number amplification and expression level upregulation ratio against copy number deletion and expression level downregulation, as well as copy number deletion and expression level downregulation ratio against copy number amplification and expression level upregulation (417). Additionally, expression quantitative trait loci (eQTL) analysis can be a valuable method for evaluating the impact of CNAs on mRNA expression, identifying associations with changes in nearby (cis-eQTLs) or distant (trans-eQTLs) genes, indicating potential regulatory interactions (418).

The first part of this chapter also aimed to maximise the potential of the G&T-seq technique, by assessing the suitability of transcriptome-based DNA copy number detection to validate chromosomal changes previously observed in single-cell genomes. Unlike WGS, RNA-seq is primarily designed to measure gene expression levels and therefore faces challenges in

distinguishing whether changes in gene expression are due to differential expression or due to actual changes in the number of DNA copies (419). This difficulty arises because RNA-seq data is inherently biased towards exonic regions—where genes are expressed—and does not uniformly cover the entire genome. This bias means that while RNA-seq can indicate the copy number states of genes, it may not accurately represent the copy number of intergenic regions, which are not captured.

Another issue specific to inferCNV is its inability to detect small structural changes. This limitation arises because CNA calls by inferCNV are based on genomic windows encompassing at least 100 genes. Given that the median length of a human protein-coding gene is approximately 26 kb (420), and recognising that genomic regions vary widely in gene density, a 500 kb stretch—which was the chosen window size for detecting CNAs in single-cell genomes—would typically accommodate far fewer genes than the 100 genes inferCNV uses to predict copy number changes. This makes inferCNV unsuitable for detecting focal or localised genomic changes. Despite these challenges leveraging the transcriptome for DNA copy number analysis remains a cost-effective alternative to genome-based approaches.

While not investigated in this thesis, a foreseeable limitation of using plate-based G&T-seq with PicoPlex Gold WGA to identify single-nucleotide variants (SNVs) would be the potential inaccuracy in detecting the frequency of alternative alleles. This arises because some mutations might be underrepresented or entirely missed depending on the variant allele frequency and the sequencing depth. Indeed, allelic dropout (ADO) events and allelic imbalances, which are very commonly observed in single-cell genome sequencing, stem from the challenges in capturing and amplifying DNA from single cells (421). These factors pose a significant challenge for accurate genotyping.

Nevertheless, single-cell sequencing still offers unparalleled resolution for studying tumour heterogeneity. A significant advantage of the G&T-seq technique is the flexibility it offers after the separation of mRNA and DNA, allowing for alternative amplification methods. Although multiple displacement amplification (MDA) (422, 423) has shown limitations for copy number analysis in the past due to over-amplification of certain regions (333), it has demonstrated greater success in calling SNVs (424). Furthermore, MDA-amplified fragments are compatible with long-read technologies such as PacBio high-fidelity (HiFi) sequencing (425), potentially overcoming some of the limitations associated with low sequencing depth.

In conclusion, the interplay between the genome and the transcriptome is inherently complex. G&T-seq facilitated the identification of differentially expressed genes on chromosome 2 (and

chromosome 1 for the MK1-resistant PDO) affected by CNAs across AKTi-resistant PDOs. However, the absence of direct CNA effects on chromosome 5 in the AZD1-resistant PDO suggests a multifaceted impact of CNAs on gene expression, implicating mechanisms beyond gene dosage changes. Despite numerous challenges, the integrated analysis of G&T-seq datasets revealed multiple similarities and differences that would have remained obscured by examining only one of these "omics" layers separately. Furthermore, the differential representation of CNAs and their effects on resistant organoids underscores the complexity of genomic responses to targeted therapies and highlights the importance of dissecting these patterns in detail to understand the molecular basis of resistance.

Chapter 6

General discussion & future directions

6.1 General discussion

Initial analyses of single-cell transcriptomes from MK1- and AZD1-resistant mCRC organoids focused on identifying clusters exhibiting drug-resistant phenotypes and characterising colonic cell types. Subsequent analyses evaluated gene expression differences between AKTi-resistant PDOs and the untreated Parental to determine the impact of AKT inhibition. Furthermore, single-cell genomic and transcriptomic analyses of Parental and AKTi-resistant cells identified DNA copy number alterations (CNAs). These analyses facilitated the exploration of clonal dynamics from untreated to AKTi-resistant organoids.

To further elucidate the impact of CNAs on gene expression, the genomic locations of differentially expressed genes between untreated and AKTi-resistant PDOs were compared to the coordinates of CNAs. This approach strengthened the findings by directly linking genomic alterations to changes in gene expression, underscoring the biological significance of CNAs in driving the drug-resistance phenotype. These insights contributed to a deeper understanding of the mechanisms by which mCRC PDOs developed resistance to AKT inhibition.

Four transcriptionally distinct cell clusters were observed in all mCRC PDOs, though one of the clusters (Cluster 3) likely emerged due to technical factors. Cluster 0 consisted of cells with low proliferative activity and an elevated gene expression in protein synthesis and ribosomal biogenesis. This cluster exhibited a slight decrease in the MK1- and AZD1-resistant PDOs compared to the untreated control, indicating a transcriptionally stable population before and after treatment. Contrarily, Cluster 1 was characterised by highly proliferative cells with transcriptional profiles indicative of stem-like or undifferentiated cells. This observation, coupled with the upregulation of genes involved in various oncogenic processes such as anchorage-independent growth and cancer stem cell maintenance, suggests that this cluster represented an aggressive phenotype. Notably, the cell abundance of this cluster significantly increased in the MK1-resistant PDO compared to the Parental PDO.

Regarding the MK1-resistant PDO, differential gene expression across conditions revealed the upregulation of genes involved in several processes, including extracellular matrix remodelling and cell motility (*ACTB*, *CFL1*, *KRT18*, *S100A4*) and cell-cell adhesion (*MUC21*) in this organoid. However, the most striking finding was the upregulation of genes encoding metabolic enzymes (e.g. *ENO1*, *PGD*, *PDK3*) and genes encoding plasma membrane transporters (e.g., *GLUT1*, *SLC6A14*, *CD36*). These genes play roles in pathways of energy metabolism, such as glycolysis or the pentose phosphate pathway. In particular, the upregulation of glycolytic genes aligns with the Warburg effect, suggesting a shift towards aerobic glycolysis to meet the increased metabolic demands associated with cancer progression and resistance.

6.1. General discussion

Cluster 2 comprised non-proliferative cells expressing genes that regulate cell-cell communication (*GJA1*) and genes associated with epithelial cell migration and tissue remodelling (*CEMIP*, *ECM1*, *LGALS1*). This cluster increased in the AZD1-resistant organoid, which also showed upregulation of genes involved in fatty acid (*CD36*, *FABP5*) and carbohydrate (*PDP1*) metabolism, extracellular matrix-remodelling genes (*S100A4*, *FNDC3A*, *ECM1* and *TACSTD2*), and genes regulating cellular stability and survival under cellular stress (*S100A6*, *MGST3*). Additionally, there was a notable increase in the expression of genes that regulate immune responses (*ENTPD1/CD39, CXCL14*). These adaptations could have enabled the AZD1-resistant PDO to maintain cellular integrity under AKT inhibition.

The observed differences in cluster abundance and gene expression profiles in the MK1- and AZD1-resistant organoids highlight the distinct transcriptional consequences of using different AKT inhibitors. The different responses to MK-2206 and AZD5363 might be attributed to their distinct mechanisms of AKT inhibition: MK-2206 is an allosteric inhibitor of AKT, while AZD5363 is an ATP-competitive inhibitor (146). Furthermore, while MK-2206 is a highly selective inhibitor of AKT, it has less affinity for the AKT1/2/3 isoforms than AZD5363 (426), which has higher affinity but may also interact with off-target proteins like P70S6K, PKA and ROCK1/2 (427). Consequently, these distinctions suggest that the choice of inhibitor can significantly influence the transcriptional and phenotypic landscape of cancer cells, affecting not only their response to treatment but also the characteristics of cells after acquiring drug resistance. This variability in drug response could also be attributed to the inherent cellular heterogeneity observed in solid tumours like CRC.

The matched genomic data indicated that the development of drug resistance in both AKTi-resistant organoids involved the expansion of one particular subclone (C2) with specific CNAs at chromosomes 2 and 5. This subclone was already present in the Parental control, albeit at a much lower frequency. The combination of CNAs at chromosomes 5 and 2 in AKTi-resistant PDOs might have worked synergistically to provide a more robust resistance mechanism. This synergy could have given these cells a competitive edge over other cells, enabling them to dominate under both treatment conditions. However, stochastic events could have also played a role.

The observation that the same subclone became dominant in both MK2206- and AZD5363-resistant PDOs despite their different mechanisms of action and cross-reactivities, suggests that the genetic alterations at chromosomes 2 and 5 in this subclone (and epigenetic modifications) equipped it with versatile resistance mechanisms effective against different

AKT inhibitors. In other words, although the inhibitors have distinct mechanisms of action, both ultimately prevent AKT activation. Therefore, if the resistance mechanisms either target downstream components within the PI3K/AKT pathway or activate alternative signalling pathways—e.g. GSK-3β is downstream AKT, and can also be inactivated by Wnt signalling (26)—leading to the transcription of genes typically regulated by AKT, it would confer resistance to both types of inhibitors. The activation of alternative oncogenic pathways that effectively bypass therapeutic blockade is a well-documented mechanism of resistance in cancer cells (66). This scenario suggests a combination of convergent evolution and phenotypic plasticity. While the same clonal population exhibited a resistant phenotype under different AKT inhibitors (convergent evolution) (428), the lack of genetic divergence points toward phenotypic plasticity, where the same genetic makeup produces different phenotypes in response to varying environmental pressures (429). This phenotypic convergence due to plasticity implies that this clonal population possessed an inherent adaptability to overcome diverse AKT inhibition strategies. However, further studies are needed to definitively confirm whether this phenotypic convergence is solely due to plasticity or if subtle (epi)genetic changes undetected by current methods might also have contributed.

Another reason for this phenomenon could be that using AKT inhibitors imposed selective pressure on the tumour cell population. This pressure could have favoured the survival and expansion of cells with the CNAs in a process akin to natural selection, where environmental pressures lead to the survival of the fittest (88). In addition to the CNAs at chromosomes 2 and 5, the MK1- and AZD1-resistant cells displayed PDO-specific CNAs, with a distinct CNA on chromosome 1 in the MK1-resistant organoid and a CNA on chromosome 18 in the AZD1-resistant organoid. These and other passenger CNAs likely provided further adaptive advantages under the different AKT inhibitors, and may be responsible for the distinct transcriptional responses observed to the two inhibitors.

In the MK1-resistant organoid, a direct correlation was observed between CNAs and the upregulation or downregulation of differentially expressed genes located within CNA regions across chromosomes 1, 2, and 5. Specifically, “gains” in CNAs were associated with gene upregulation, while “losses” were linked to downregulation, except for *CXCL14* on chromosome 5, which was upregulated despite being in a region with fewer copies compared to the Parental PDO. Conversely, in the AZD1-resistant organoid, the CNA-gene expression relationship was consistent only for chromosome 2, which indicates that CNAs might not directly influence gene expression across all scenarios or that other compensatory mechanisms could modulate gene dosage effects. Furthermore, this highlights the significance of the CNA on chromosome 2 in

conferring resistance to AKT inhibition. Notably, *PCYOX1* was the only gene upregulated in both organoids within this region, pointing to its possible role in resistance. However, it is also plausible that CNAs in these areas could influence the expression of genes beyond those directly encompassed by the CNA, though such effects were beyond the scope of this study.

This study has yielded insightful findings, yet it is crucial to acknowledge that the experimental design of this study did not represent the clinical scenario of drug resistance development in patients. In this experiment, PDOs were unintermittedly exposed to the inhibitors in the media, leading to the fast development of resistance. In reality, patients often undergo treatment cycles with breaks, a strategy intended to prevent or delay the emergence of drug resistance. This difference in treatment approach could influence the development and characteristics of drug resistance observed in clinical settings versus this experimental model. Moreover, chemotherapy in clinical practice is commonly administered with other therapeutic agents rather than as a monotherapy, as in this study. This combination therapy is typically designed to target multiple pathways within cancer cells, such as DNA replication and other cellular processes. By attacking cancer cells from different angles, this approach increases the likelihood of destroying both sensitive and resistant cells, potentially delaying the onset of resistance. The exclusive use of monotherapy in this project's experimental setup may not capture the complexities and potential benefits of combination treatments in effectively managing cancer progression and resistance.

In addition, mRNA expression levels serve as an indirect indicator of protein abundance. The correlation between mRNA and protein levels is complex and influenced by post-transcriptional, translational, and post-translational modifications. These layers of regulation could significantly alter protein abundance and, therefore, the phenotype of cells.

6.2 Possible mechanisms of drug resistance to AKT inhibitors

6.2.1 Resistance mechanisms to MK-2206 reported in the literature

Several studies have explored the mechanisms by which cancer cells develop resistance to AKT inhibitors like MK-2206, uncovering various strategies that enable cancer cells to evade treatment.

Qi *et al*., investigated the mechanisms of acquired resistance to MK-2206 in neuroblastoma (NB) cell lines, and found that resistance was primarily associated with the activation of alternative signalling pathways, specifically the PDK1-mTOR-S6K pathway, and in some cases, the MAPK pathway (430). Inhibiting these pathways with PDK1 (GSK2334470) or mTOR (AZD8055) inhibitors effectively suppressed cell growth by arresting MK-2206-resistant cells in the G0-G1 phase. These findings suggest that combination therapy may be necessary to address the complex resistance mechanisms to MK-2206 in NB.

The development of MK-2206-resistant breast cancer cell lines with a PIK3CA mutation highlighted a notable upregulation of *AKT3*, while *AKT1* and *AKT2* levels remained unchanged at both the mRNA and protein levels (431). Functional assays confirmed that this *AKT3* upregulation was the key driver of resistance, as its depletion restored the sensitivity of cells to MK-2206. The upregulation of *AKT3* was found to be mediated through epigenetic mechanisms involving bromodomain and extra-terminal domain (BET) proteins, which are known to regulate various components of the PI3K pathway, including IGF1R (430, 432). Resistant cells also exhibited an epithelial-to-mesenchymal transition (EMT) phenotype, characterised by increased invasiveness and reduced E-cadherin expression. Notably, depleting AKT3 not only reversed the EMT phenotype but also reduced the aggressive and resistant nature of the cells. These findings present *AKT3* upregulation as a novel mechanism of acquired resistance to MK-2206 in breast cancer, pointing to the potential of targeting this kinase or its epigenetic regulators to overcome resistance. Additionally, this research highlights the need for developing AKT3-selective inhibitors to mitigate the potential side effects associated with pan-AKT inhibitors like MK-2206.

In Tsang's research, MK-2206-resitant breast cancer cells exhibited cross-resistance to the ATP-competitive AKT inhibitor GDC0068 (433). In these resistant cells, there was a notable upregulation of *AKT3* and *IGF1R* expression—matching the results of Stottrup's research (431)—and an upregulation of phosphorylated EGFR (pEGFR). Treating resistant cells with a combination of MK-2206 and the EGFR inhibitor gefitinib significantly reduced cell viability

and AKT phosphorylation, confirming that EGFR activation played a key role in resistance. Resistant cells also displayed enhanced cancer stem cell properties, including increased mammosphere formation—mammary epithelial stem cell aggregates—and upregulation of the transcription factors Slug and ID4, a key regulator of mammary stem cells that is associated with a CSC-like phenotype and poor prognosis in triple-negative breast cancer (TNBC) (434). The upregulation of ID4 was accompanied by the downregulation of key downstream targets that are typically suppressed by ID4, such as Brca1 and components of the Notch signalling pathway (e.g., Hey1, Notch1). Interestingly, knocking down *ID4* in resistant cells led to a decrease in the expression of several stemness-related genes (including Twist1, Twist2, Snail, Slug), suggesting that ID4 is crucial for maintaining the enhanced CSC properties observed in resistant cells (433).

6.2. Possible mechanisms of drug resistance to AKT inhibitors

A. Potential resistance mechanisms in MK1-resistant mCRC PDOs and therapeutic strategies to overcome it

While the studies above identified critical pathways and molecular changes contributing to MK-2206 resistance, the transcriptional changes observed in MK1-resistant mCRC PDOs revealed distinct, yet complementary, mechanisms of resistance. In the MK1-resistant PDO, our research identified a resistance mechanism primarily contributing to metabolic reprogramming through the Warburg effect. Additionally, the increased frequency of Cluster 1 cells in this organoid, which was linked to a stem-like gene expression, suggests that stemness played a role in driving resistance in this organoid.

The Warburg effect provides cancer cells with metabolic flexibility, allowing them to produce ATP via aerobic glycolysis even in oxygen-rich conditions, instead of relying on the more energy-efficient oxidative phosphorylation. This metabolic adaptation facilitates the uptake and incorporation of essential biosynthetic intermediates into the biomass—the total mass of organic material that makes up a cell—such as ribose for nucleotide synthesis and glycerol and citrate for lipid synthesis, thereby providing the necessary materials to support rapidly proliferating cells (435, 436). However, this leads to an increased uptake of glucose—facilitated in MK1-resistant PDO by the overexpression of the glucose transported *GLUT1*—and increased production of lactate. The upregulation of *PDK3* in the MK1-resistant organoid further promoted lactate production by inhibiting the activity of the pyruvate dehydrogenase (PDH) complex, which normally converts pyruvate into acetyl-CoA before it enters the mitochondria for oxidative phosphorylation (243).

The Warburg effect significantly contributes to various mechanisms of drug resistance. In the first instance, the release of high volumes of lactate leads to a decrease in the pH of the tumour microenvironment (TME), resulting in acidosis (437). An acidic TME acts as a chemical barrier, leading to the accumulation of weakly basic chemotherapeutic drugs outside the cells (e.g., doxorubicin), which in turn reduces their effectiveness (438). This accumulation is due to an ion trapping effect, where the drug becomes ionised and is less able to pass through the cell membrane effectively.

The acidification of the TME also promotes several cancer-related processes, such as metastasis, angiogenesis and immunosuppression. Indeed, there is a strong correlation between elevated lactate levels and metastasis in several human cancers, including colorectal adenocarcinoma (439). Lactate produced primarily by tumour cells, with a minor contribution from stromal cells, weakens the immune system by preventing the maturation of dendritic cells,

thereby hindering their ability to activate CD4+ and CD8+ T cells (438). Additionally, lactate induces the apoptosis of immune cells, particularly natural killer (NK) and natural killer T (NKT) cells, and promotes the development of myeloid-derived suppressor cells (MDSCs), which in turn support the growth of regulatory T cells (Tregs). Consequently, lactate is considered an important oncometabolite in certain cancers (440).

The upregulation of *PGD* in the MK1-resistant organoid suggested an increased activation of the pentose phosphate pathway (PPP), which is closely linked to glycolysis. Consequently, elevated levels of aerobic glycolysis in tumours not only generate precursors for nucleotide and amino acid production via the PPP but also help manage oxidative stress by producing ample NADPH (441). This NADPH provides the reducing power necessary for various cellular processes, including the repair of damaged DNA, enabling tumour cells to survive treatments that induce DNA damage, such as radiation therapy and certain chemotherapeutic agents.

In CRC, the Warburg effect is intricately linked to the maintenance of cancer stem cells (CSCs). CSCs often possess characteristics similar to those of normal stem cells, including a higher number of drug transporters, enhanced DNA damage repair capacity, and protective niches, which make them more likely to survive conventional treatments (442). Therefore, understanding the basis of this differential sensitivity is critical for developing more effective cancer therapies that can potentially prevent tumour relapse and metastasis.

Emmink *et al.,* compared the proteins secreted by CSCs with those secreted by their differentiated counterparts and found that the proteins enriched in CSCs included various glycolysis-related enzymes, such as *GPI*, *PGM1*, and *PGM2*, suggesting that these cells rely on aerobic glycolysis for their cancer-promoting functions (443). Furthermore, studies have shown that increasing glucose concentration can elevate the percentage of colon CSCs, indicating a strong link between glucose availability and CSC proliferation (444, 445). Notably, treatment of several human cancer cell lines, including CRC, with the glycolysis inhibitor 3-BrOP significantly reduced the percentage of CSCs and inhibited tumour development (444).

Overall, the Warburg effect plays a pivotal role in various drug resistance mechanisms within CRC. Therefore, targeting glycolysis in resistant cells with elevated glycolytic activity, such as in the MK1-resistant mCRC organoid, presents a potent strategy for overcoming drug resistance.

6.2. Possible mechanisms of drug resistance to AKT inhibitors

6.2.2 Resistance mechanisms to AZD5363/capivasertib reported in the literature

Gris-Oliver *et al*., investigated potential biomarkers and mechanisms of response and resistance to AZD5363/capivasertib in HER2-negative breast cancer using patient-derived xenograft (PDX) models (446). The study revealed that *PIK3CA*/*AKT1* mutations, in the absence of mTORC1-activating alterations, predicted sensitivity to the drug, while low phosphorylated AKT or the presence of mTORC1-activating mutations indicated resistance. Capivasertib's mechanism of action was primarily through cyclin D1 downregulation and cell cycle arrest. Acquired resistance to capivasertib could arise from cyclin D1 amplification or loss of the *AKT1 E17K* mutation. Additionally, the combination of capivasertib and paclitaxel showed promise in *TP53* wild-type tumours, with p53 activation potentially contributing to its efficacy.

Similarly, Dunn *et al*., investigated capivasertib resistance in PTEN-deficient breast cancer cell lines (447). They discovered that resistance to capivasertib specifically was associated with the loss of *TSC1/2* or *STK11*, leading to persistent mTORC1 signalling activation. Importantly, the study demonstrated that the dual inhibition of AKT with capivasertib and the anti-apoptotic protein Mcl-1 with AZD5991 could restore sensitivity to PI3K/AKT targeted therapy, ultimately inducing apoptosis in resistant cells.

Jakubowski investigated mechanisms of acquired resistance to capivasertib in ovarian cancer cell lines with hyperactivated PI3K/AKT/mTOR (PAM) signalling, and found that resistance was associated with increased cap-dependent protein synthesis (CDPS), a critical process for translating mRNA into proteins (448). This increase in CDPS was driven by reduced activity of 4EBP1, a protein that normally suppresses translation initiation. Restoring 4EBP1 function partially reversed resistance. This connection between protein synthesis and drug resistance is particularly intriguing given the downregulation of ribosomal-related proteins observed in both AKTi-resistant mCRC PDOs compared to the untreated Parental organoid.

The mutational landscape of the AZD1-resistant mCRC PDO, characterised by mutations in *TP53* and *AKT1* (Figure 2.1), but lacking known *mTORC1*-activating alterations, suggests that the resistance mechanisms seen in breast and ovarian cancers may not fully explain the resistance in our model.

6.2. Possible mechanisms of drug resistance to AKT inhibitors

A. Potential resistance mechanisms in AZD1-resistant mCRC PDOs and therapeutic strategies to overcome it

The AZD1-resistant PDO exhibited a complex molecular profile, indicating multiple resistance mechanisms. Notably, this organoid showed increased activity in pathways related to tumour progression, including cell adhesion, migration, and extracellular matrix remodelling. There was also upregulation of genes linked to immune evasion, suggesting mechanisms that suppress immune cell function and create an immunosuppressive microenvironment. Additionally, overactive Wnt/β-catenin and EGFR signalling pathways were observed, along with a metabolic shift characterised by downregulation of the protein synthesis machinery, potentially reflecting compensatory mechanisms to meet increased energy demands.

In the AZD1-resistant PDO, immune suppression particularly emerged as a critical factor in the development of drug resistance. The tumour microenvironment (TME) in CRC is often immunosuppressive, shaped by a complex network of factors, including manipulation of immune cell populations, alteration of cell surface proteins and dysregulation of cytokines and chemokines (449). For example, tumour cells exploit immune checkpoints like the PD-1/PD-L1 axis to evade immune surveillance (449). By upregulating PD-L1, tumours can engage PD-1 on T cells, inhibiting their anti-tumour activity. This mechanism is more prevalent in metastatic CRC compared to primary CRC, resulting in decreased responsiveness to PD-1/PD-L1-targeting immunotherapies (449, 450).

Additionally, the tumour actively recruits immunosuppressive cells, such as T regulatory cells (Tregs), tumour-associated macrophages (TAMs), and myeloid-derived suppressor cells (MDSCs), into the TME to establish a protective niche (449). Chemokines and cytokines secreted by the tumour, such as CCL20 from TAMs, facilitate the recruitment of Tregs, further amplifying immunosuppression in CRC (449, 451).

Metabolic reprogramming within the TME also contributes to drug resistance. For example, the enzyme IDO1 plays a significant role in creating an immunosuppressive environment in CRC (449). IDO1, induced by the cytokine IFN-γ in various immune and non-immune cells (such as dendritic cells, macrophages, and fibroblasts), breaks down tryptophan into kynurenine, a metabolite that inhibits T cell function and promotes Treg development. TAMs can exacerbate this by causing T cell starvation through products generated via the IDO1/2 pathway, further suppressing T cell activity (449, 452). *IDO1* is often overexpressed in CRC and is associated with poorer patient outcomes (449, 453). Recent research showed that reducing *IDO1* expression using short hairpin RNAs (shRNAs) led to a higher presence of neutrophils in tumours and

slowed their growth in murine models of CRC, highlighting its role as a systemic immunosuppressor (449, 453, 454).

Hypoxia is a common feature of the TME that exacerbates immunosuppression by promoting the recruitment and function of inhibitory immune cells such as Tregs, TAMs, and MDSCs, while also enhancing the suppressive activity of dendritic cells (449). Other factors within the TME, particularly the accumulation of extracellular adenosine, significantly contribute to immune evasion (252, 449). Adenosine, produced by cell surface enzymes like CD39 (*ENTPD1*)—which was overexpressed in the AZD1-resistant PDO (Table 15)—acts as a potent immunosuppressive signal. Hypoxia itself can trigger increased adenosine production, further amplifying its inhibitory effects (449).

To counteract the immunosuppressive TME in CRC, several therapeutic strategies are being actively pursued. While immune checkpoint blockade with anti-PD-1/PD-L1 antibodies has proven effective in microsatellite instability-high (MSI-H) CRCs, patients with advanced microsatellite stable (MSS) CRCs often exhibit resistance when it is used as a monotherapy (449, 455). To improve treatment outcomes, researchers have explored combining anti-PD-1 or anti-PD-L1 therapies (e.g., durvalumab, atezolizumab) with anti-CTLA-4 therapy (e.g., tremelimumab). Although this combination demonstrated a modest increase in overall survival for MSS CRC patients who had previously undergone chemotherapy, the results were not statistically significant (449, 456).

Another promising approach involves directly targeting immunosuppressive cytokines and cells. For instance, blocking VEGF with bevacizumab has shown efficacy in reducing Tregs in both mouse models and CRC patients (449). However, for patients with MSS CRC, combining atezolizumab (an anti-PD-L1 treatment) with fluorouracil and bevacizumab as a first-line treatment did not yield improved effectiveness.

Researchers are also investigating ways to actively recruit T cells into the immunosuppressive TME using T cell bispecific (TCB) antibodies (449). For example, carcinoembryonic antigen (CEA) TCB antibodies (e.g., RG7802, RO6958688) are designed to bind both CEA on tumour cells and CD3 on T cells, facilitating the activation of effector T cells to target and kill CEA-expressing cancer cells (449, 457). In a phase 1 clinical trial, CEA-TCB was tested alone and in combination with the PD-L1 inhibitor atezolizumab in patients with MSS CRC (457). The results showed that CEA-TCB monotherapy led to increased CD3+ T cell infiltration and antitumor activity, while its combination with atezolizumab demonstrated enhanced efficacy and maintained a manageable safety profile. This study is particularly relevant given the

upregulation of *CEACAM6*, a member of the CEA family, in the MK1-resistant organoid (Table 13), further supporting the role of dysregulated immune regulation in both AKTi-resistant mCRC organoids.

Finally, adoptive cell transfer (ACT)—which involves the transfer of ex-vivo expanded or engineered T cells or natural killer (NK) cells—and chimeric antigen receptor (CAR)-T cell therapy, represent approaches aimed at enhancing anti-tumour immunity within the TME by increasing T cell trafficking and persistence in solid tumours (449). In a Phase I/II clinical study, T lymphocytes from sentinel lymph nodes (SLNs) were expanded outside the body and then infused into CRC patients who had undergone either radical or palliative surgery (458). The results showed that the 24-month survival rate for patients receiving SLN-T lymphocytes was significantly higher (55.6%) compared to the control group (17.5%), indicating that SLN-T lymphocyte immunotherapy is safe and may enhance overall survival in patients with metastatic CRC.

Although enhanced immune-suppression could be one possible mechanism of resistance in AZD1-resistant cells, another potential mechanism could involve off-target effects of AZD5363. Whelan investigated the mechanisms of acquired resistance to AZD5363 in breast cancer cell lines with PI3K/AKT pathway mutations and found no cross-resistance to other AKT inhibitors, such as MK-2206 (459). Furthermore, there were no significant changes in PI3K/AKT signalling, indicating that the resistance was not directly linked to AKT inhibition. This finding aligns with the current thesis, where no genes directly downstream of the AKT protein were found to be differentially expressed in either MK1- or AZD1-resistant mCRC PDOs. Whelan's study suggested that the resistance could be due to the inhibition of cAMP-dependent protein kinase (PKA), another target of AZD5363, with the potential upregulation of PKA-proximal signalling pathways circumventing AKT inhibition. Notably, AZD5363 inhibits PKA with similar potency to its inhibition of AKT2/3, but not AKT1 (427). This observation is particularly relevant to the current research, where the development of AZD5363-resistant mCRC PDOs with an *AKT1* mutation and amplification suggests that the drug's multi-target activity may have contributed to resistance.

Overall, the diverse mechanisms of resistance to the MK-2206 and AZD5363 AKT inhibitors, as demonstrated by both preclinical studies and the findings of this thesis, highlight the complexity of targeting the PI3K/AKT/mTOR pathway in CRC. In particular, the metabolic reprogramming observed in the MK1-resistant organoid, alongside the dysregulation of immune responses in the AZD1-resistant organoid, underscores the multifaceted nature of drug resistance to the same drug target. These findings emphasise the need for personalised

treatment strategies that account not only for the genetic alterations driving tumour growth but also for the dynamic interplay between cancer cells and their microenvironment.

6.3 Novel findings and their potential clinical and research implications

This study provides valuable insights into the complex mechanisms of resistance to AKT inhibition in mCRCs, with potential implications for both clinical practice and the development of future AKT-targeted therapies.

Notably, the presence of a pre-existing resistant clone harbouring unique CNAs at chr2 and chr5 in the Parental organoid derived from a heavily pre-treated patient reinforces the clinical relevance of pre-existing resistance and its role in treatment failure. This finding underscores the need for early detection and characterisation of such clones to guide personalised treatment decisions, potentially incorporating combination therapies or alternative treatment modalities from the outset. In this case, the specific CNA at chr2 emerges as a potential multi-drug resistance marker. The consistent upregulation of *PCYOX1* in the chr2 CNA gain region in both AKTi-resistant organoids highlights it as a potential novel player in the resistance to AKT inhibiton. Further investigation into its function and potential as a therapeutic target may reveal novel strategies to overcome resistance, either by directly targeting *PCYOX1* or developing predictive biomarkers.

The identification of an inherently resistant clone, expanding under the selective pressure of both allosteric (MK-2206) and ATP-competitive (AZD5363) AKT inhibitors, suggests a novel resistance mechanism independent of specific drug-target interactions. Distinct gene expression profiles and enriched pathways in the MK1- and AZ1-resistant organoids further underscore the importance of personalised treatment selection, demonstrating that the choice of AKT inhibitor can significantly influence the transcriptional landscape and phenotype of resistant cells.

While CNAs are known to influence gene expression, this study showed variations in the relationship between CNAs and gene expression depending on the specific AKT inhibitor used. The discordance observed between CNAs and gene expression patterns in the AZD1-resistant organoid, particularly the upregulation of genes on chr5 despite a CNA gain, underscores the complexity of gene regulation in the context of drug resistance. This suggests that the impact of CNAs on gene expression may be context-dependent, and highlights the importance of considering epigenetic modifications and other regulatory factors that may influence gene expression beyond simple gene dosage effects. Understanding these intricacies could lead to the identification of novel therapeutic targets or the development of epigenetic therapies to overcome resistance. The identification of potential resistance mechanisms to AZD5363 in

mCRC PDOs is particularly timely and clinically relevant, as this drug has recently been approved by the FDA for the treatment of certain metastatic breast cancers (150). To our knowledge, this is the first study to characterise resistance mechanisms to AZD5363 in colorectal cancers harbouring a mutation and an amplification in *AKT1*. This research's findings on the potential role of clonal evolution, epigenetic modifications, and compensatory pathways in driving resistance to AKT inhibition could inform strategies to improve the efficacy of AZD5363, and potentially expand its use to other cancer types, including mCRC.

The development of resistant PDO models provides a unique and powerful platform to address the challenges posed by drug resistance. By generating mCRC models of acquired resistance to the AKT inhibitors MK-2206 and AZD5363, this research recapitulated not only the cellular heterogeneity typically observed in human tumours, but also the adaptive responses that lead to treatment evasion. These models also offer a valuable platform for directly testing potential therapies, assessing their efficacy in resensitising resistant tumours or circumventing the resistance mechanisms entirely. Furthermore, creating resistant organoids from individual patients opens doors to explore personalised medicine and drug screening approaches, aiding in the selection of the most effective treatment options for each patient. Investigating cross-resistance between different drugs in these models can also inform the development of more potent drug combinations. Finally, the use of resistant organoids has the potential to reduce reliance on animal models in preclinical research, thus accelerating drug discovery and addressing ethical concerns.

In conclusion, this research advances our understanding of AKT inhibitor resistance in mCRC, and offers potential avenues for improving clinical outcomes. By elucidating the complex interplay between genetic and epigenetic factors, identifying novel resistance mechanisms, and highlighting potential therapeutic targets like *PCYOX1*, this work may pave the way for the development of more personalised and effective treatment strategies for patients with metastatic colorectal cancers.

6.4 Future perspectives

Future research directions would include a comprehensive analysis of single nucleotide variants (SNVs), which this study did not investigate. SNVs could offer valuable insights into the genetic foundations of drug resistance and enable tracking of the evolution of cancer subclones.

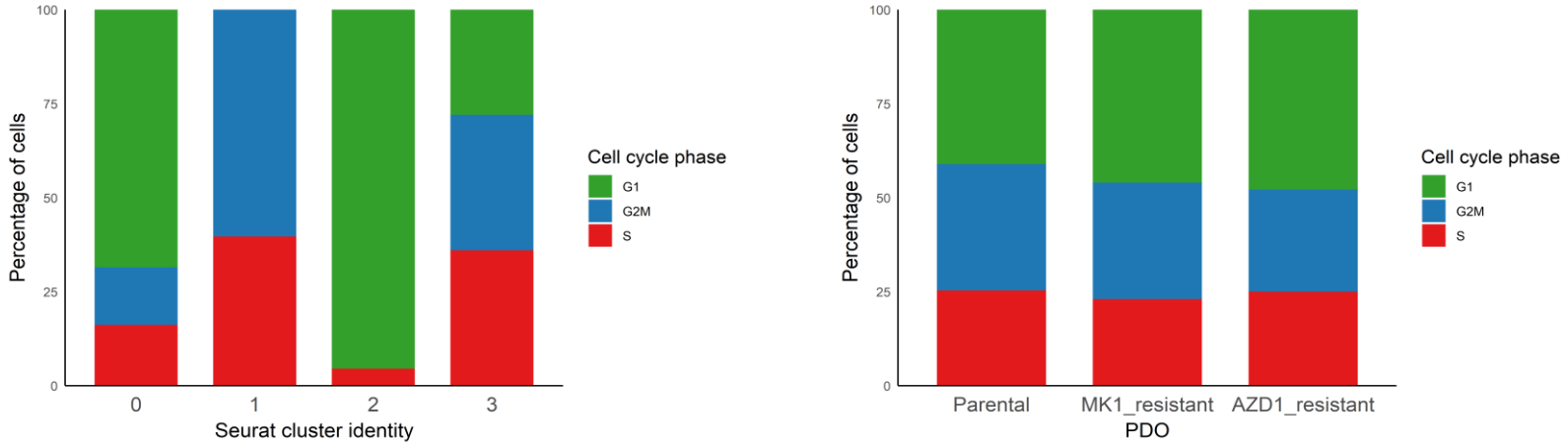
A significant advancement would be extending G&T-seq analyses to primary tumours, distant metastases, and normal mucosa from the same patient. This strategy would permit the study of subclonal evolution from cancer initiation to metastasis, potentially identifying early markers of resistance. Moreover, integrating G&T-seq with laser capture microdissection, as previously demonstrated in a multiregional breast cancer study (460), would allow an analysis of the role of the tumour microenvironment on the development of drug resistance. This aspect is particularly intriguing given that numerous genes involved in extracellular matrix remodelling and the regulation of immune responses were dysregulated in the present study, pointing to the role of the tumour microenvironment in the emergence of drug resistance.

Additionally, employing targeted genetic manipulation techniques such as CRISPR-mediated transcriptional activation (CRISPRa) or CRISPR-Cas9-mediated gene knockout, alongside RNA interference (RNAi) strategies like siRNA or shRNA (as previously suggested in Chapter 5), could be used to overexpress or inhibit potential driver genes of AKTi resistance identified on chromosomes 2 (*PCYOX1*) and 5 (*CXCL14*, *NPM1*, *RACK1, RPS14*). This could involve targeting genes on each chromosome separately, as well as simultaneously, to elucidate their individual contributions and potential synergistic effects on resistance. This functional validation approach could confirm the role of these genes in resistance and potentially uncover new therapeutic targets.

In conclusion, this thesis demonstrates the potential of G&T-seq and multiomics data integration to unravel the complex mechanisms of resistance in metastatic colorectal cancer. While challenges remain in data interpretation, this powerful approach allows researchers to identify genetic alterations and their impact on gene expression, paving the way for the discovery of novel therapeutic targets. The insights gained from this research underscore the importance of considering both genomic and transcriptomic landscapes when developing new treatment strategies and highlight the potential of personalised medicine for overcoming drug resistance in metastatic colorectal cancer. Future research building upon these findings will further refine our understanding of resistance mechanisms and ultimately contribute to improved patient outcomes.

Appendix A. Supplementary material for Chapter 3

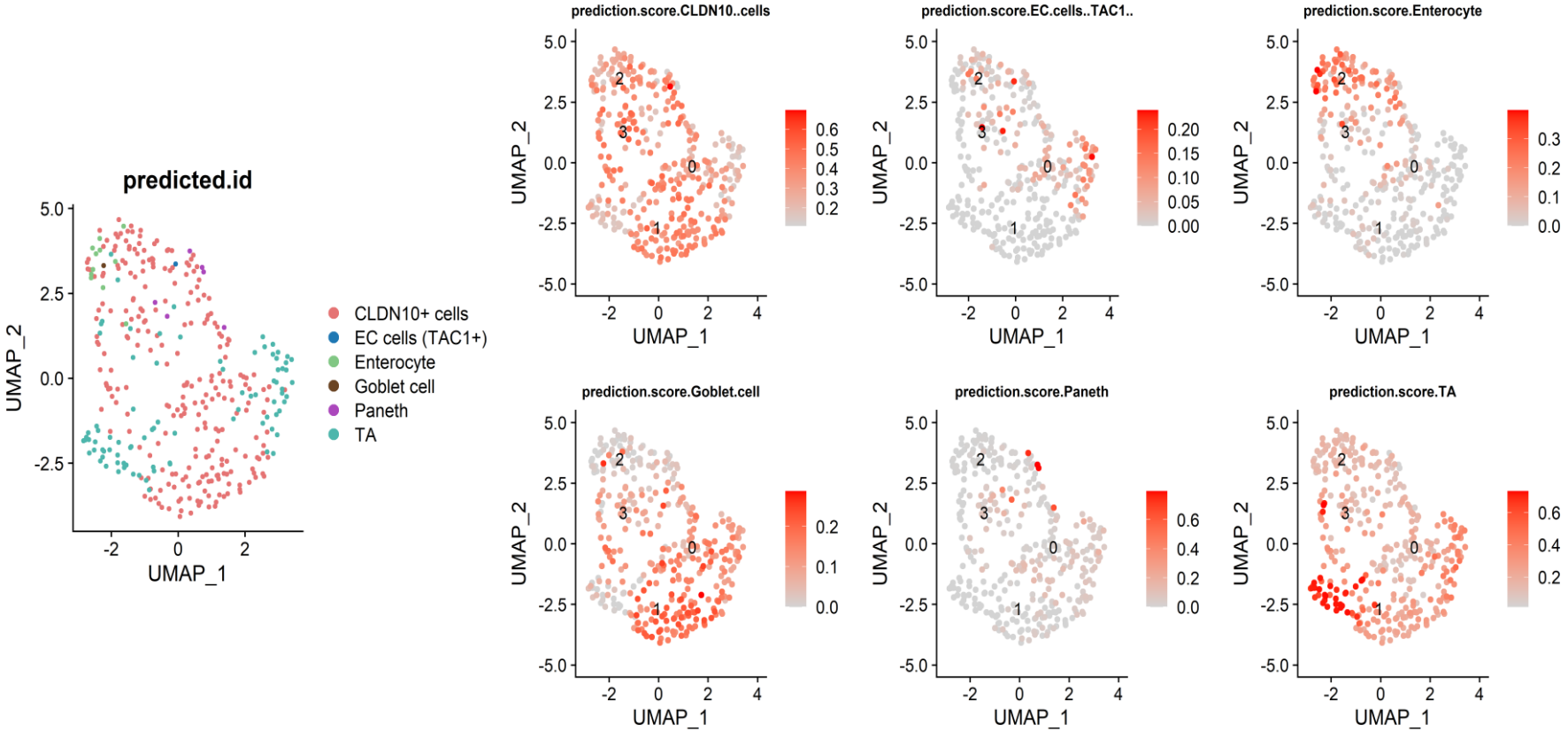
A.1 Distribution of cell cycle phases in Smart-seq2 data



Supplementary Figure 1. Distribution of cell cycle phases across Seurat clusters and mCRC PDOs in Smart-seq2 data.

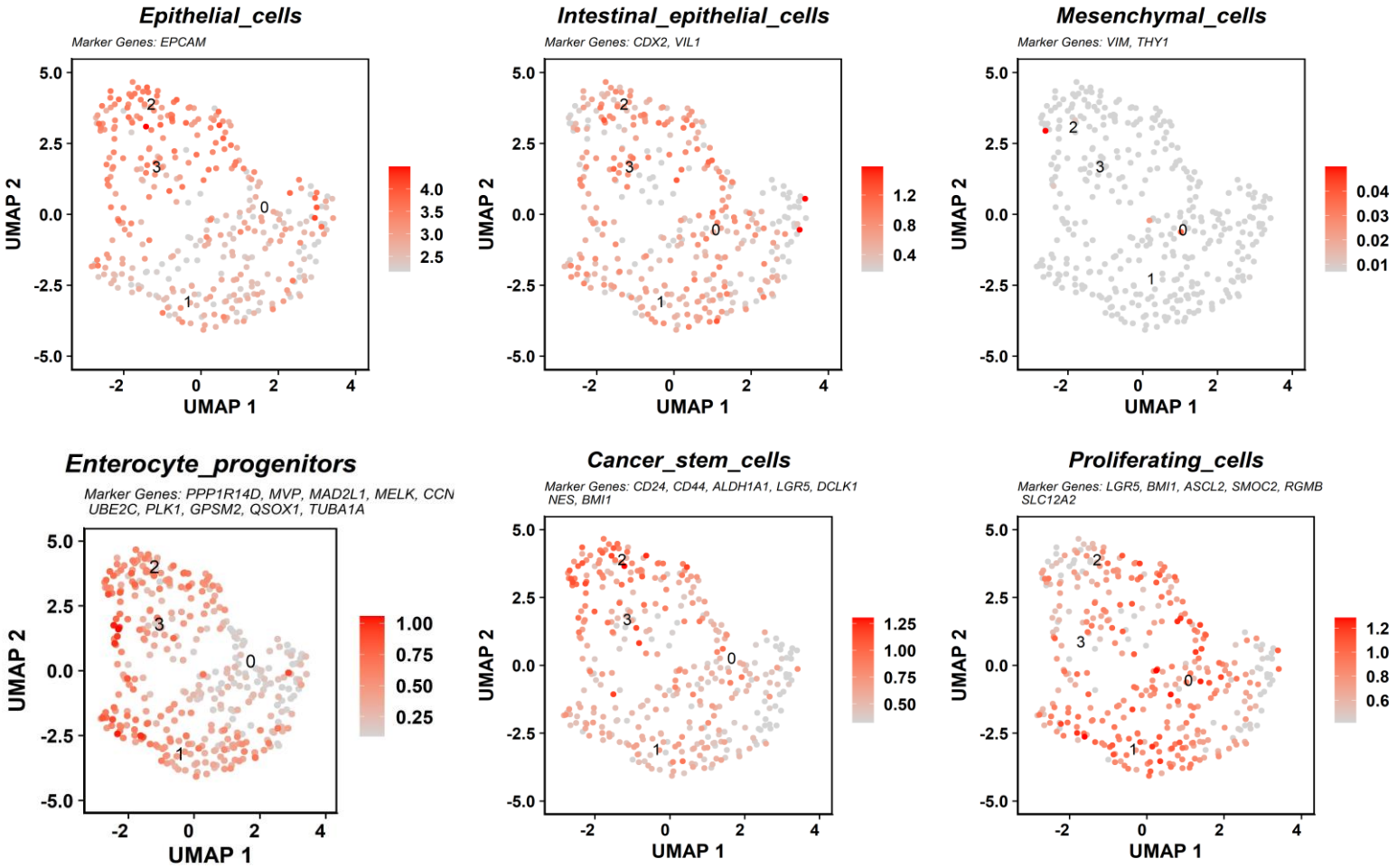
Stacked bar plot represents the distribution of cell cycle phases across clusters identified in Smart-seq2 data (left panel) and the general distribution of cell cycle phases across the three mCRC PDOs (right panel).

A.2 Cell-type annotation of Smart-seq2 data using the Human Gut Cell Atlas



Supplementary Figure 2. Prediction scores for colonic cell types identified in mCRC PDOs using the Human Gut Cell Atlas.

UMAP plots display the distribution of colonic cell types in mCRC PDOs identified in the Smart-seq2 dataset (left) and the corresponding prediction scores (right). These scores reflect the confidence in cell type classification and the gene expression similarity between cells in the Smart-seq2 queried dataset and the reference cell types in the Human Gut Cell Atlas.



Supplementary Figure 3. UMAP plots showing the expression patterns of selected marker genes used to identify different cell types

A.3 Smart-seq2 differential gene expression analysis of AKTi-resistant mCRC PDOs

Supplementary Table 1. Full list of statistically significant dysregulated genes in the MK1-resistant PDO

Gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	pct_diff	Gene_id	Expression
MUC21	2.02E-08	2.384874	0.747	0.47	0.000487	0.277	ENSG00000204544	Upregulated
CXCL14	2.05E-07	2.240981	0.954	0.978	0.004935	-0.024	ENSG00000145824	Upregulated
TACSTD2	3.43E-08	2.199259	0.897	0.858	0.000826	0.039	ENSG00000184292	Upregulated
LGALS1	1.63E-09	1.95564	0.92	0.873	3.93E-05	0.047	ENSG00000100097	Upregulated
CEACAM6	1.15E-06	1.953189	0.828	0.784	0.027699	0.044	ENSG00000086548	Upregulated
LRATD1	4.27E-08	1.4727	0.793	0.582	0.001028	0.211	ENSG00000162981	Upregulated
FABP5	6.71E-09	1.315372	0.931	0.97	0.000162	-0.039	ENSG00000164687	Upregulated
SDR16C5	2.35E-08	1.22146	0.678	0.433	0.000566	0.245	ENSG00000170786	Upregulated
EPAS1	2.13E-10	1.173951	0.874	0.903	5.12E-06	-0.029	ENSG00000116016	Upregulated
SLC2A1	2.14E-07	1.155772	0.851	0.731	0.005162	0.12	ENSG00000117394	Upregulated
PDK3	4.26E-07	1.065709	0.943	1	0.01026	-0.057	ENSG00000067992	Upregulated
LGALS3	8.50E-09	1.045865	0.92	0.978	0.000205	-0.058	ENSG00000131981	Upregulated
ENO1	4.63E-14	1.004288	0.966	0.985	1.12E-09	-0.019	ENSG00000074800	Upregulated
S100A4	3.09E-09	0.982926	0.977	0.985	7.44E-05	-0.008	ENSG00000196154	Upregulated
PRSS8	3.46E-08	0.971759	0.851	0.784	0.000834	0.067	ENSG00000052344	Upregulated
CD320	4.47E-10	0.883349	0.874	0.843	1.08E-05	0.031	ENSG00000167775	Upregulated
MT-ND1	4.92E-18	0.846994	1	0.993	1.19E-13	0.007	ENSG00000198888	Upregulated
MT-ND2	2.18E-19	0.846213	0.989	1	5.24E-15	-0.011	ENSG00000198763	Upregulated
YWHAZ	3.91E-11	0.840387	0.977	1	9.43E-07	-0.023	ENSG00000164924	Upregulated
MT-ND6	1.67E-12	0.835848	0.874	0.94	4.02E-08	-0.066	ENSG00000198695	Upregulated
DOCK11	1.34E-09	0.82746	0.667	0.358	3.23E-05	0.309	ENSG00000147251	Upregulated
ATP6AP1	1.36E-06	0.824049	0.874	0.94	0.032849	-0.066	ENSG00000071553	Upregulated
KLK10	2.45E-10	0.816882	0.425	0.082	5.90E-06	0.343	ENSG00000129451	Upregulated
PGD	5.84E-07	0.794954	0.874	0.873	0.014073	0.001	ENSG00000142657	Upregulated
MT-CYB	1.82E-16	0.782423	1	1	4.40E-12	0	ENSG00000198727	Upregulated
TSPAN1	6.82E-07	0.780536	0.54	0.246	0.016427	0.294	ENSG00000117472	Upregulated
PREB	1.45E-07	0.779376	0.851	0.888	0.003483	-0.037	ENSG00000138073	Upregulated

ACTB	3.64E-13	0.771538	1	1	8.76E-09	0	ENSG00000075624	Upregulated
PDIA6	1.97E-11	0.763966	0.908	0.933	4.76E-07	-0.025	ENSG00000143870	Upregulated
GPRC5A	9.25E-08	0.755597	0.667	0.328	0.002229	0.339	ENSG00000013588	Upregulated
MT-ND3	4.68E-13	0.751087	0.931	0.97	1.13E-08	-0.039	ENSG00000198840	Upregulated
ELP1	1.95E-06	0.738345	0.816	0.672	0.047071	0.144	ENSG00000070061	Upregulated
PCYOX1	1.28E-07	0.7335	0.885	0.903	0.003083	-0.018	ENSG00000116005	Upregulated
CSTB	1.62E-06	0.707238	0.851	0.858	0.038961	-0.007	ENSG00000160213	Upregulated
SYVN1	1.19E-06	0.681213	0.747	0.627	0.028645	0.12	ENSG00000162298	Upregulated
S100A16	1.21E-06	0.644071	0.897	0.91	0.029076	-0.013	ENSG00000188643	Upregulated
MT-ND5	1.24E-14	0.642089	0.989	1	2.98E-10	-0.011	ENSG00000198786	Upregulated
TMSB4X	1.99E-10	0.640867	1	1	4.80E-06	0	ENSG00000205542	Upregulated
PDIA3	1.78E-07	0.640724	0.966	0.985	0.004295	-0.019	ENSG00000167004	Upregulated
SPINT2	1.90E-06	0.640339	0.874	0.925	0.045719	-0.051	ENSG00000167642	Upregulated
CFL1	4.49E-09	0.635387	0.897	0.933	0.000108	-0.036	ENSG00000172757	Upregulated
RPN2	6.53E-09	0.629856	0.897	0.97	0.000157	-0.073	ENSG00000118705	Upregulated
CLDN1	1.49E-06	0.623463	0.506	0.246	0.035846	0.26	ENSG00000163347	Upregulated
SYNGR2	1.00E-07	0.613673	0.908	0.97	0.002411	-0.062	ENSG00000108639	Upregulated
MLEC	3.39E-07	0.611429	0.874	0.925	0.008179	-0.051	ENSG00000110917	Upregulated
POLR2L	2.22E-07	0.603688	0.885	0.918	0.005352	-0.033	ENSG00000177700	Upregulated
BCAP31	4.47E-07	0.599635	0.885	0.94	0.01078	-0.055	ENSG00000185825	Upregulated
PARK7	2.32E-07	0.595866	0.908	0.955	0.005587	-0.047	ENSG00000116288	Upregulated
SULF2	5.06E-07	0.584913	0.483	0.209	0.012184	0.274	ENSG00000196562	Upregulated
YWHAQ	6.93E-08	0.578208	0.943	1	0.00167	-0.057	ENSG00000134308	Upregulated
KRT18	9.28E-07	0.573409	1	1	0.022365	0	ENSG00000111057	Upregulated
GSTP1	7.88E-11	0.550828	0.966	0.993	1.90E-06	-0.027	ENSG00000084207	Upregulated
MT-ND4	1.19E-12	0.547273	1	1	2.86E-08	0	ENSG00000198886	Upregulated
GPX1	1.61E-07	0.540385	0.897	0.948	0.00387	-0.051	ENSG00000233276	Upregulated
TM7SF2	9.61E-08	0.534202	0.667	0.433	0.002316	0.234	ENSG00000149809	Upregulated
SLC6A14	1.10E-10	0.527858	0.655	0.246	2.64E-06	0.409	ENSG00000268104	Upregulated
BSG	1.26E-07	0.505266	0.954	0.985	0.003024	-0.031	ENSG00000172270	Upregulated
RPS16	1.77E-09	-0.50063	1	1	4.27E-05	0	ENSG00000105193	Downregulated

RPL13	3.10E-07	-0.50106	0.943	1	0.007467	-0.057	ENSG00000167526	Downregulated
RPL11	4.28E-09	-0.50337	0.977	1	0.000103	-0.023	ENSG00000142676	Downregulated
RPL4	7.44E-08	-0.51417	0.931	0.985	0.001793	-0.054	ENSG00000174444	Downregulated
RPL35A	1.41E-12	-0.51548	0.966	1	3.39E-08	-0.034	ENSG00000182899	Downregulated
SNHG6	9.59E-09	-0.52521	0.851	0.888	0.000231	-0.037	ENSG00000245910	Downregulated
TMED7	2.11E-08	-0.53765	0.77	0.881	0.000507	-0.111	ENSG00000134970	Downregulated
EEF1A1	2.31E-11	-0.54214	1	1	5.57E-07	0	ENSG00000156508	Downregulated
RPLP0	2.45E-17	-0.54366	1	1	5.91E-13	0	ENSG00000089157	Downregulated
PTPRG	5.54E-09	-0.55266	0.471	0.746	0.000134	-0.275	ENSG00000144724	Downregulated
RPS15A	1.91E-14	-0.55476	0.874	0.94	4.61E-10	-0.066	ENSG00000134419	Downregulated
TMPRSS11E	2.61E-08	-0.55906	0.506	0.821	0.000629	-0.315	ENSG00000087128	Downregulated
RPS11	7.34E-14	-0.56733	0.966	1	1.77E-09	-0.034	ENSG00000142534	Downregulated
RPS9	1.89E-09	-0.57374	0.954	0.978	4.56E-05	-0.024	ENSG00000170889	Downregulated
RPS13	9.24E-13	-0.57705	0.989	1	2.23E-08	-0.011	ENSG00000110700	Downregulated
RPS4X	2.02E-11	-0.58727	1	1	4.88E-07	0	ENSG00000198034	Downregulated
EIF3E	3.62E-09	-0.59133	0.897	0.978	8.73E-05	-0.081	ENSG00000104408	Downregulated
HNRNPH1	4.88E-09	-0.59396	0.874	0.978	0.000118	-0.104	ENSG00000169045	Downregulated
RPL10	1.53E-10	-0.59423	0.966	1	3.68E-06	-0.034	ENSG00000147403	Downregulated
RPL15	8.54E-17	-0.59504	0.989	1	2.06E-12	-0.011	ENSG00000174748	Downregulated
SETD5	1.79E-06	-0.59507	0.736	0.851	0.043232	-0.115	ENSG00000168137	Downregulated
RPL27A	1.43E-12	-0.59616	0.954	1	3.45E-08	-0.046	ENSG00000166441	Downregulated
MRPL22	1.27E-07	-0.60432	0.828	0.955	0.003059	-0.127	ENSG00000082515	Downregulated
SLC25A6	1.58E-08	-0.60879	0.862	0.97	0.000381	-0.108	ENSG00000169100	Downregulated
EEF1A1P5	1.05E-12	-0.60988	0.828	0.903	2.53E-08	-0.075	ENSG00000196205	Downregulated
RPL3	6.34E-09	-0.61117	0.966	0.993	0.000153	-0.027	ENSG00000100316	Downregulated
SSB	1.05E-06	-0.62007	0.839	0.955	0.025272	-0.116	ENSG00000138385	Downregulated
RACK1	1.76E-10	-0.62172	0.989	1	4.23E-06	-0.011	ENSG00000204628	Downregulated
RPL37	1.34E-11	-0.62558	0.977	1	3.22E-07	-0.023	ENSG00000145592	Downregulated
RPL30	3.68E-13	-0.62675	0.885	0.993	8.87E-09	-0.108	ENSG00000156482	Downregulated
RPL7A	9.83E-13	-0.62787	0.943	1	2.37E-08	-0.057	ENSG00000148303	Downregulated
KRR1	3.41E-07	-0.64454	0.793	0.896	0.008218	-0.103	ENSG00000111615	Downregulated

RPL24	1.31E-14	-0.65311	0.943	0.993	3.16E-10	-0.05	ENSG00000114391	Downregulated
SNHG29	5.82E-12	-0.65359	0.805	0.896	1.40E-07	-0.091	ENSG00000175061	Downregulated
TBCA	5.82E-08	-0.67622	0.897	0.993	0.001401	-0.096	ENSG00000171530	Downregulated
RPL5	5.29E-14	-0.68163	0.966	1	1.27E-09	-0.034	ENSG00000122406	Downregulated
GDI2	8.57E-08	-0.68785	0.828	0.91	0.002064	-0.082	ENSG00000057608	Downregulated
SH3BP4	1.23E-06	-0.68826	0.655	0.821	0.029546	-0.166	ENSG00000130147	Downregulated
RPL32	2.04E-12	-0.69811	0.943	1	4.92E-08	-0.057	ENSG00000144713	Downregulated
NPM1	3.44E-12	-0.69942	0.977	1	8.29E-08	-0.023	ENSG00000181163	Downregulated
RPS18	1.20E-16	-0.70107	0.989	1	2.90E-12	-0.011	ENSG00000231500	Downregulated
RPS23	1.39E-14	-0.7071	0.874	0.94	3.35E-10	-0.066	ENSG00000186468	Downregulated
NAP1L1	6.26E-11	-0.70967	0.874	0.97	1.51E-06	-0.096	ENSG00000187109	Downregulated
SLC12A2	4.77E-10	-0.71599	0.897	1	1.15E-05	-0.103	ENSG00000064651	Downregulated
RPL6	3.33E-14	-0.71739	0.966	1	8.01E-10	-0.034	ENSG00000089009	Downregulated
RPL12	2.15E-11	-0.72887	0.954	0.985	5.17E-07	-0.031	ENSG00000197958	Downregulated
RPS20	1.80E-18	-0.73203	0.966	1	4.33E-14	-0.034	ENSG00000008988	Downregulated
RPS3	2.64E-13	-0.73325	0.931	0.993	6.36E-09	-0.062	ENSG00000149273	Downregulated
RPL27	1.11E-13	-0.73412	0.943	0.993	2.68E-09	-0.05	ENSG00000131469	Downregulated
RPS27	6.54E-13	-0.73518	1	1	1.58E-08	0	ENSG00000177954	Downregulated
RPS24	8.18E-18	-0.74385	0.989	1	1.97E-13	-0.011	ENSG00000138326	Downregulated
RPL10A	1.05E-13	-0.74422	0.931	1	2.53E-09	-0.069	ENSG00000198755	Downregulated
EEF1B2	2.96E-14	-0.7582	0.931	1	7.13E-10	-0.069	ENSG00000114942	Downregulated
RPS25	3.11E-15	-0.75868	0.931	1	7.49E-11	-0.069	ENSG00000118181	Downregulated
RPS12	8.80E-12	-0.77391	0.954	1	2.12E-07	-0.046	ENSG00000112306	Downregulated
GAS5	9.68E-14	-0.77893	0.897	1	2.33E-09	-0.103	ENSG00000234741	Downregulated
LGR5	1.77E-10	-0.80647	0.529	0.836	4.26E-06	-0.307	ENSG00000139292	Downregulated
RPS14	1.08E-15	-0.80903	1	1	2.59E-11	0	ENSG00000164587	Downregulated
RPL31	4.84E-20	-0.82388	1	1	1.17E-15	0	ENSG00000071082	Downregulated
RPL21	2.60E-19	-0.82715	0.989	1	6.26E-15	-0.011	ENSG00000122026	Downregulated
RPS8	3.69E-19	-0.82849	0.92	1	8.88E-15	-0.08	ENSG00000142937	Downregulated
TPT1	1.11E-15	-0.83226	0.897	0.985	2.68E-11	-0.088	ENSG00000133112	Downregulated
RPL34	3.85E-18	-0.8358	0.954	1	9.27E-14	-0.046	ENSG00000109475	Downregulated

FOXP1	2.35E-14	-0.84289	0.575	0.896	5.66E-10	-0.321	ENSG00000114861	Downregulated
RDH10	2.25E-10	-0.88481	0.747	0.91	5.42E-06	-0.163	ENSG00000121039	Downregulated
RPL9	1.57E-20	-0.89717	0.943	1	3.78E-16	-0.057	ENSG00000163682	Downregulated
RPL13A	3.48E-19	-0.91091	0.977	1	8.39E-15	-0.023	ENSG00000142541	Downregulated
RPL22L1	5.88E-13	-0.93615	0.759	0.94	1.42E-08	-0.181	ENSG00000163584	Downregulated
RPL23	1.57E-18	-0.98915	0.908	1	3.78E-14	-0.092	ENSG00000125691	Downregulated
IGFBP2	3.22E-18	-1.00172	0.345	0.851	7.76E-14	-0.506	ENSG00000115457	Downregulated
RPS3A	8.11E-23	-1.02767	0.943	1	1.95E-18	-0.057	ENSG00000145425	Downregulated
RPL7	5.61E-17	-1.13727	1	1	1.35E-12	0	ENSG00000147604	Downregulated
RPS6	4.39E-25	-1.1452	0.977	1	1.06E-20	-0.023	ENSG00000137154	Downregulated
DEFA5	1.23E-07	-2.63208	0.276	0.627	0.002952	-0.351	ENSG00000164816	Downregulated

Supplementary Table 2. Full list of statistically significant dysregulated genes in the AZD1-resistant PDO

Gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	pct_diff	Gene_id	Expression
C6orf15	7.57E-12	2.225761	0.793	0.552	1.82E-07	0.241	ENSG00000204542	Upregulated
CXCL14	4.46E-12	2.02619	0.986	0.978	1.07E-07	0.008	ENSG00000145824	Upregulated
S100A3	5.14E-11	1.860376	0.957	0.955	1.24E-06	0.002	ENSG00000188015	Upregulated
LINC00867	2.27E-12	1.485194	0.736	0.433	5.47E-08	0.303	ENSG00000232139	Upregulated
CD36	4.88E-12	1.464117	0.836	0.612	1.17E-07	0.224	ENSG00000135218	Upregulated
FNDC3A	1.93E-09	1.417007	0.929	0.903	4.64E-05	0.026	ENSG00000102531	Upregulated
AP1S2	3.46E-16	1.385132	0.871	0.776	8.33E-12	0.095	ENSG00000182287	Upregulated
NEAT1	1.13E-11	1.338621	0.95	1	2.72E-07	-0.05	ENSG00000245532	Upregulated
SEMA3C	1.20E-10	1.280454	0.907	0.843	2.89E-06	0.064	ENSG00000075223	Upregulated
ECM1	5.72E-09	1.230542	0.836	0.724	0.000138	0.112	ENSG00000143369	Upregulated
RBP1	2.19E-07	1.200961	0.507	0.269	0.005286	0.238	ENSG00000114115	Upregulated
SDR16C5	1.63E-12	1.191399	0.693	0.433	3.92E-08	0.26	ENSG00000170786	Upregulated
PRSS23	2.95E-08	1.190268	0.907	0.963	0.00071	-0.056	ENSG00000150687	Upregulated
TACSTD2	1.50E-07	1.146665	0.843	0.858	0.003606	-0.015	ENSG00000184292	Upregulated
CHPF	4.44E-09	1.013184	0.836	0.791	0.000107	0.045	ENSG00000123989	Upregulated
XIST	4.14E-12	0.998503	0.921	0.978	9.98E-08	-0.057	ENSG00000229807	Upregulated
ADAM19	1.58E-08	0.996656	0.707	0.537	0.000381	0.17	ENSG00000135074	Upregulated
PTHLH	7.05E-11	0.960781	0.707	0.403	1.70E-06	0.304	ENSG00000087494	Upregulated
PDP1	5.72E-10	0.917062	0.843	0.716	1.38E-05	0.127	ENSG00000164951	Upregulated
CAMK4	1.51E-06	0.902235	0.507	0.313	0.036458	0.194	ENSG00000152495	Upregulated
TGFA	1.09E-07	0.900429	0.814	0.716	0.002617	0.098	ENSG00000163235	Upregulated
PDK3	5.67E-08	0.900373	0.95	1	0.001367	-0.05	ENSG00000067992	Upregulated
YWHAZ	2.98E-16	0.892769	1	1	7.19E-12	0	ENSG00000164924	Upregulated
S100A4	1.06E-09	0.88518	0.986	0.985	2.56E-05	0.001	ENSG00000196154	Upregulated
FABP5	3.74E-07	0.8718	0.9	0.97	0.009012	-0.07	ENSG00000164687	Upregulated
B3GNT2	1.16E-10	0.868774	0.871	0.799	2.78E-06	0.072	ENSG00000170340	Upregulated
NELL2	4.93E-07	0.86629	0.793	0.657	0.011873	0.136	ENSG00000184613	Upregulated
TPD52L1	8.96E-10	0.856973	0.807	0.604	2.16E-05	0.203	ENSG00000111907	Upregulated

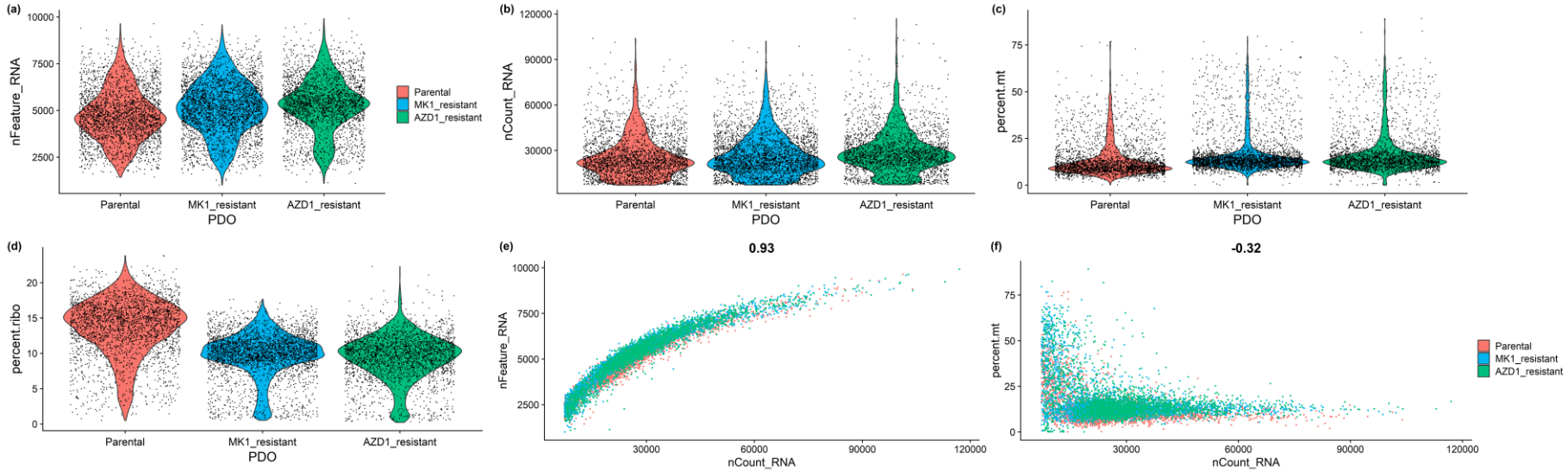
ATP6AP1	1.91E-09	0.847683	0.907	0.94	4.61E-05	-0.033	ENSG00000071553	Upregulated
OSBPL6	3.01E-09	0.825911	0.614	0.291	7.25E-05	0.323	ENSG00000079156	Upregulated
TMSB4X	1.11E-12	0.802858	1	1	2.69E-08	0	ENSG00000205542	Upregulated
CRABP2	1.40E-08	0.766386	0.7	0.478	0.000338	0.222	ENSG00000143320	Upregulated
RCAN2	8.70E-09	0.765857	0.543	0.269	0.00021	0.274	ENSG00000172348	Upregulated
MGST3	1.04E-10	0.744736	0.971	1	2.50E-06	-0.029	ENSG00000143198	Upregulated
DOCK11	3.69E-08	0.735501	0.579	0.358	0.000888	0.221	ENSG00000147251	Upregulated
PPP1CB	9.53E-13	0.733123	0.929	0.985	2.30E-08	-0.056	ENSG00000213639	Upregulated
ENTPD1	1.23E-07	0.730499	0.714	0.522	0.00297	0.192	ENSG00000138185	Upregulated
DEGS1	2.84E-09	0.727505	0.814	0.56	6.85E-05	0.254	ENSG00000143753	Upregulated
BSG	5.18E-12	0.708225	0.979	0.985	1.25E-07	-0.006	ENSG00000172270	Upregulated
ETNK1	6.25E-08	0.707039	0.907	0.94	0.001505	-0.033	ENSG00000139163	Upregulated
DLG1	1.00E-06	0.701573	0.836	0.754	0.024212	0.082	ENSG00000075711	Upregulated
MRPL33	1.47E-08	0.699351	0.914	0.933	0.000355	-0.019	ENSG00000243147	Upregulated
SLC7A8	1.80E-06	0.687463	0.5	0.276	0.043384	0.224	ENSG00000092068	Upregulated
KLHL42	5.11E-08	0.68138	0.814	0.627	0.001231	0.187	ENSG00000087448	Upregulated
TMEM9	4.12E-07	0.670195	0.929	0.948	0.009929	-0.019	ENSG00000116857	Upregulated
EIF1AX	3.13E-08	0.667052	0.921	0.933	0.000753	-0.012	ENSG00000173674	Upregulated
ROBO2	6.16E-09	0.666023	0.807	0.627	0.000148	0.18	ENSG00000185008	Upregulated
PRCP	1.22E-06	0.662187	0.857	0.813	0.029404	0.044	ENSG00000137509	Upregulated
SELENOT	1.06E-09	0.653764	0.907	0.881	2.55E-05	0.026	ENSG00000198843	Upregulated
PTGFRN	1.41E-07	0.644416	0.829	0.731	0.003391	0.098	ENSG00000134247	Upregulated
ATP5F1E	2.28E-09	0.634915	0.986	1	5.50E-05	-0.014	ENSG00000124172	Upregulated
HSPH1	1.55E-07	0.631229	0.893	0.91	0.003731	-0.017	ENSG00000120694	Upregulated
SNRPG	5.86E-09	0.624809	0.986	1	0.000141	-0.014	ENSG00000143977	Upregulated
NRP2	3.97E-10	0.624054	0.529	0.194	9.56E-06	0.335	ENSG00000118257	Upregulated
EPCAM	2.09E-10	0.607049	0.986	1	5.04E-06	-0.014	ENSG00000119888	Upregulated
PCYOX1	2.84E-08	0.600497	0.907	0.903	0.000685	0.004	ENSG00000116005	Upregulated
SERINC1	2.15E-07	0.600056	0.821	0.754	0.005185	0.067	ENSG00000111897	Upregulated
PDE4D	2.07E-06	0.594876	0.864	0.851	0.049939	0.013	ENSG00000113448	Upregulated
CALR	8.90E-07	0.59105	0.921	0.963	0.021449	-0.042	ENSG00000179218	Upregulated

BIRC6	2.75E-09	0.581963	0.9	0.866	6.62E-05	0.034	ENSG00000115760	Upregulated
XPO1	1.58E-06	0.570453	0.857	0.866	0.038102	-0.009	ENSG00000082898	Upregulated
CSTB	7.27E-07	0.564414	0.864	0.858	0.017504	0.006	ENSG00000160213	Upregulated
SLC35F5	9.82E-07	0.561794	0.85	0.799	0.02367	0.051	ENSG00000115084	Upregulated
HDDC2	1.39E-06	0.553189	0.864	0.754	0.033429	0.11	ENSG00000111906	Upregulated
FYTTD1	3.29E-07	0.550304	0.871	0.851	0.007924	0.02	ENSG00000122068	Upregulated
CD44	4.98E-07	0.544319	0.943	0.918	0.012009	0.025	ENSG00000026508	Upregulated
DNAJB1	1.80E-07	0.544009	0.893	0.888	0.004325	0.005	ENSG00000132002	Upregulated
HDAC9	3.86E-08	0.531605	0.686	0.41	0.00093	0.276	ENSG00000048052	Upregulated
RAB6A	1.24E-07	0.531235	0.886	0.925	0.002986	-0.039	ENSG00000175582	Upregulated
S100A6	5.38E-07	0.529534	1	1	0.012955	0	ENSG00000197956	Upregulated
PRSS56	1.94E-09	0.527679	0.421	0.097	4.67E-05	0.324	ENSG00000237412	Upregulated
LMNTD1	2.08E-07	0.519623	0.4	0.149	0.005001	0.251	ENSG00000152936	Upregulated
EXOC6	6.26E-07	0.505075	0.814	0.634	0.015094	0.18	ENSG00000138190	Upregulated
RPL14	1.28E-10	-0.50557	0.979	1	3.09E-06	-0.021	ENSG00000188846	Downregulated
GPR108	1.47E-06	-0.51105	0.6	0.754	0.035475	-0.154	ENSG00000125734	Downregulated
NPM1	2.38E-12	-0.51975	0.993	1	5.73E-08	-0.007	ENSG00000181163	Downregulated
GAPDH	4.87E-10	-0.52108	0.964	1	1.17E-05	-0.036	ENSG00000111640	Downregulated
SLC25A3	5.33E-08	-0.52365	0.9	0.978	0.001283	-0.078	ENSG00000075415	Downregulated
ACSL3	3.40E-08	-0.52845	0.843	0.933	0.000818	-0.09	ENSG00000123983	Downregulated
RPL7A	3.20E-14	-0.53934	0.993	1	7.72E-10	-0.007	ENSG00000148303	Downregulated
IGFBP2	4.20E-10	-0.54624	0.564	0.851	1.01E-05	-0.287	ENSG00000115457	Downregulated
HNRNPA1	1.43E-11	-0.55375	0.936	0.978	3.44E-07	-0.042	ENSG00000135486	Downregulated
PTMA	1.25E-11	-0.57921	0.993	1	3.02E-07	-0.007	ENSG00000187514	Downregulated
MGST1	3.31E-10	-0.581	0.736	0.918	7.97E-06	-0.182	ENSG00000008394	Downregulated
RPL18A	1.00E-11	-0.58221	0.964	1	2.42E-07	-0.036	ENSG00000105640	Downregulated
EIF3F	1.53E-08	-0.58357	0.771	0.918	0.00037	-0.147	ENSG00000175390	Downregulated
IL1R2	6.68E-07	-0.58896	0.271	0.537	0.016103	-0.266	ENSG00000115590	Downregulated
RPL7	2.63E-08	-0.59932	1	1	0.000634	0	ENSG00000147604	Downregulated
RPS3	5.60E-16	-0.61928	1	0.993	1.35E-11	0.007	ENSG00000149273	Downregulated
MTCO1P12	2.61E-08	-0.62391	0.807	0.821	0.000629	-0.014	ENSG00000237973	Downregulated

RPL22L1	1.60E-09	-0.62627	0.871	0.94	3.86E-05	-0.069	ENSG00000163584	Downregulated
RPS3A	1.31E-16	-0.62855	0.986	1	3.14E-12	-0.014	ENSG00000145425	Downregulated
CD46	9.60E-09	-0.63549	0.964	0.993	0.000231	-0.029	ENSG00000117335	Downregulated
SH3BP4	4.23E-07	-0.63616	0.707	0.821	0.010185	-0.114	ENSG00000130147	Downregulated
RPL13A	5.96E-15	-0.63889	0.993	1	1.44E-10	-0.007	ENSG00000142541	Downregulated
RPS9	2.02E-13	-0.6443	0.971	0.978	4.87E-09	-0.007	ENSG00000170889	Downregulated
RPL8	4.70E-16	-0.64614	0.993	1	1.13E-11	-0.007	ENSG00000161016	Downregulated
RPS6	1.27E-18	-0.65793	1	1	3.06E-14	0	ENSG00000137154	Downregulated
RPS14	1.09E-16	-0.66142	1	1	2.63E-12	0	ENSG00000164587	Downregulated
CALCA	3.20E-07	-0.66291	0.129	0.396	0.00771	-0.267	ENSG00000110680	Downregulated
TESC	6.09E-08	-0.67436	0.821	0.925	0.001468	-0.104	ENSG00000088992	Downregulated
RPL13	2.09E-14	-0.6865	0.964	1	5.04E-10	-0.036	ENSG00000167526	Downregulated
PLEKHB1	3.03E-09	-0.73659	0.614	0.799	7.31E-05	-0.185	ENSG00000021300	Downregulated
EEF1A1	1.35E-22	-0.7544	1	1	3.26E-18	0	ENSG00000156508	Downregulated
SLC25A6	7.54E-15	-0.79627	0.921	0.97	1.82E-10	-0.049	ENSG00000169100	Downregulated
HES6	5.65E-09	-0.83116	0.814	0.91	0.000136	-0.096	ENSG00000144485	Downregulated
RACK1	4.22E-22	-0.8878	0.986	1	1.02E-17	-0.014	ENSG00000204628	Downregulated
RPL3	1.17E-23	-1.09246	0.979	0.993	2.82E-19	-0.014	ENSG00000100316	Downregulated
MT-RNR2	1.61E-14	-1.25719	1	1	3.88E-10	0	ENSG00000210082	Downregulated

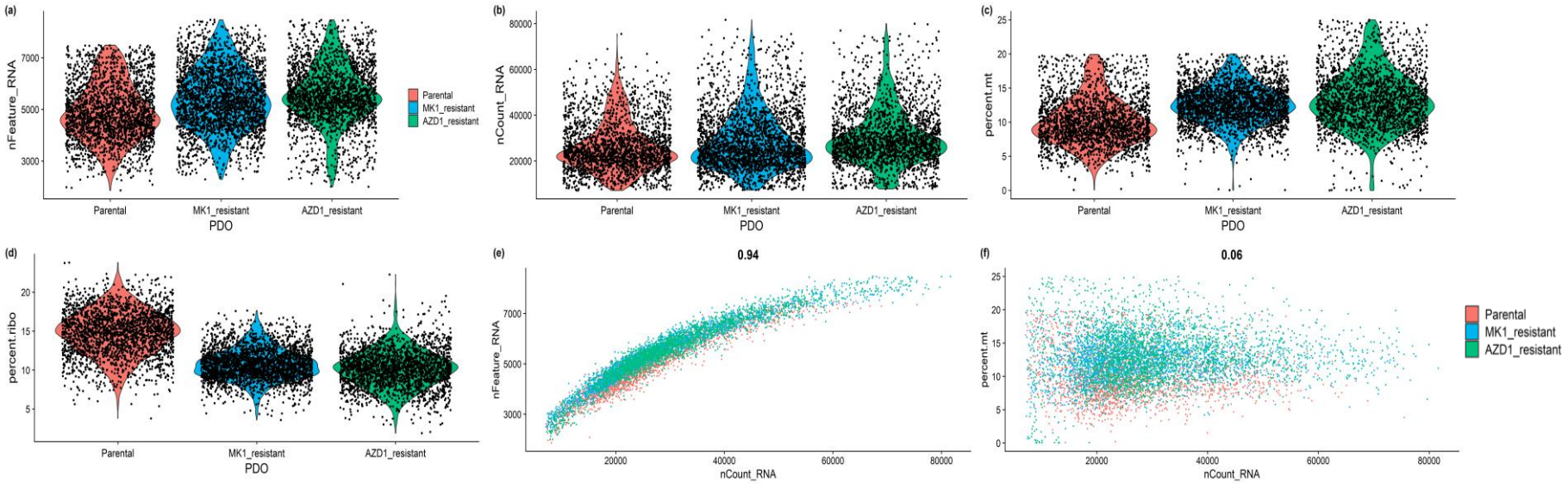
A.4 10x Genomics scRNA-seq analysis of colorectal cancer patient-derived organoids

A.4.1 10x scRNA-seq data quality control, batch correction and clustering analysis



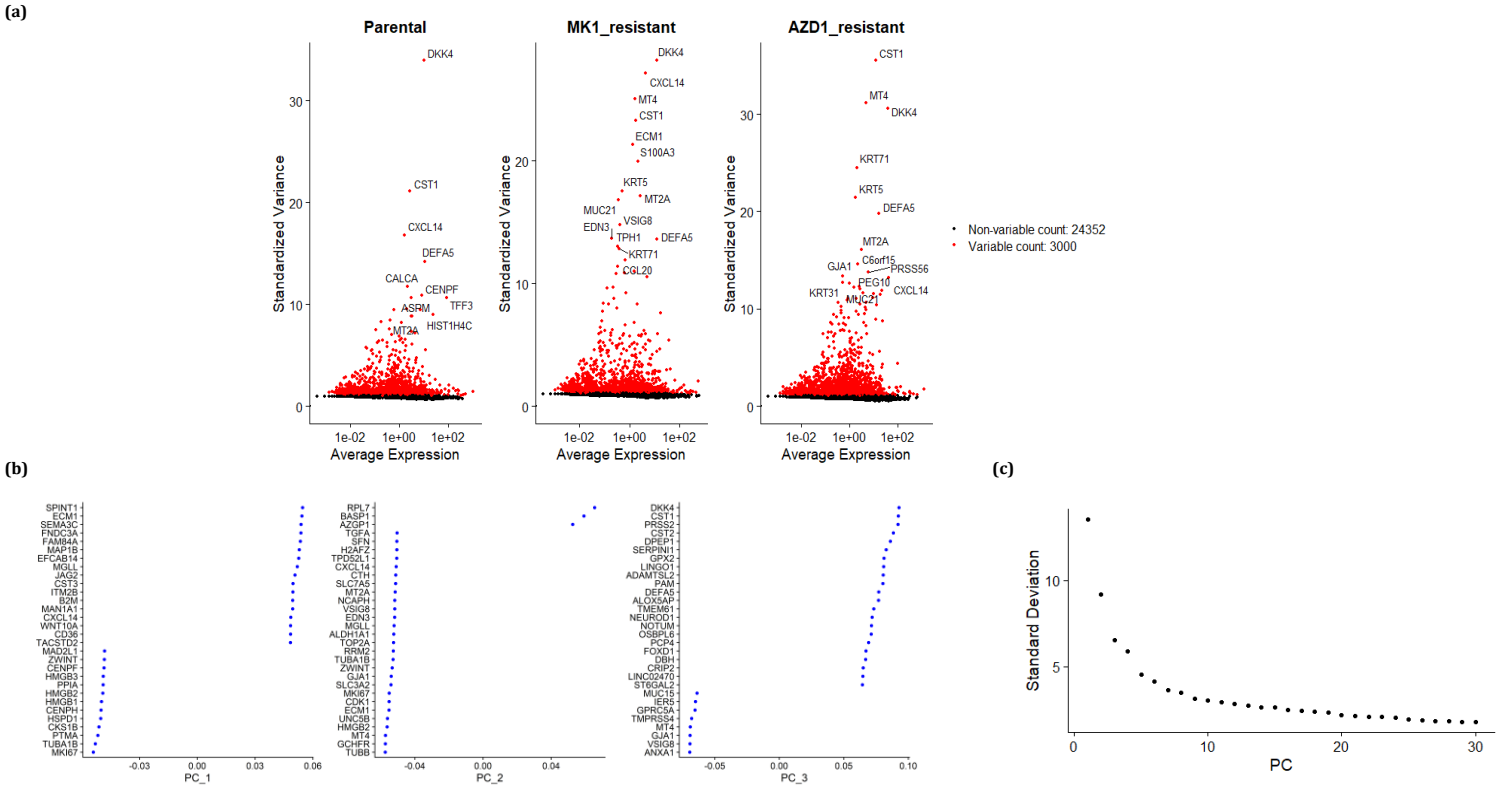
Supplementary Figure 4. Quality assessment of 8,577 10x Genomics scRNA-seq libraries from three mCRC PDOs before excluding low-quality cells.

Violin plots illustrating the distribution of single cells derived from the Parental, MK1-resistant and AZD1-resistant PDOs based on various Seurat quality control metrics before filtering. These metrics include: ***(a)*** *number of genes detected (nFeature_RNA),* ***(b)*** *number of RNA molecules (nCount_RNA),* ***(c)*** *mitochondrial RNA percentage (percent.mt), and* ***(d)*** *ribosomal RNA percentage (percent.ribo) for each cell. All metrics were evaluated across each PDO and over two sequencing runs. Scatter plots depict the relationship between RNA content and* ***(e)*** *number of genes detected or* ***(f)*** *mitochondrial content for each cell. 8,577 cells in total: 2,726 Parental, 3,220 MK1-resistant, and 2,631 AZD1-resistant cells. In this multi-panel figure, colour-coding distinguishes the sequencing runs, with coral representing the first and sky-blue denoting the second batch of sequenced cells.*



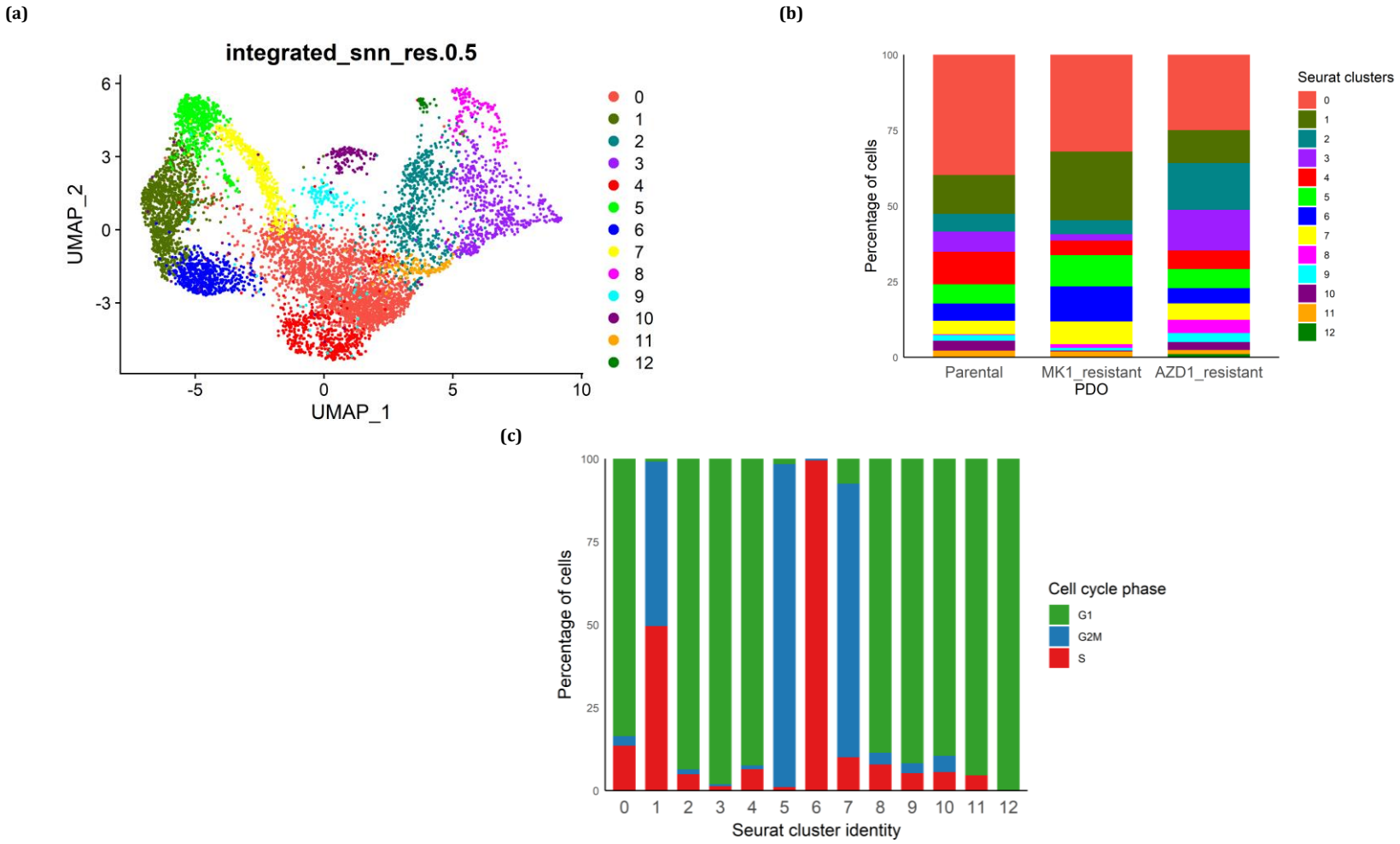
Supplementary Figure 5. Quality metrics for 7,242 scRNA-seq libraries from three mCRC PDOs after Seurat filtering.

Violin showing: ***(a)*** *number of genes detected (nFeature_RNA),* ***(b)*** *number of RNA molecules (nCount_RNA),* ***(c)*** *mitochondrial read percentage (percent.mt), and* ***(d)*** *ribosomal read percentage (percent.ribo) per single cell after the removal of low-quality cells. All metrics were evaluated across each PDO and over two sequencing runs. Scatter plots depict the correlation between RNA content and* ***(e)*** *number of genes detected or* ***(f)*** *mitochondrial content for each cell. 7,242 cells in total: 2,245 Parental, 2,708 MK1-resistant, and 2,289 AZD1-resistant cells. In this multi-panel figure, colour-coding distinguishes the sequencing runs, with coral representing the first and sky-blue denoting the second batch of sequenced cell.*



Supplementary Figure 6. Seurat integration aligns three scRNA-seq datasets using variably expressed anchor genes, followed by principal component analysis to identify the main axes of variation for further analyses.

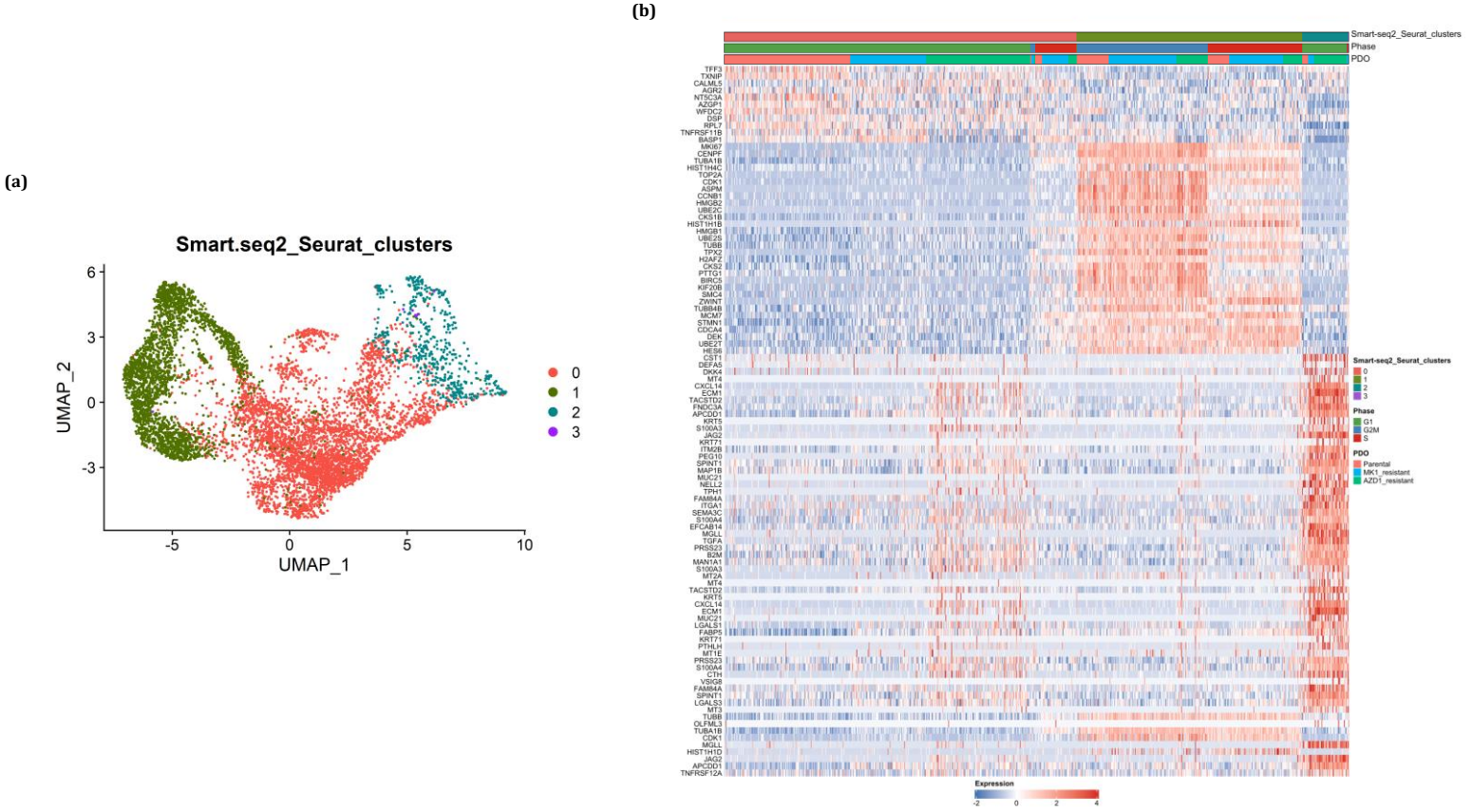
(a) *Scatter plots reflect the standard variance of genes against average expression, highlighting the top 3,000 genes (red) with the most significant variability in gene expression in the first (left) and second (right) sequencing runs. Top 15 genes in each batch are annotated.* ***(b)*** *Representation of genes in the first three principal component space in the integrated dataset.* ***(c)*** *Scree plot shows the variation captured by the first 30 principal components. This visualisation aids in selecting the optimal number of principal components for cell clustering, with PCs 1-25 chosen based on the plot's "elbow", where the line begins to flatten.*



Supplementary Figure 7. Non-linear dimensionality reduction identifies 12 transcriptionally distinct cell clusters in mCRC PDOs at a resolution of 0.5.

(a) *In the UMAP plot cells are grouped by gene expression similarity resulting in 12 clusters at a resolution of 0.5.* ***(b)*** *Distribution of clusters across mCRC organoids.* ***(c)*** *Percentage of cells in various cell cycle phases across identified clusters.*

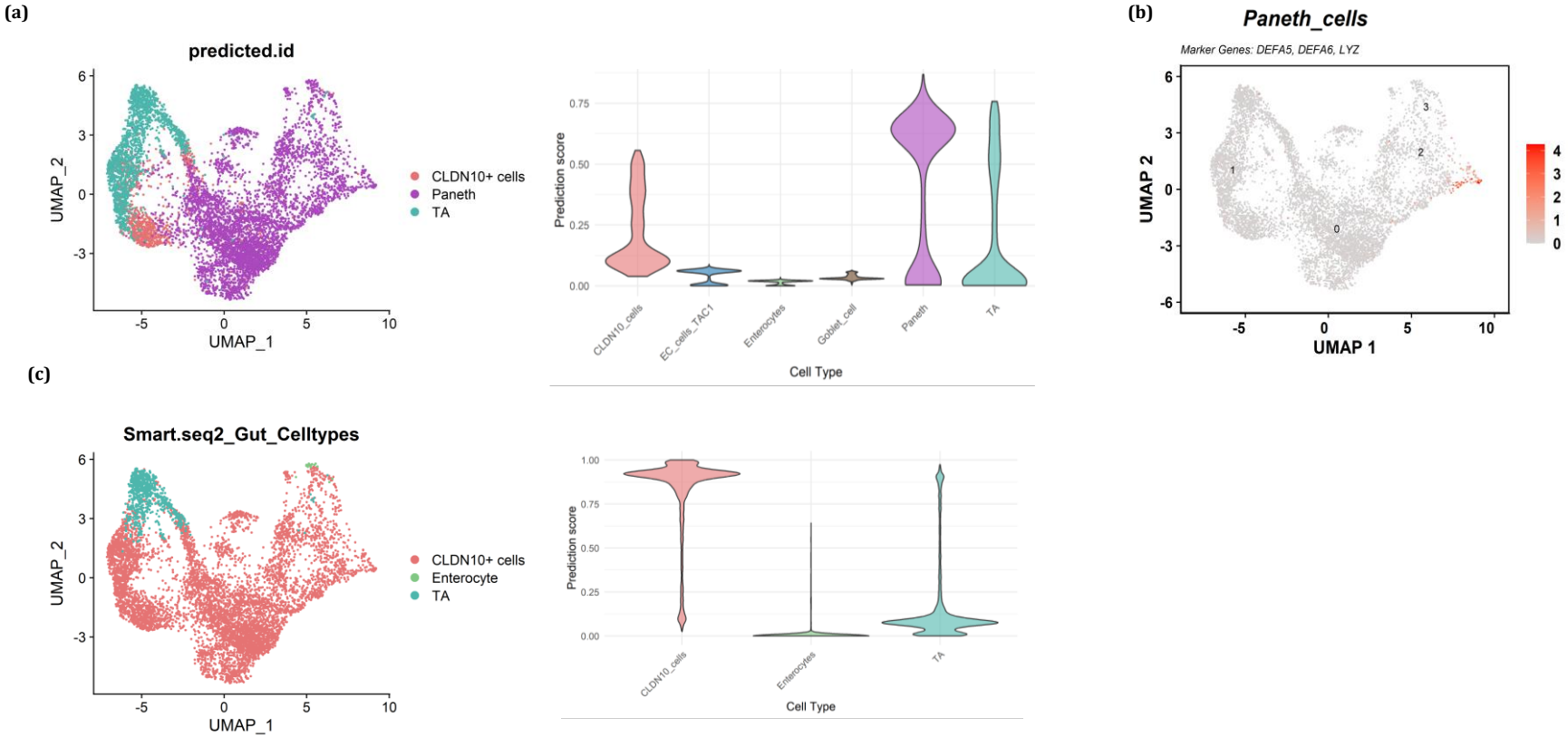
A.4.2 Projection of Smart-seq2 clusters onto the 10x scRNA-seq dataset



Supplementary Figure 8. Gene expression analysis of clusters identified in mCRC PDOs.

(a) *UMAP plots displays clusters identified when Smart-seq2 clusters are projected onto the 10x scRNA-seq dataset.* ***(b)*** *Heatmap of key differentially expressed genes for each cluster, with gene expression levels transitioning from blue (low expression) to red (high expression). Annotations above the heatmap indicate cell cycle stage and PDO source.*

A.4.3 Projection of Smart-seq2 cell type labels onto the 10x scRNA-seq dataset

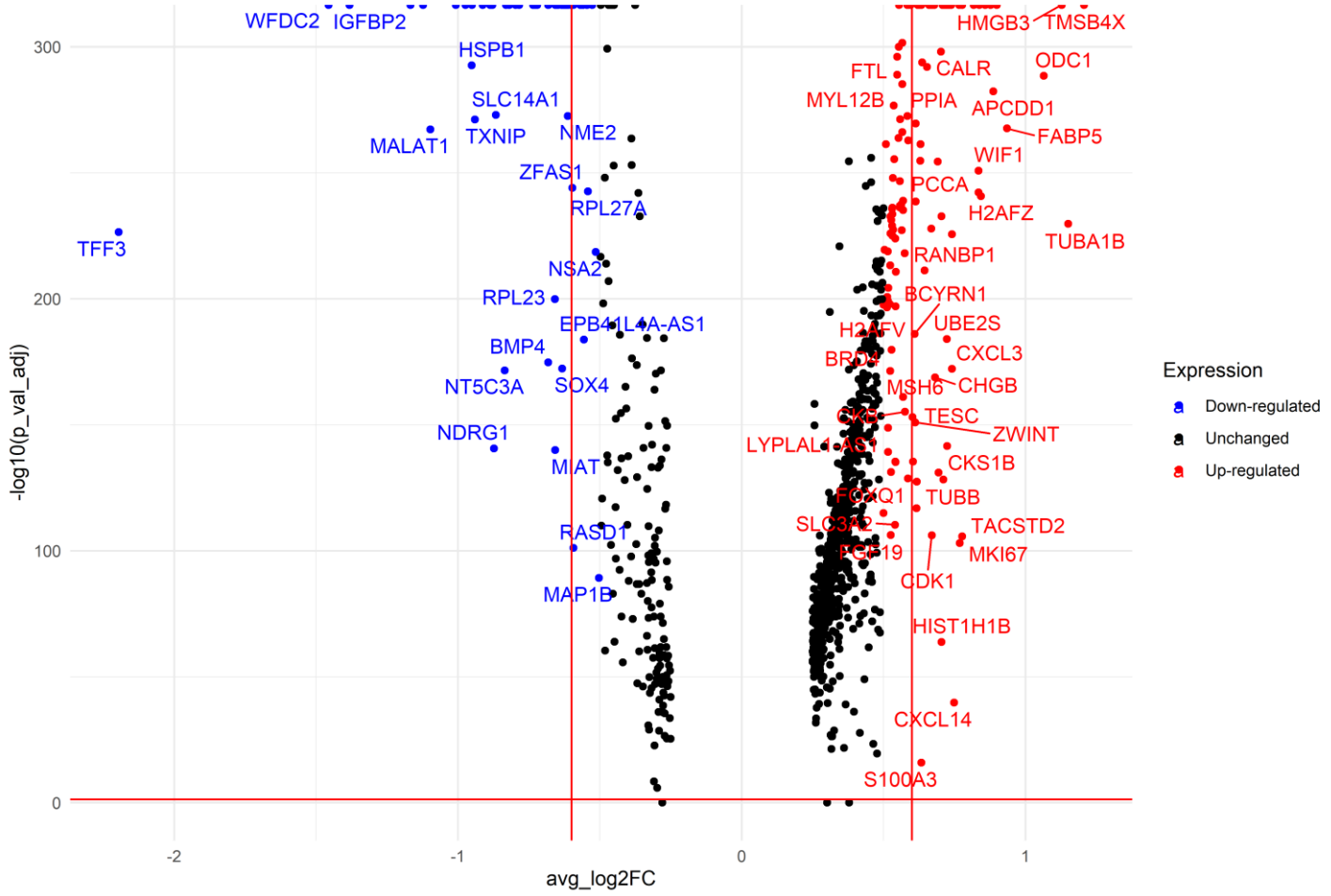


Supplementary Figure 9. Prediction scores for colonic cell types identified in mCRC PDOs using two annotation references

(a) *UMAP plot of mCRC PDOs from 10x scRNA-seq data, labelled with the Human Gut Cell Atlas (left), and distribution of prediction scores across annotated cell types (right).* ***(b)*** *Expression patterns of selected marker genes used to identify Paneth cells.* ***(b)*** *UMAP plot of mCRC PDOs from 10x scRNA-seq data, labelled with cell types identified in a reference Smart-seq2 dataset previously annotated using the Human Gut Cell Atlas (left), and distribution of prediction scores across annotated cell types (right).*

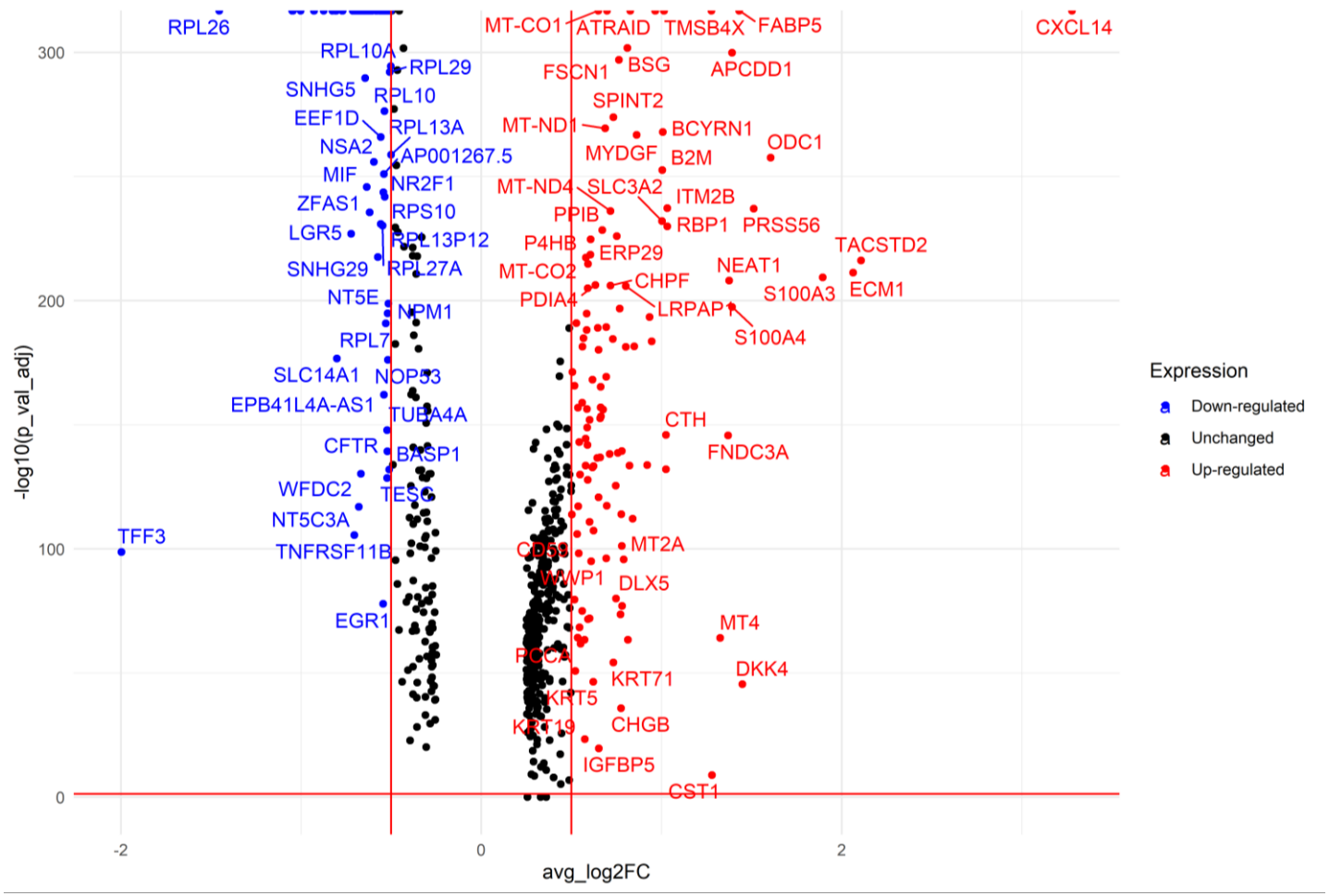
A.4.4 10x scRNA-seq differential gene expression analysis across mCRC PDOs

I. Pairwise DGE analysis between MK1-resistant and Parental PDOs



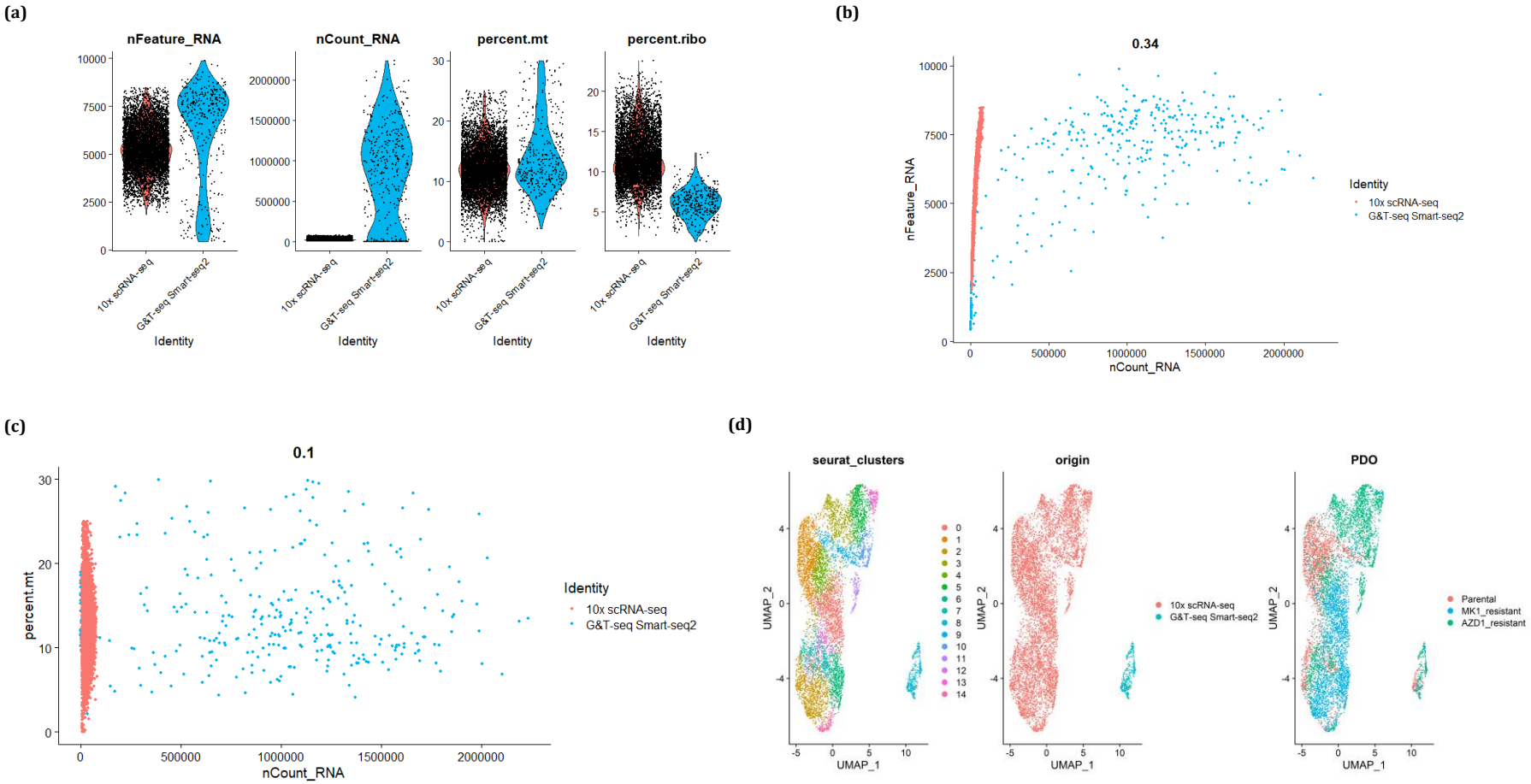
Supplementary Figure 10. Differential expression analysis between MK1-resistant and Parental PDOs.

II. Pairwise DGE analysis between AZD1-resistant and Parental PDOs



Supplementary Figure 11. Differential expression analysis between AZD1-resistant and Parental PDOs.

A.5 Comparison between Smart-seq2 and 10x scRNA-seq datasets

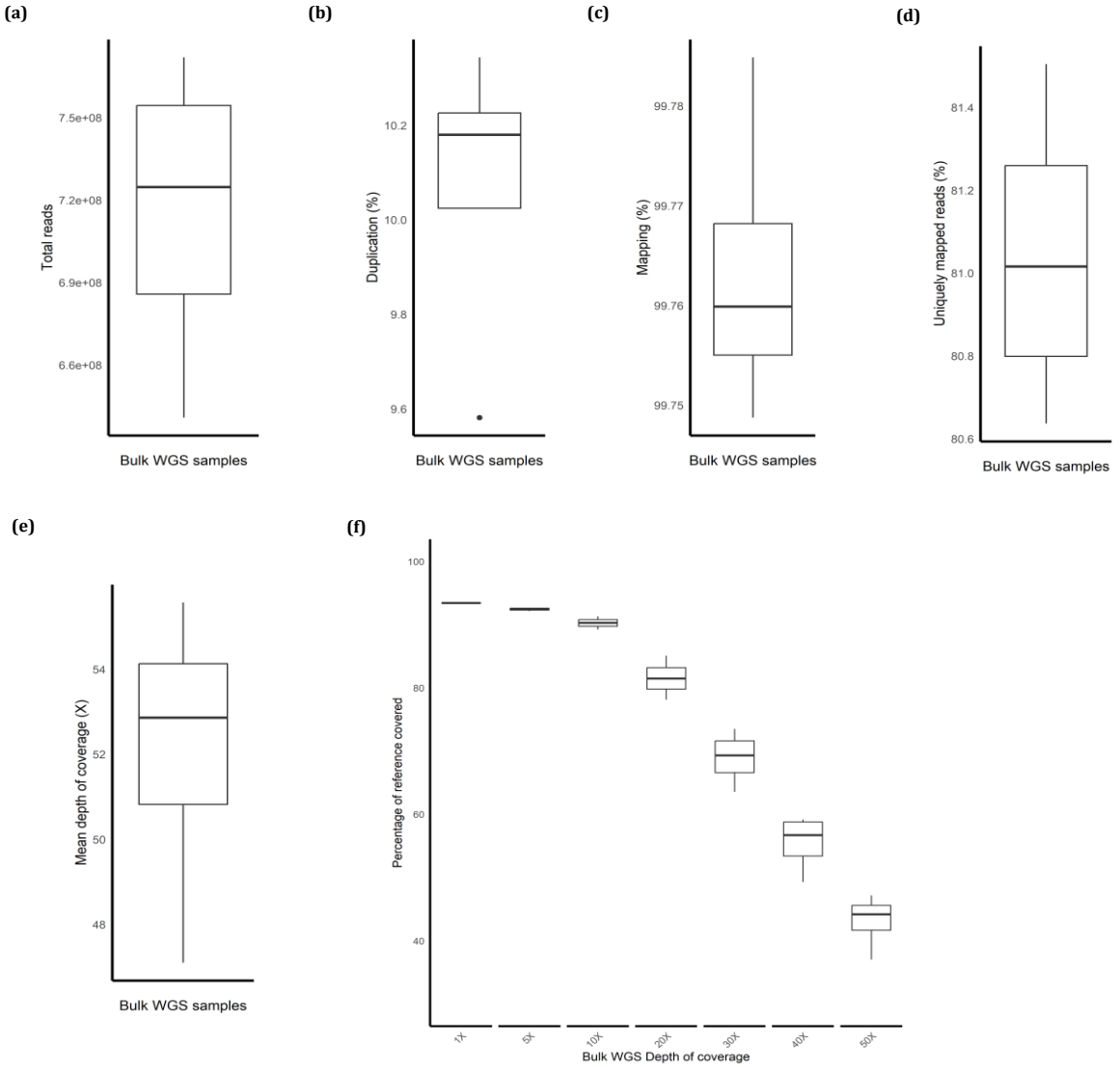


Supplementary Figure 12. Technical comparison between Smart-seq2 and 10x scRNA-seq datasets from mCRC PDOs.

In each plot, data points represent cells derived from mCRC PDOs: n= 361 cells for Smart-seq2; n = 7,242 cells for 10x Genomics.

Appendix B. Supplementary material for Chapter 4

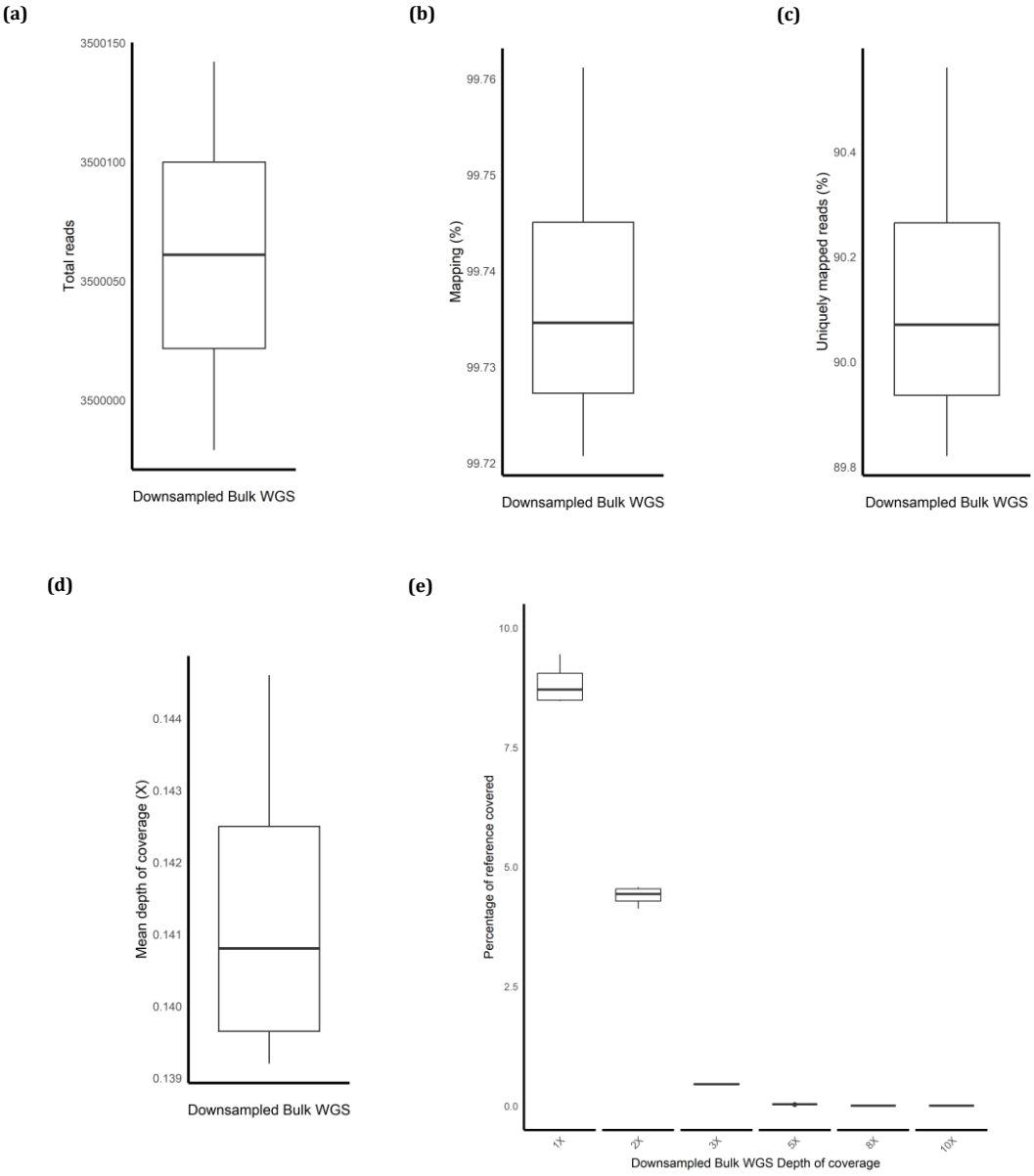
B.1 Processing of bulk WGS data from mCRC organoids and blood control



Supplementary Figure 13. Distribution of bulk whole-genome sequencing data derived from mCRC tumoroids and matching blood control, across various quality control metrics.

Quality control metrics of bulk WGS data derived from Parental, MK1-resistant, AZD1-resistant mCRC organoids, along with a matched blood sample from the patient from whom the organoids were derived. Metrics evaluated include: ***(a)*** *total number of reads,* ***(b)*** *percentage of duplicated reads,* ***(c)*** *overall mapping rate,* ***(d)*** *percentage of uniquely mapping reads,* ***(e)*** *mean depth of coverage, and (f) breadth of coverage at various coverage depths. Each box plot displays the median value (central line), the 25th and 75th percentiles (box boundaries), and the 1.5× interquartile range (whiskers). Outliers are shown as individual points beyond the whiskers.*

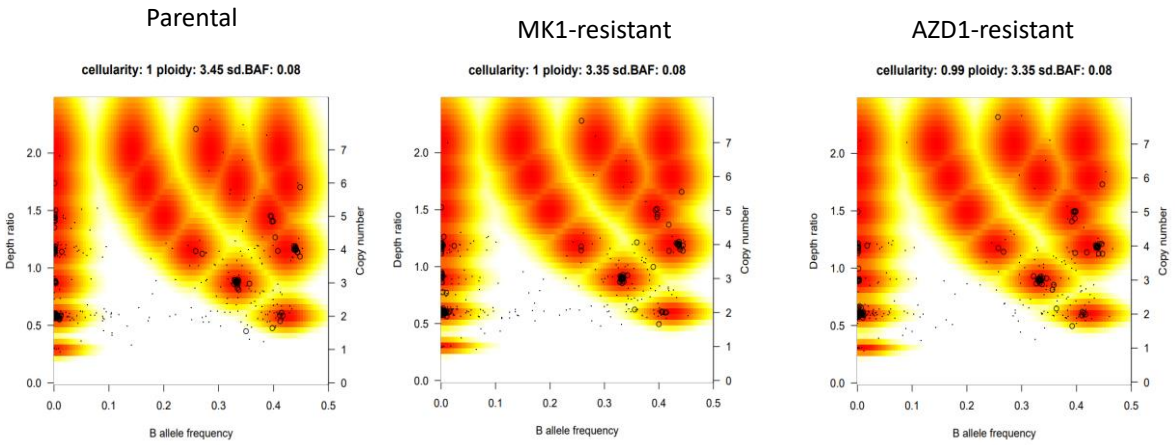
B.1.1 Quality control metrics for the downsampled bulk WGS dataset



Supplementary Figure 14. Distribution of downsampled bulk whole-genome sequencing data derived from mCRC tumoroids and matching blood control, across various quality control metrics.

Quality control metrics for downsampled bulk WGS data derived from Parental, MK1-resistant, AZD1-resistant mCRC organoids, along with a matched blood sample from the patient from whom the organoids were derived. Metrics evaluated include: ***(a)*** *total number of reads,* ***(b)*** *percentage of duplicated reads,* ***(c)*** *overall mapping rate,* ***(d)*** *percentage of uniquely mapping reads,* ***(e)*** *mean depth of coverage, and (f) breadth of coverage at various coverage depths. Each box plot displays the median value (central line), the 25th and 75th percentiles (box boundaries), and the 1.5× interquartile range (whiskers).*

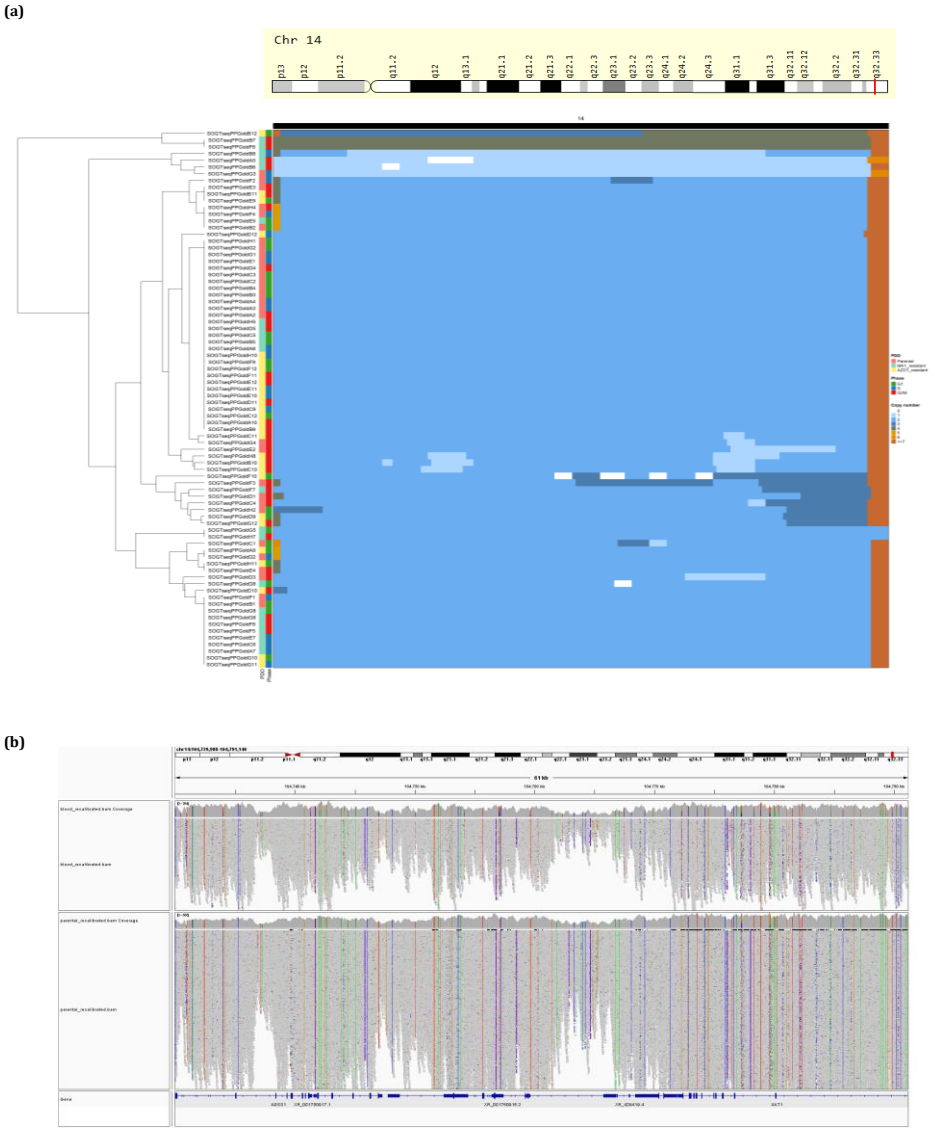
B.2 Sequenza CNA analysis of bulk WGS data



Supplementary Figure 15. Sequenza cellularity and ploidy estimates for bulk Parental, MK1- and AZD1-resistant PDOs.

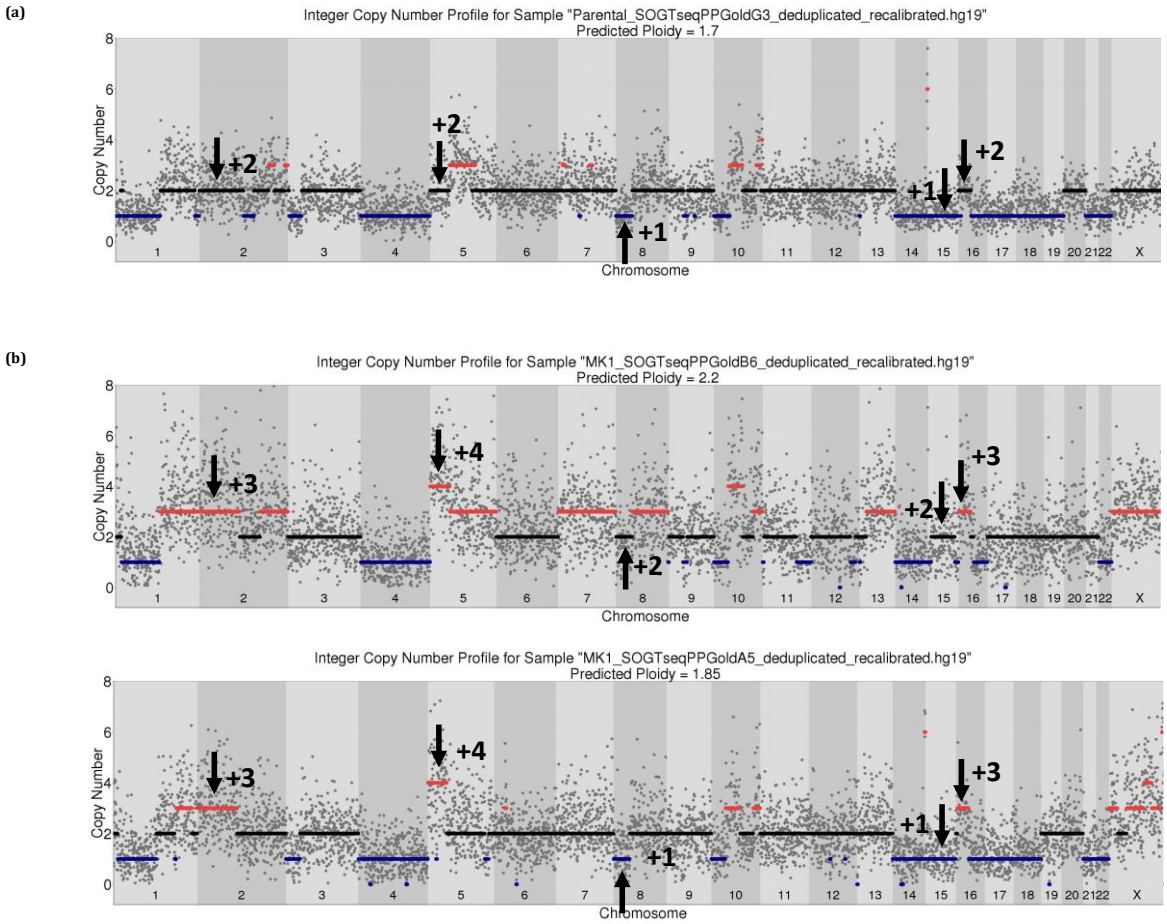
The x-axis shows the B allele frequency (BAF), which is the proportion of sequencing reads that contain the variant (non-reference) allele at heterozygous loci within the genomic segments analysed (black circles and dots). The y-axis represents the observed to expected read count ratio. This depth ratio helps infer the copy number (right-side scale) at each genome segment. The colour gradient illustrates the log posterior probability density of observing specific depth ratio, B allele frequency and ploidy combinations across genomic segments.

B.3 Subclonal characterisation of mCRC PDOs



Supplementary Figure 16. Copy number profile of mCRC tumoroids at chromosome 14 highlighting the *AKT1* locus

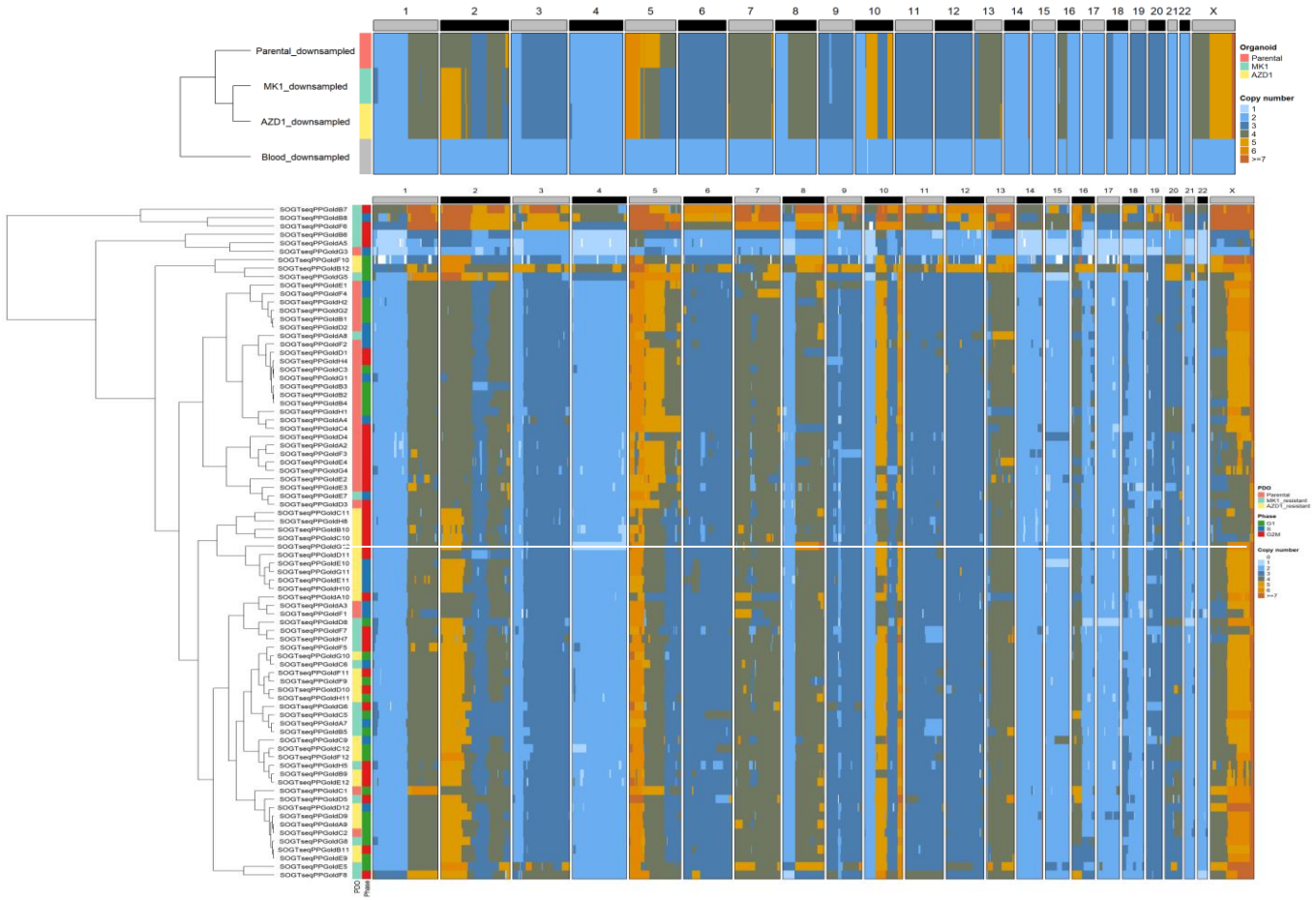
(a) *Heatmap shows copy number profiles of mCRC cells across chromosome 14. The cytoband region displayed above the heatmap, extracted from genecards.org, highlights the AKT1 locus at 14q32.33. Rows correspond to 80 single-cell genomes from Parental (n=30), MK1- (n=22), and AZD1-resistant (n=28) organoids.* ***(b)*** *IGV snapshot focusing (bulk data) on the AKT1 locus, comparing read depth between the matched blood control (top track) and Parental organoid (bottom track). The amplification of AKT1 in the Parental sample is evident as an increased number of reads compared to the blood control*



Supplementary Figure 17. Copy number profiles of single-cell genomes derived from a common ancestor.

Plots display genome-wide copy number profiles of a single-cell genome derived from ***(a)*** *the Parental organoid, along with* ***(b)*** *two MK1-resistant cells. These resistant cells likely originated and subsequently diverged from the Parental line based on shared similarities in their copy number profiles. Arrows indicate regions of CNA gains or loses where the MK1-reistant cells diverge with respect to the Parental cell. Coloured horizontal lines represent median copy number states: black for diploid segments, red for amplifications, and blue for deletions.*

B.4 Comparison between bulk and single-cell WGS

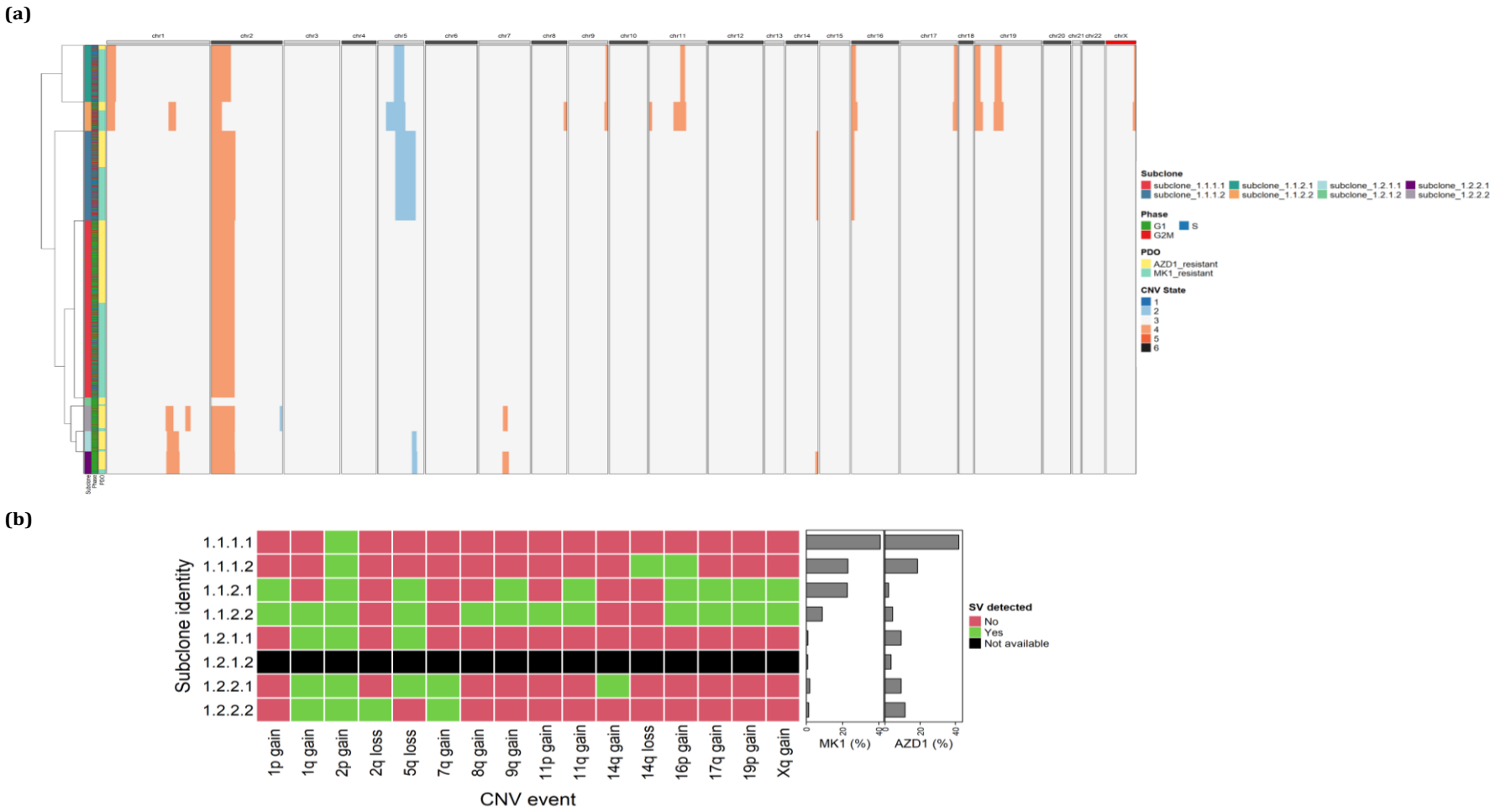


Supplementary Figure 18. Comparison of copy number profiles of mCRC PDOs by bulk and single-cell WGS.

*Heatmaps depicting **(top)** genome-wide copy number profiles for downsampled bulk WGS data and **(bottom)** 80 single-cell genomes from Parental (n=30), MK1-resistant (n=22), and AZD1-resistant (n=28) organoids (both analyses performed using 500 kb genomic bins). Hierarchical clustering of the genomes was performed using Euclidean distance and Ward's linkage method. The colour scale indicates integer copy number state.*

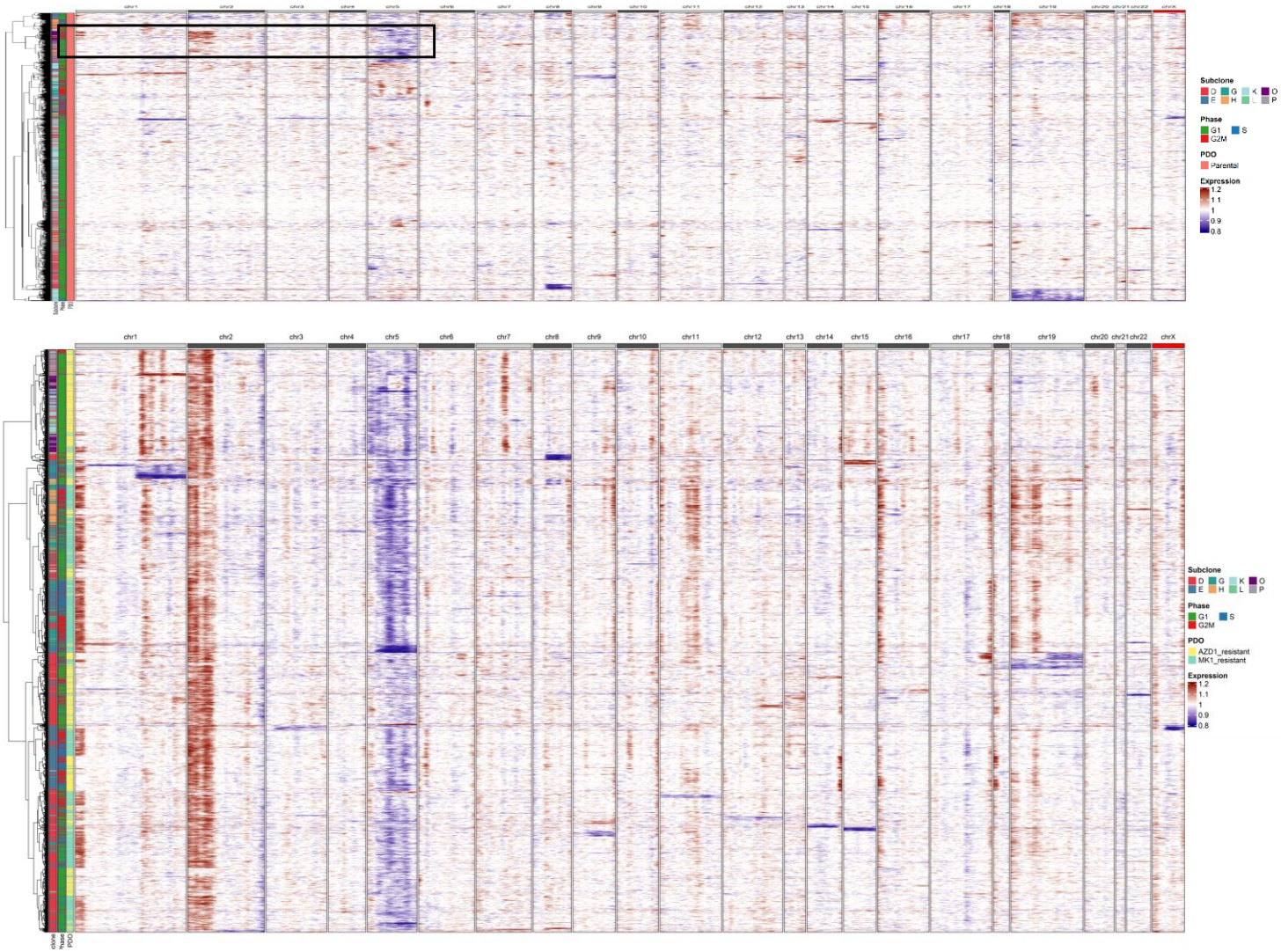
Appendix C. Supplementary material for Chapter 5

C.1 Transcriptome-based DNA copy number inference from 10x scRNA-seq data



Supplementary Figure 19. Genome-wide copy number states of AKTi-resistant cells inferred from 10x scRNA-seq data.

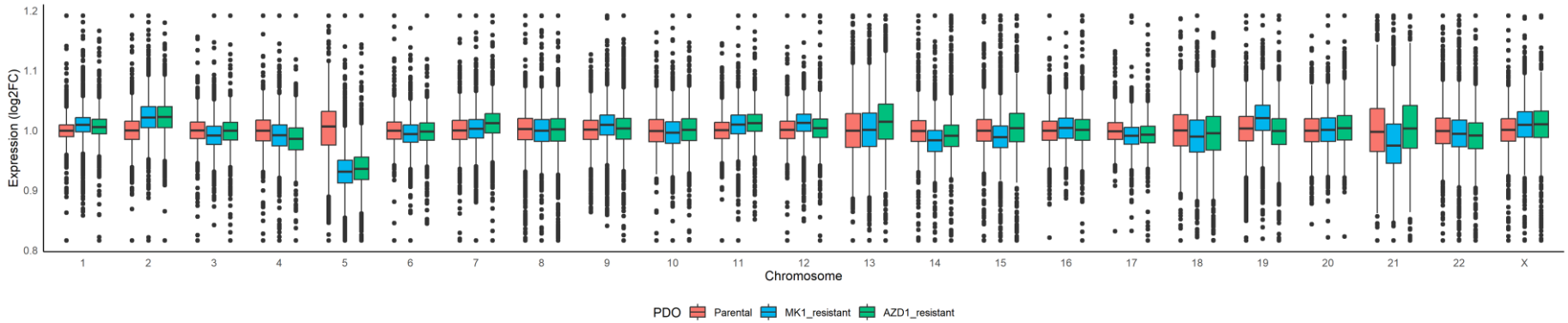
(a) *Heatmap displays predicted CNV regions as identified by inferCNV using a six-state Hidden Markov Model (HMM). Heatmap colours correspond to one of the six HMM states, indicating varying degrees of copy number alterations: State 1 represents the loss of two copies; State 2 indicates the loss of one copy; State 3 denotes a neutral state with no change; State 4 represents the addition of one copy; State 5 represents the addition of two copies; and State 6 denotes the addition of more than two copies. Columns correspond to genomic windows covering 100 adjacent genes, providing an averaged view of CNV states. Rows represent 4,997 single-cell transcriptomes derived from MK1-resistant (n=2,708) and AZD1-resistant (n=2,289) mCRC organoids.* ***(b)*** *CNV events characterising inferCNV subclones, with adjacent bar plots denoting their frequency (%) in AKTi-resistant PDOs.* ***(c)*** *Frequency of inferCNV subclones in the Parental organoid.*



Supplementary Figure 20. Heatmap of relative expression values of genes across mCRC single-cell transcriptomes from 10x scRNA-seq.

Top heatmap displays the expression values of Parental cells (n=2,245) along the rows, clustered based on Euclidean distance and Ward's linkage.

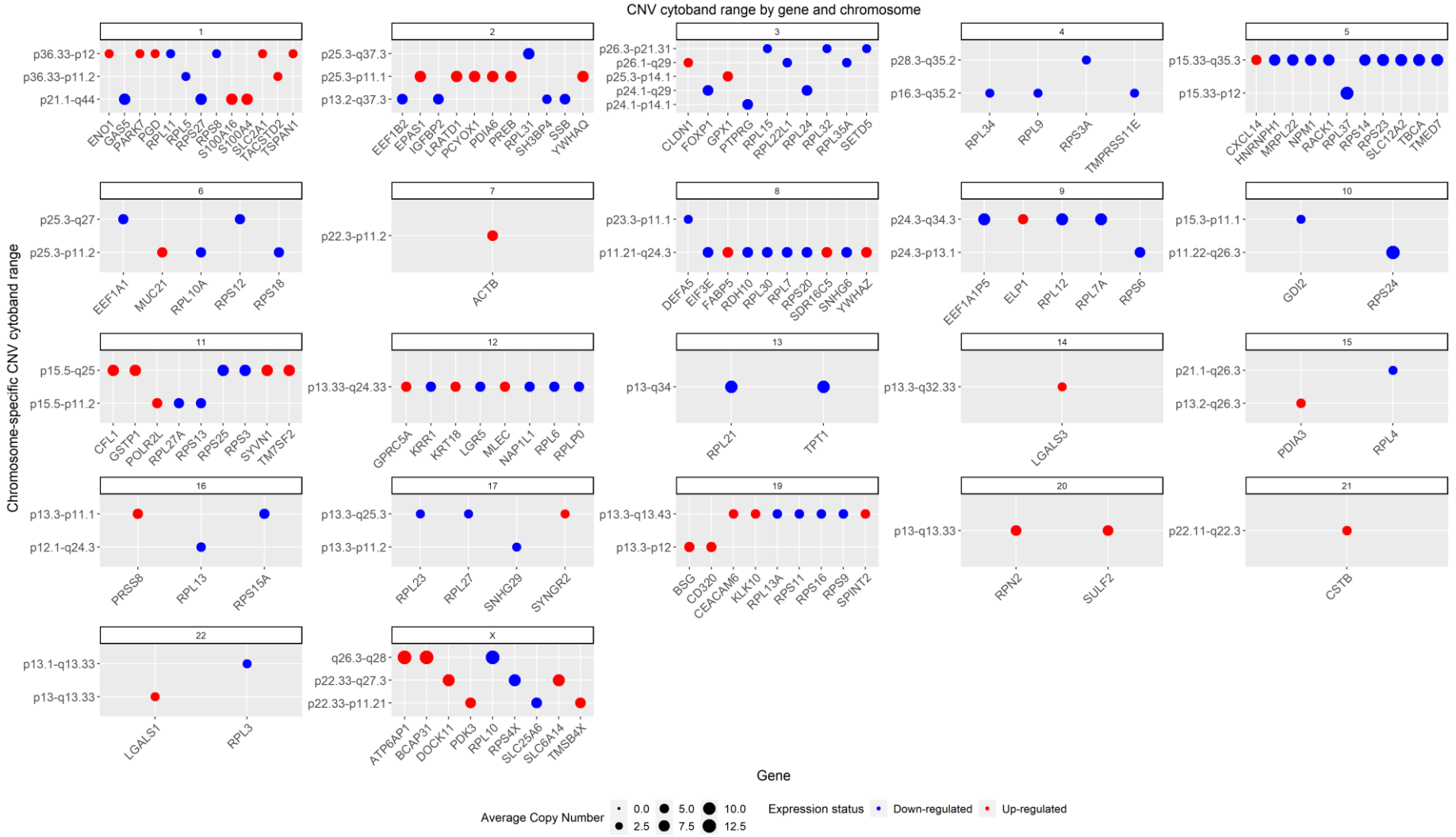
Columns represent genes ordered by their genomic position across chromosomes. Bottom heatmap shows the residual expression values for MK1- and AZD1-resistant cells (n=2,708, n=2,289 cells, respectively). Colour intensities indicate chromosomal regions with significantly higher or lower expression. Red indicates regions likely containing large, amplified segments, while blue denotes regions with potential deletions. Rows (cells) are organised using hierarchical clustering based on Euclidean distance and average linkage. The black box on the top heatmap highlights Parental cells with amplifications at chromosomes 1 and 2, and a deletion at chromosome 5. This suggests that these cells were already present in the untreated control and subsequently expanded in the AKTi-resistant organoids.



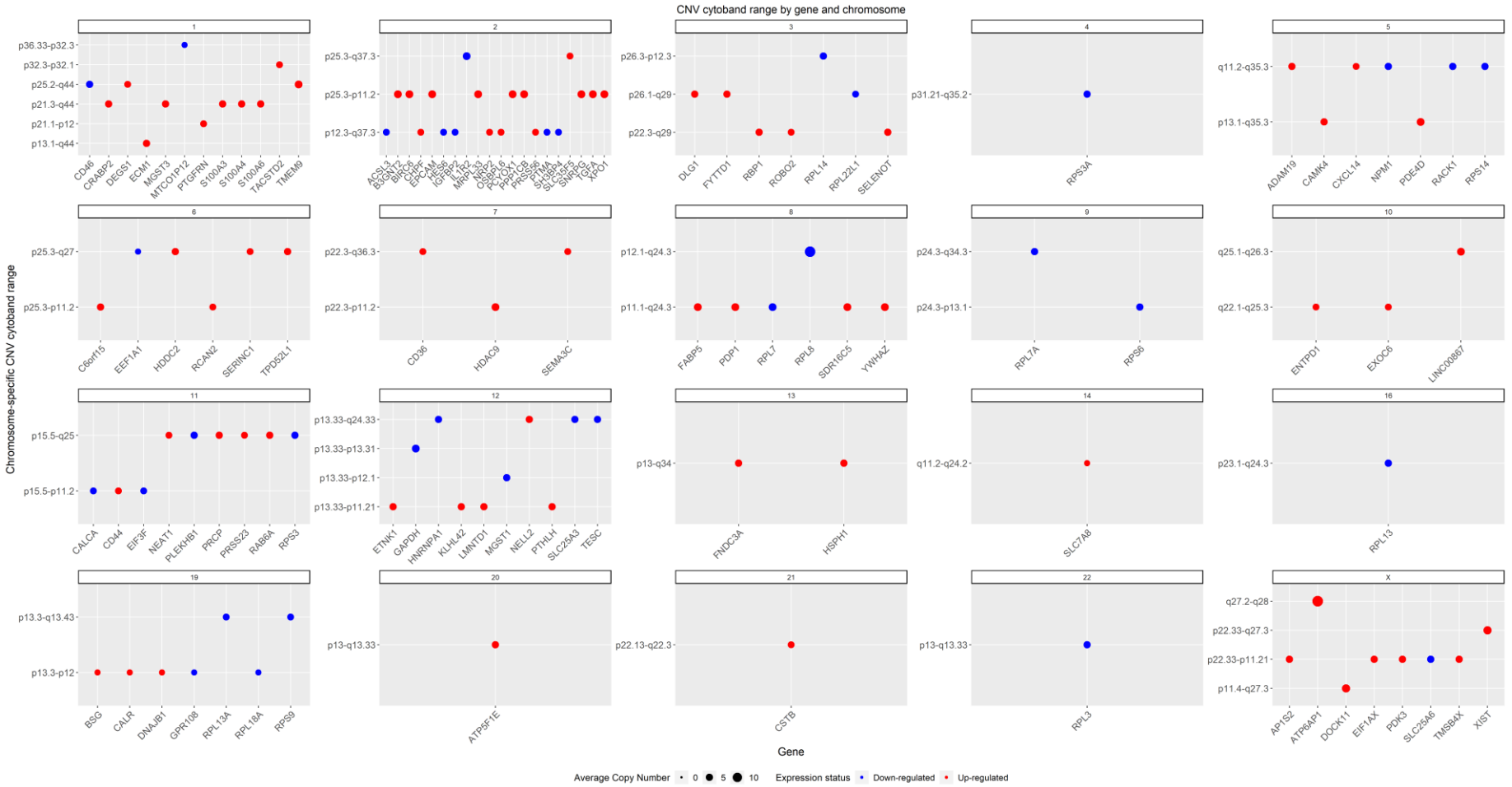
Supplementary Figure 21. Genome-wide gene expression binned per chromosome in mCRC single-cell transcriptomes derived from 10x high-throughput scRNA-seq.

Box plots represent the median expression values across each chromosome for the Parental (n=2,245), MK1-resistant (n=2,708), and AZD1-resistant (2,289) mCRC PDOs. Expression data was obtained by subtracting the average log2 fold change values in Parental cells from those in AKTi-resistant cells.

C.2. Relationship between copy number alterations and differentially expressed genes in AKTi-resistant organoids



Supplementary Figure 22. Genome-wide landscape of copy number alterations affecting differentially expressed genes in MK1-resistant cells.



Supplementary Figure 23 Genome-wide landscape of copy number alterations affecting differentially expressed genes in AZD1-resistant cells.

Bibliography

1. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. CA Cancer J Clin. 2023;73(3):233-54.

2. Cancer Research UK. Bowel cancer statistics: Cancer Research UK; 2023 [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero.

3. Aarons CB, Shanmugan S, Bleier JI. Management of malignant colon polyps: current status and controversies. World J Gastroenterol. 2014;20(43):16178-83.

4. Shussman N, Wexner SD. Colorectal polyps and polyposis syndromes. Gastroenterol Rep (Oxf). 2014;2(1):1-15.

5. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. J Mol Diagn. 2008;10(1):13-27.

6. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-7.

7. Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite instability: a review of what the oncologist should know. Cancer Cell International. 2020;20(1):16.

8. Kawakami H, Zaanan A, Sinicrope FA. Microsatellite instability testing and its role in the management of colorectal cancer. Curr Treat Options Oncol. 2015;16(7):30.

9. Nguyen HT, Duong HQ. The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy. Oncol Lett. 2018;16(1):9-18.

10. Cao Y. Tumorigenesis as a process of gradual loss of original cell identity and gain of properties of neural precursor/progenitor cells. Cell Biosci. 2017;7:61.

11. Williams MJ, Sottoriva A, Graham TA. Measuring Clonal Evolution in Cancer with Genomics. Annu Rev Genomics Hum Genet. 2019;20:309-29.

12. Menter DG, Davis JS, Broom BM, Overman MJ, Morris J, Kopetz S. Back to the Colorectal Cancer Consensus Molecular Subtype Future. Current Gastroenterology Reports. 2019;21(2):5.

13. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.

14. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022;12(1):31-46.

15. Malki A, ElRuz RA, Gupta I, Allouch A, Vranic S, Al Moustafa AE. Molecular Mechanisms of Colon Cancer Progression and Metastasis: Recent Insights and Advancements. Int J Mol Sci. 2020;22(1).

16. Jelsig AM, Qvist N, Brusgaard K, Nielsen CB, Hansen TP, Ousager LB. Hamartomatous polyposis syndromes: a review. Orphanet J Rare Dis. 2014;9:101.

17. Fearon ER. Molecular genetics of colorectal cancer. Annu Rev Pathol. 2011;6:479-507.

18. Meyer LA, Broaddus RR, Lu KH. Endometrial Cancer and Lynch Syndrome: Clinical and Pathologic Considerations. Cancer Control. 2009;16(1):14-22.

19. Bennedsen ALB, Furbo S, Bjarnsholt T, Raskov H, Gogenur I, Kvich L. The gut microbiota can orchestrate the signaling pathways in colorectal cancer. APMIS. 2022;130(3):121-39.

20. Basu S, Haase G, Ben-Ze'ev A. Wnt signaling in cancer stem cells and colon cancer metastasis. F1000Res. 2016;5.

21. Marshman E, Booth C, Potten CS. The intestinal epithelial stem cell. Bioessays. 2002;24(1):91-8.

22. Mah AT, Yan KS, Kuo CJ. Wnt pathway regulation of intestinal stem cells. J Physiol. 2016;594(17):4837-47.

23. Du H, Nie Q, Holmes WR. The Interplay between Wnt Mediated Expansion and Negative Regulation of Growth Promotes Robust Intestinal Crypt Structure and Homeostasis. PLoS Comput Biol. 2015;11(8):e1004285.

24. Kok SY, Nakayama M, Morita A, Oshima H, Oshima M. Genetic and nongenetic mechanisms for colorectal cancer evolution. Cancer Sci. 2023;114(9):3478-86.

25. Teeuwssen M, Fodde R. Cell Heterogeneity and Phenotypic Plasticity in Metastasis Formation: The Case of Colon Cancer. Cancers [Internet]. 2019; 11(9).

26. Wu D, Pan W. GSK3: a multifaceted kinase in Wnt signaling. Trends Biochem Sci. 2010;35(3):161-8.

27. Flack JE, Mieszczanek J, Novcic N, Bienz M. Wnt-Dependent Inactivation of the Groucho/TLE Co-repressor by the HECT E3 Ubiquitin Ligase Hyd/UBR5. Mol Cell. 2017;67(2):181-93 e5.

28. Morgan RG, Mortensson E, Williams AC. Targeting LGR5 in Colorectal Cancer: therapeutic gold or too plastic? Br J Cancer. 2018;118(11):1410-8.

29. Mukherjee A, Dhar N, Stathos M, Schaffer DV, Kane RS. Understanding How Wnt Influences Destruction Complex Activity and β-Catenin Dynamics. iScience. 2018;6:13-21.

30. San Roman AK, Jayewickreme CD, Murtaugh LC, Shivdasani RA. Wnt secretion from epithelial cells and subepithelial myofibroblasts is not required in the mouse intestinal stem cell niche in vivo. Stem Cell Reports. 2014;2(2):127-34.

31. Liu J, Xiao Q, Xiao J, Niu C, Li Y, Zhang X, et al. Wnt/beta-catenin signalling: function, biological mechanisms, and therapeutic opportunities. Signal Transduct Target Ther. 2022;7(1):3.

32. Nguyen LH, Goel A, Chung DC. Pathways of Colorectal Carcinogenesis. Gastroenterology. 2020;158(2):291-302.

33. Kosmidou V, Oikonomou E, Vlassi M, Avlonitis S, Katseli A, Tsipras I, et al. Tumor heterogeneity revealed by KRAS, BRAF, and PIK3CA pyrosequencing: KRAS and PIK3CA intratumor mutation profile differences and their therapeutic implications. Hum Mutat. 2014;35(3):329-40.

34. Nitulescu GM, Van De Venter M, Nitulescu G, Ungurianu A, Juzenas P, Peng Q, et al. The Akt pathway in oncology therapy and beyond (Review). Int J Oncol. 2018;53(6):2319-31.

35. McCubrey JA, Steelman LS, Bertrand FE, Davis NM, Abrams SL, Montalto G, et al. Multifaceted roles of GSK-3 and Wnt/beta-catenin in hematopoiesis and leukemogenesis: opportunities for therapeutic intervention. Leukemia. 2014;28(1):15-33.

36. Jeong WJ, Ro EJ, Choi KY. Interaction between Wnt/beta-catenin and RAS-ERK pathways and an anti-cancer strategy via degradations of beta-catenin and RAS by targeting the Wnt/beta-catenin pathway. NPJ Precis Oncol. 2018;2(1):5.

37. Hossain MS, Karuniawati H, Jairoun AA, Urbi Z, Ooi J, John A, et al. Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies. Cancers (Basel). 2022;14(7).

38. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, et al. Genetic alterations during colorectal-tumor development. N Engl J Med. 1988;319(9):525-32.

39. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61(5):759-67.

40. Levy DB, Smith KJ, Beazer-Barclay Y, Hamilton SR, Vogelstein B, Kinzler KW. Inactivation of both APC alleles in human and mouse tumors. Cancer Res. 1994;54(22):5953-8.

41. Oshima M, Oshima H, Kitagawa K, Kobayashi M, Itakura C, Taketo M. Loss of Apc heterozygosity and abnormal tissue building in nascent intestinal polyps in mice carrying a truncated Apc gene. Proc Natl Acad Sci U S A. 1995;92(10):4482-6.

42. Oshima M, Takahashi M, Oshima H, Tsutsumi M, Yazawa K, Sugimura T, et al. Effects of docosahexaenoic acid (DHA) on intestinal polyp development in ApcΔ716 knockout mice. Carcinogenesis. 1995;16(11):2605-7.

43. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8.

44. Sakai E, Nakayama M, Oshima H, Kouyama Y, Niida A, Fujii S, et al. Combined Mutation of Apc, Kras, and Tgfbr2 Effectively Drives Metastasis of Intestinal Cancer. Cancer Res. 2018;78(5):1334-46.

45. Margonis GA, Kim Y, Spolverato G, Ejaz A, Gupta R, Cosgrove D, et al. Association Between Specific Mutations in KRAS Codon 12 and Colorectal Liver Metastasis. JAMA Surg. 2015;150(8):722-9.

46. Jones RP, Sutton PA, Evans JP, Clifford R, McAvoy A, Lewis J, et al. Specific mutations in KRAS codon 12 are associated with worse overall survival in patients with advanced and recurrent colorectal cancer. Br J Cancer. 2017;116(7):923-9.

47. Chen J, Guo F, Shi X, Zhang L, Zhang A, Jin H, et al. BRAF V600E mutation and KRAS codon 13 mutations predict poor survival in Chinese colorectal cancer patients. BMC Cancer. 2014;14:802.

48. Kwak MS, Cha JM, Yoon JY, Jeon JW, Shin HP, Chang HJ, et al. Prognostic value of KRAS codon 13 gene mutation for overall survival in colorectal cancer: Direct and indirect comparison meta-analysis. Medicine (Baltimore). 2017;96(35):e7882.

49. Renaud S, Guerrera F, Seitlinger J, Costardi L, Schaeffer M, Romain B, et al. KRAS exon 2 codon 13 mutation is associated with a better prognosis than codon 12 mutation following lung metastasectomy in colorectal cancer. Oncotarget. 2017;8(2):2514-24.

50. Walther A, Johnstone E, Swanton C, Midgley R, Tomlinson I, Kerr D. Genetic prognostic and predictive markers in colorectal cancer. Nat Rev Cancer. 2009;9(7):489-99.

51. Gonzalez-Gonzalez M, Fontanillo C, Abad MM, Gutierrez ML, Mota I, Bengoechea O, et al. Identification of a characteristic copy number alteration profile by high-resolution single nucleotide polymorphism arrays associated with metastatic sporadic colorectal cancer. Cancer. 2014;120(13):1948-59.

52. Mehlen P, Fearon ER. Role of the dependence receptor DCC in colorectal cancer pathogenesis. J Clin Oncol. 2004;22(16):3420-8.

53. Fleming NI, Jorissen RN, Mouradov D, Christie M, Sakthianandeswaren A, Palmieri M, et al. SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. Cancer Res. 2013;73(2):725-35.

54. Feroz W, Sheikh AMA. Exploring the multiple roles of guardian of the genome: P53. Egyptian Journal of Medical Human Genetics. 2020;21(1):49.

55. Tan BS, Tiong KH, Choo HL, Chung FF, Hii LW, Tan SH, et al. Mutant p53-R273H mediates cancer cell survival and anoikis resistance through AKT-dependent suppression of BCL2-modifying factor (BMF). Cell Death Dis. 2015;6(7):e1826.

56. Velho S, Moutinho C, Cirnes L, Albuquerque C, Hamelin R, Schmitt F, et al. BRAF, KRAS and PIK3CA mutations in colorectal serrated polyps and cancer: primary or secondary genetic events in colorectal carcinogenesis? BMC Cancer. 2008;8:255.

57. Gabelli SB, Huang CH, Mandelker D, Schmidt-Kittler O, Vogelstein B, Amzel LM. Structural effects of oncogenic PI3Kalpha mutations. Curr Top Microbiol Immunol. 2010;347:43-53.

58. Wang H, Liang L, Fang JY, Xu J. Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers. Oncogene. 2016;35(16):2011-9.

59. Diep CB, Kleivi K, Ribeiro FR, Teixeira MR, Lindgjaerde OC, Lothe RA. The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. Genes Chromosomes Cancer. 2006;45(1):31-41.

60. Duan B, Zhao Y, Bai J, Wang J, Duan X, Luo X, et al. Colorectal Cancer: An Overview. In: Morgado-Diaz JA, editor. Gastrointestinal Cancers. Brisbane (AU)2022.

61. Tian J, Afebu KO, Bickerdike A, Liu Y, Prasad S, Nelson BJ. Fundamentals of Bowel Cancer for Biomedical Engineers. Ann Biomed Eng. 2023;51(4):679-701.

62. GOV.UK. Bowel cancer screening: programme overview: GOV.UK; 2021 [updated 17 March 2021. Available from: https://www.gov.uk/guidance/bowel-cancer-screening-

programme-overview#:~:text=population%20screening%20programmes.-,Target%20population,invited%20for%20bowel%20scope%20screening.

63. The American Cancer Society. Treatment of Colon Cancer, by Stage: The American Cancer Society; 2024 [updated 06 February 2024. Available from: https://www.cancer.org/cancer/types/colon-rectal-cancer/treating/by-stage-colon.html.

64. Serrano D, Bonanni B, Brown K. Therapeutic cancer prevention: achievements and ongoing challenges - a focus on breast and colorectal cancer. Mol Oncol. 2019;13(3):579-90.

65. The American Cancer Society. Colorectal Cancer Stages: The American Cancer Society; 2024 [updated 29 Januray 2024. Available from: https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/staged.html.

66. Wang Q, Shen X, Chen G, Du J. Drug Resistance in Colorectal Cancer: From Mechanism to Clinic. Cancers (Basel). 2022;14(12).

67. Cancer Research UK. Chemotherapy for colon cancer: Cancer Research UK; 2022 [updated 04 February 2022. Available from: https://www.cancerresearchuk.org/about-cancer/bowel-cancer/treatment/treatment-colon/colon-chemotherapy.

68. Zhang N, Yin Y, Xu SJ, Chen WS. 5-Fluorouracil: mechanisms of resistance and reversal strategies. Molecules. 2008;13(8):1551-69.

69. Raymond E, Faivre S, Chaney S, Woynarowski J, Cvitkovic E. Cellular and Molecular Pharmacology of Oxaliplatin1. Molecular Cancer Therapeutics. 2002;1(3):227-35.

70. Ozawa S, Miura T, Terashima J, Habano W. Cellular irinotecan resistance in colorectal cancer and overcoming irinotecan refractoriness through various combination trials including DNA methyltransferase inhibitors: a review. Cancer Drug Resist. 2021;4(4):946-64.

71. Howe JR. The impact of DNA testing on management of patients with colorectal cancer. Ann Gastroenterol Surg. 2022;6(1):17-28.

72. Berner A. Presentation: Patient with localised colorectal cancer: NHS Genomics Education Programme; 2023 [updated 28 March 2023. Available from: https://www.genomicseducation.hee.nhs.uk/genotes/in-the-clinic/presentation-patient-with-localised-colorectal-cancer/.

73. Berner A. Mismatch repair deficiency and microsatellite instability: NHS Genomics Education Programme; 2022 [updated 25 March 2022. Available from: https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/mismatch-repair-deficiency-and-microsatellite-instability/#genomic-testing.

74. Berner A. Presentation: Patient with metastatic colorectal cancer: NHS Genomics Education Programme; 2023 [updated 28 March 2023. Available from: https://www.genomicseducation.hee.nhs.uk/genotes/in-the-clinic/presentation-patient-with-metastatic-colorectal-cancer/.

75. Modest DP, Stintzing S, Weikersthal LFv, Decker T, Kiani A, Vehling-Kaiser U, et al. Impact of Subsequent Therapies on Outcome of the FIRE-3/AIO KRK0306 Trial: First-Line Therapy With FOLFIRI Plus Cetuximab or Bevacizumab in Patients With KRAS Wild-Type Tumors in Metastatic Colorectal Cancer. Journal of Clinical Oncology. 2015;33(32):3718-26.

76. Fares CM, Allen EMV, Drake CG, Allison JP, Hu-Lieskovan S. Mechanisms of Resistance to Immune Checkpoint Blockade: Why Does Checkpoint Inhibitor Immunotherapy Not Work for All Patients? American Society of Clinical Oncology Educational Book. 2019(39):147-64.

77. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol. 2005;23(3):609-18.

78. Linnebacher M, Gebert J, Rudy W, Woerner S, Yuan YP, Bork P, et al. Frameshift peptide-derived T-cell epitopes: a source of novel tumor-specific antigens. Int J Cancer. 2001;93(1):6-11.

79. Kosuri KV, Wu X, Wang L, Villalona-Calero MA, Otterson GA. An epigenetic mechanism for capecitabine resistance in mesothelioma. Biochem Biophys Res Commun. 2010;391(3):1465-70.

80. Haibe Y, Kreidieh M, El Hajj H, Khalifeh I, Mukherji D, Temraz S, et al. Resistance Mechanisms to Anti-angiogenic Therapies in Cancer. Front Oncol. 2020;10:221.

81. Michaud M, Martins I, Sukkurwala AQ, Adjemian S, Ma Y, Pellegatti P, et al. Autophagy-Dependent Anticancer Immune Responses Induced by Chemotherapeutic Agents in Mice. Science. 2011;334(6062):1573-7.

82. Hsu HH, Chen MC, Baskaran R, Lin YM, Day CH, Lin YJ, et al. Oxaliplatin resistance in colorectal cancer cells is mediated via activation of ABCG2 to alleviate ER stress induced apoptosis. J Cell Physiol. 2018;233(7):5458-67.

83. Fiszman GL, Jasnis MA. Molecular Mechanisms of Trastuzumab Resistance in HER2 Overexpressing Breast Cancer. Int J Breast Cancer. 2011;2011:352182.

84. Chen X, Zhang W, Yang W, Zhou M, Liu F. Acquired resistance for immune checkpoint inhibitors in cancer immunotherapy: challenges and prospects. Aging (Albany NY). 2022;14(2):1048-64.

85. Nagasaki J, Ishino T, Togashi Y. Mechanisms of resistance to immune checkpoint inhibitors. Cancer Sci. 2022;113(10):3303-12.

86. Jiang L, Li L, Liu Y, Lu L, Zhan M, Yuan S, et al. Drug resistance mechanism of kinase inhibitors in the treatment of hepatocellular carcinoma. Front Pharmacol. 2023;14:1097277.

87. Luebker SA, Koepsell SA. Diverse Mechanisms of BRAF Inhibitor Resistance in Melanoma Identified in Clinical and Preclinical Studies. Front Oncol. 2019;9:268.

88. Fortunato A, Boddy A, Mallo D, Aktipis A, Maley CC, Pepper JW. Natural Selection in Cancer Biology: From Molecular Snowflakes to Trait Hallmarks. Cold Spring Harb Perspect Med. 2017;7(2).

89. Shlush LI, Hershkovitz D. Clonal evolution models of tumor heterogeneity. Am Soc Clin Oncol Educ Book. 2015:e662-5.

90. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? Biochim Biophys Acta Rev Cancer. 2017;1867(2):151-61.

91. Heppner GH, Miller BE. Tumor heterogeneity: biological implications and therapeutic consequences. Cancer Metastasis Rev. 1983;2(1):5-23.

92. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306-13.

93. Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat Med. 2016;22(1):105-13.

94. Grzywa TM, Paskal W, Wlodarski PK. Intratumor and Intertumor Heterogeneity in Melanoma. Transl Oncol. 2017;10(6):956-75.

95. Klein CA. Parallel progression of primary tumours and metastases. Nat Rev Cancer. 2009;9(4):302-12.

96. Geyer FC, Weigelt B, Natrajan R, Lambros MB, de Biase D, Vatcheva R, et al. Molecular analysis reveals a genetic basis for the phenotypic diversity of metaplastic breast carcinomas. J Pathol. 2010;220(5):562-73.

97. Zito Marino F, Liguori G, Aquino G, La Mantia E, Bosari S, Ferrero S, et al. Correction: Intratumor Heterogeneity of ALK-Rearrangements and Homogeneity of EGFR-Mutations in Mixed Lung Adenocarcinoma. PLoS One. 2015;10(10):e0141521.

98. Giaretti W, Monaco R, Pujic N, Rapallo A, Nigro S, Geido E. Intratumor heterogeneity of K-ras2 mutations in colorectal adenocarcinomas: association with degree of DNA aneuploidy. Am J Pathol. 1996;149(1):237-45.

99. Friemel J, Rechsteiner M, Frick L, Bohm F, Struckmann K, Egger M, et al. Intratumor heterogeneity in hepatocellular carcinoma. Clin Cancer Res. 2015;21(8):1951-61.

100. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. Nat Genet. 2015;47(3):209-16.

101. Ling S, Hu Z, Yang Z, Yang F, Li Y, Lin P, et al. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. Proc Natl Acad Sci U S A. 2015;112(47):E6496-505.

102. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nat Genet. 2016;48(3):238-44.

103. Sun R, Hu Z, Curtis C. Big Bang Tumor Growth and Clonal Evolution. Cold Spring Harb Perspect Med. 2018;8(5).

104. Ashouri A, Zhang C, Gaiti F. Decoding Cancer Evolution: Integrating Genetic and Non-Genetic Insights. Genes (Basel). 2023;14(10).

105. Shackleton M, Quintana E, Fearon ER, Morrison SJ. Heterogeneity in cancer: cancer stem cells versus clonal evolution. Cell. 2009;138(5):822-9.

106. Niida A, Mimori K, Shibata T, Miyano S. Modeling colorectal cancer evolution. J Hum Genet. 2021;66(9):869-78.

107. Raghavan S. How inclusive are cell lines in preclinical engineered cancer models? Dis Model Mech. 2022;15(5).

108. Liu Y, Wu W, Cai C, Zhang H, Shen H, Han Y. Patient-derived xenograft models in cancer therapy: technologies and applications. Signal Transduct Target Ther. 2023;8(1):160.

109. Vlachogiannis G, Hedayat S, Vatsiou A, Jamin Y, Fernandez-Mateos J, Khan K, et al. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. Science. 2018;359(6378):920-6.

110. Grivel JC, Margolis L. Use of human tissue explants to study human infectious agents. Nat Protoc. 2009;4(2):256-69.

111. Gunti S, Hoke ATK, Vu KP, London NR, Jr. Organoid and Spheroid Tumor Models: Techniques and Applications. Cancers (Basel). 2021;13(4).

112. Li M, Izpisua Belmonte JC. Organoids - Preclinical Models of Human Disease. N Engl J Med. 2019;380(6):569-79.

113. Lancaster MA, Knoblich JA. Generation of cerebral organoids from human pluripotent stem cells. Nat Protoc. 2014;9(10):2329-40.

114. Wu Q, Liu J, Wang X, Feng L, Wu J, Zhu X, et al. Organ-on-a-chip: recent breakthroughs and future prospects. Biomed Eng Online. 2020;19(1):9.

115. DrugBank. Regorafenib: DrugBank; 2013 [updated 22 February 2024. Available from: https://go.drugbank.com/drugs/DB08896.

116. Ogbeide S, Giannese F, Mincarelli L, Macaulay IC. Into the multiverse: advances in single-cell multiomic profiling. Trends Genet. 2022;38(8):831-43.

117. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361(6400).

118. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. eLife. 2017;6:e27041.

119. Zhang J, Spath SS, Marjani SL, Zhang W, Pan X. Characterization of cancer genomic heterogeneity by next-generation sequencing advances precision medicine in cancer treatment. Precis Clin Med. 2018;1(1):29-48.

120. Mincarelli L, Lister A, Lipscombe J, Macaulay IC. Defining Cell Identity with Single-Cell Omics. Proteomics. 2018;18(18):e1700312.

121. Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nature Genetics. 2006;38(4):468-73.

122. Bozic I, Nowak MA. Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. Proc Natl Acad Sci U S A. 2014;111(45):15964-8.

123. Schmitt MW, Loeb LA, Salk JJ. The influence of subclonal resistance mutations on targeted cancer therapy. Nat Rev Clin Oncol. 2016;13(6):335-47.

124. Lim B, Lin Y, Navin N. Advancing Cancer Research and Medicine with Single-Cell Genomics. Cancer Cell. 2020;37(4):456-70.

125. Lei Y, Tang R, Xu J, Wang W, Zhang B, Liu J, et al. Applications of single-cell sequencing in cancer research: progress and perspectives. J Hematol Oncol. 2021;14(1):91.

126. Venkatesan S, Swanton C, Taylor BS, Costello JF. Treatment-Induced Mutagenesis and Selective Pressures Sculpt Cancer Evolution. Cold Spring Harb Perspect Med. 2017;7(8).

127. Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell. 2018;173(4):879-93 e13.

128. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015;33(3):285-9.

129. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12(6):519-22.

130. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016;13(3):229-32.

131. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016;26(3):304-19.

132. Macaulay IC, Teng MJ, Haerty W, Kumar P, Ponting CP, Voet T. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. Nat Protoc. 2016;11(11):2081-103.

133. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865-8.

134. Han KY, Kim KT, Joung JG, Son DS, Kim YJ, Jo A, et al. SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. Genome Res. 2018;28(1):75-87.

135. Gu C, Liu S, Wu Q, Zhang L, Guo F. Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes. Cell Res. 2019;29(2):110-23.

136. Hu Y, An Q, Guo Y, Zhong J, Fan S, Rao P, et al. Simultaneous Profiling of mRNA Transcriptome and DNA Methylome from a Single Cell. Methods Mol Biol. 2019;1979:363-77.

137. Rodriguez-Meira A, Buck G, Clark SA, Povinelli BJ, Alcolea V, Louka E, et al. Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. Mol Cell. 2019;73(6):1292-305 e8.

138. Hwang B, Lee DS, Tamaki W, Sun Y, Ogorodnikov A, Hartoularos GC, et al. SCITO-seq: single-cell combinatorial indexed cytometry sequencing. Nat Methods. 2021;18(8):903-11.

139. Prathamesh Dhamale SJ. Challenges and Solutions in Single Cell RNA-seq Data Analysis: Elucidata; 2024 [updated 10 February 2023. Available from: https://www.elucidata.io/blog/challenges-and-solutions-in-single-cell-rna-seq-data-analysis.

140. Lahnemann D, Koster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):31.

141. Moffitt JR, Zhuang X. RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). Methods Enzymol. 2016;572:1-49.

142. Okamoto T, duVerle D, Yaginuma K, Natsume Y, Yamanaka H, Kusama D, et al. Comparative Analysis of Patient-Matched PDOs Revealed a Reduction in OLFM4-Associated Clusters in Metastatic Lesions in Colorectal Cancer. Stem Cell Reports. 2021;16(4):954-67.

143. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10(11):1096-8.

144. Moorcraft SY, Gonzalez de Castro D, Cunningham D, Jones T, Walker BA, Peckitt C, et al. Investigating the feasibility of tumour molecular profiling in gastrointestinal malignancies in routine clinical practice. Ann Oncol. 2018;29(1):230-6.

145. Zhang X, Duan R, Wang Y, Liu X, Zhang W, Zhu X, et al. FOLFIRI (folinic acid, fluorouracil, and irinotecan) increases not efficacy but toxicity compared with single-agent irinotecan as a second-line treatment in metastatic colorectal cancer patients: a randomized clinical trial. Therapeutic Advances in Medical Oncology. 2022;14:17588359211068737.

146. Sommer EM, Dry H, Cross D, Guichard S, Davies BR, Alessi DR. Elevated SGK1 predicts resistance of breast cancer cells to Akt inhibitors. Biochem J. 2013;452(3):499-508.

147. Hirai H, Sootome H, Nakatsuru Y, Miyama K, Taguchi S, Tsujioka K, et al. MK-2206, an allosteric Akt inhibitor, enhances antitumor efficacy by standard chemotherapeutic agents or molecular targeted drugs in vitro and in vivo. Mol Cancer Ther. 2010;9(7):1956-67.

148. Hyman DM, Smyth LM, Donoghue MTA, Westin SN, Bedard PL, Dean EJ, et al. AKT Inhibition in Solid Tumors With AKT1 Mutations. J Clin Oncol. 2017;35(20):2251-9.

149. Xing Y, Lin NU, Maurer MA, Chen H, Mahvash A, Sahin A, et al. Phase II trial of AKT inhibitor MK-2206 in patients with advanced breast cancer who have tumors with PIK3CA or AKT mutations, and/or PTEN loss/PTEN mutation. Breast Cancer Res. 2019;21(1):78.

150. Mullard A. FDA approves first-in-class AKT inhibitor. Nat Rev Drug Discov. 2024;23(1):9.

151. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature Methods. 2013;10(11):1096-8.

152. Hahaut V, Picelli S. Full-Length Single-Cell RNA-Sequencing with FLASH-seq. Methods Mol Biol. 2023;2584:123-64.

153. Hammond WA, Swaika A, Mody K. Pharmacologic resistance in colorectal cancer: a review. Ther Adv Med Oncol. 2016;8(1):57-84.

154. Ma SC, Zhang JQ, Yan TH, Miao MX, Cao YM, Cao YB, et al. Novel strategies to reverse chemoresistance in colorectal cancer. Cancer Med. 2023;12(10):11073-96.

155. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021;2(3):100141.

156. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. GigaScience. 2018;7(7).

157. Gu Z. Complex heatmap visualization. iMeta. 2022;1(3):e43.

158. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Github. Available online at https://github.com/s-andrews/FastQC 2010 [

159. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv. 2016:060012.

160. Gennady K, Vladimir S, Nikolay B, Boris S, Maxim NA, Alexey S. Fast gene set enrichment analysis. bioRxiv. 2021:060012.

161. Wickham H, Sievert C, SpringerLink. ggplot2: Elegant Graphics for Data Analysis. Use R! 2nd 2016. ed. Cham: Springer

Springer International Publishing : Imprint: Springer; 2016.

162. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9.

163. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25.

164. Dolgalev I. msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format 2022 [Available from: https://CRAN.R-project.org/package=msigdbr.

165. Dharmesh D. Bhuva GKS, Alexandra Garnham. MSigDB: An ExperimentHub Package for the Molecular Signatures Database. CRAN. Available online at https://cran.r-project.org/web/packages/msigdbr/index.html 2023 [

166. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8.

167. Carlson M. org.Hs.eg.db: Genome wide annotation for Human 2023 [Available from: https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.

168. Rossum V, Guido , Drake L, Fred. Python 3 Reference Manual Scotts Valley, CA: CreateSpace; 2009 [Available from: https://www.python.org/.

169. Garcia-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Gotz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 2012;28(20):2678-9.

170. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292-4.

171. Mingchu X, Zhongqi G. Scillus: Seurat wrapper package enhancing the processing and visualization of single cell data 2021 [Available from: https://github.com/xmc811/Scillus.

172. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018;36(5):411-20.

173. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902.e21.

174. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.e29.

175. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

176. Wickham H AM, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686.

177. Krueger F. Trim Galore 2021 [Available from: https://github.com/FelixKrueger/TrimGalore.

178. Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, et al. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. Github. Available online at https://github.com/FelixKrueger/TrimGalore 2023 [

179. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res. 2018;7:1338.

180. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721-9.

181. Ensembl. Ensembl FTP site: Github; 2023 [Available from: https://ftp.ensembl.org/pub/.

182. Satija Lab. Seurat - Guided Clustering Tutorial: Satija Lab; 2023 [Available from: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html.

183. Satija Lab. Introduction to scRNA-seq integration: Satija Lab; 2023 [Available from: https://satijalab.org/seurat/articles/integration_introduction.html.

184. Andreatta M, Carmona SJ. STACAS: Sub-Type Anchor Correction for Alignment in Seurat to integrate single-cell RNA-seq data. Bioinformatics. 2021;37(6):882-4.

185. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202.

186. Zhu X, Zhang J, Xu Y, Wang J, Peng X, Li HD. Single-Cell Clustering Based on Shared Nearest Neighbor and Graph Partitioning. Interdiscip Sci. 2020;12(2):117-30.

187. Tang M, Kaymaz Y, Logeman BL, Eichhorn S, Liang ZS, Dulac C, et al. Evaluating single-cell cluster stability using the Jaccard similarity index. Bioinformatics. 2020;37(15):2212-4.

188. Metcalf D. On hematopoietic stem cell fate. Immunity. 2007;26(6):669-73.

189. Yang Y, Sun H, Zhang Y, Zhang T, Gong J, Wei Y, et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Reports. 2021;36(4):109442.

190. Elmentaite R, Kumasaka N, Roberts K, Fleming A, Dann E, King HW, et al. Cells of the human intestinal tract mapped across space and time. Nature. 2021;597(7875):250-5.

191. Wellcome Sanger Institute. Gut Cell Survey: Gut Cell Atlas; 2021 [Available from: https://www.gutcellatlas.org/#publications.

192. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nature Biotechnology. 2011;29(12):1120-7.

193. Gehart H, van Es JH, Hamer K, Beumer J, Kretzschmar K, Dekkers JF, et al. Identification of Enteroendocrine Regulators by Real-Time Single-Cell Differentiation Mapping. Cell. 2019;176(5):1158-73.e16.

194. Parikh K, Antanaviciute A, Fawkner-Corbett D, Jagielowicz M, Aulicino A, Lagerholm C, et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. Nature. 2019;567(7746):49-55.

195. Fawkner-Corbett D, Antanaviciute A, Parikh K, Jagielowicz M, Gerós AS, Gupta T, et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. Cell. 2021;184(3):810-26.e23.

196. Tascini AS. iGCA_scRNAseq_analysis - Azoospermia Data Integration Script: Github; 2021 [Available from: https://github.com/volpesofi/iGCA_scRNAseq_analysis/blob/master/Seurat/azoospermia.integration.seurat.ipynb.

197. 10x Genomics. Cell Ranger analysis pipelines: Github; 2024 [Available from: https://github.com/10XGenomics/cellranger.

198. Bioinformatics & Evolutionary Genomics. Calculate and draw custom Venn diagrams 2024 [Available from: https://bioinformatics.psb.ugent.be/webtools/Venn/.

199. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.

200. Chao HP, Chen Y, Takata Y, Tomida MW, Lin K, Kirk JS, et al. Systematic evaluation of RNA-Seq preparation protocol performance. BMC Genomics. 2019;20(1):571.

201. Shi H, Zhou Y, Jia E, Pan M, Bai Y, Ge Q. Bias in RNA-seq Library Preparation: Current Challenges and Solutions. Biomed Res Int. 2021;2021:6647597.

202. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352(6282):189-96.

203. Barbachano A, Fernandez-Barral A, Bustamante-Madrid P, Prieto I, Rodriguez-Salas N, Larriba MJ, et al. Organoids and Colorectal Cancer. Cancers (Basel). 2021;13(11).

204. Yin H, Wang J, Li H, Yu Y, Wang X, Lu L, et al. Extracellular matrix protein-1 secretory isoform promotes ovarian cancer through increasing alternative mRNA splicing and stemness. Nature Communications. 2021;12(1):4230.

205. Glatz JC, Luiken JFP. Dynamic role of the transmembrane glycoprotein CD36 (SR-B2) in cellular fatty acid uptake and utilization. Journal of Lipid Research. 2018;59(7):1084-93.

206. Zhang H, Guo W, Zhang F, Li R, Zhou Y, Shao F, et al. Monoacylglycerol Lipase Knockdown Inhibits Cell Proliferation and Metastasis in Lung Adenocarcinoma. Front Oncol. 2020;10:559568.

207. Liedtke D, Orth M, Meissler M, Geuer S, Knaup S, Köblitz I, et al. ECM alterations in Fndc3a (Fibronectin Domain Containing Protein 3A) deficient zebrafish cause temporal fin development and regeneration defects. Scientific Reports. 2019;9(1):13383.

208. Vidula N, Yau C, Rugo H. Trophoblast Cell Surface Antigen 2 gene (TACSTD2) expression in primary breast cancer. Breast Cancer Res Treat. 2022;194(3):569-75.

209. Yu C-Y, Chang W-C, Zheng J-H, Hung W-H, Cho E-C. Transforming growth factor alpha promotes tumorigenesis and regulates epithelial-mesenchymal transition modulation in colon cancer. Biochemical and Biophysical Research Communications. 2018;506(4):901-6.

210. Mahdi MR, Georges RB, Ali DM, Bedeer RF, Eltahry HM, Gabr AHZ, et al. Modulation of the Endothelin System in Colorectal Cancer Liver Metastasis: Influence of Epigenetic Mechanisms? Front Pharmacol. 2020;11:180.

211. Sun G, Wu L, Sun G, Shi X, Cao H, Tang W. WNT5a in Colorectal Cancer: Research Progress and Challenges. Cancer Manag Res. 2021;13:2483-98.

212. Küçükköse E, Peters NA, Ubink I, van Keulen VAM, Daghighian R, Verheem A, et al. KIT promotes tumor stroma formation and counteracts tumor-suppressive TGFβ signaling in colorectal cancer. Cell Death & Disease. 2022;13(7):617.

213. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017;8(1):14049.

214. Narayanankutty A. PI3K/ Akt/ mTOR Pathway as a Therapeutic Target for Colorectal Cancer: A Review of Preclinical and Clinical Evidence. Curr Drug Targets. 2019;20(12):1217-26.

215. HUGO Gene Nomenclature Committee. Gene group: Ribosomal proteins: HGNC; 2023 [Available from: https://www.genenames.org/data/genegroup/#!/group/1054.

216. Pelletier J, Sonenberg N. The Organizing Principles of Eukaryotic Ribosome Recruitment. Annual Review of Biochemistry. 2019;88(1):307-35.

217. Zeng Q, Lei F, Chang Y, Gao Z, Wang Y, Gao Q, et al. An oncogenic gene, SNRPA1, regulates PIK3R1, VEGFC, MKI67, CDK1 and other genes in colorectal cancer. Biomed Pharmacother. 2019;117:109076.

218. Fang H, Sheng S, Chen B, Wang J, Mao D, Han Y, et al. A Pan-Cancer Analysis of the Oncogenic Role of Cell Division Cycle-Associated Protein 4 (CDCA4) in Human Tumors. Front Immunol. 2022;13:826337.

219. Wang L, Hu XD, Li SY, Liang XY, Ren L, Lv SX. ASPM facilitates colorectal cancer cells migration and invasion by enhancing beta-catenin expression and nuclear translocation. Kaohsiung J Med Sci. 2022;38(2):129-38.

220. Shaath H, Vishnubalaji R, Elango R, Velayutham D, Jithesh PV, Alajez NM. Therapeutic targeting of the TPX2/TTK network in colorectal cancer. Cell Commun Signal. 2023;21(1):265.

221. Zhu C, Zhang L, Zhao S, Dai W, Xu Y, Zhang Y, et al. UPF1 promotes chemoresistance to oxaliplatin through regulation of TOP2A activity and maintenance of stemness in colorectal cancer. Cell Death & Disease. 2021;12(6):519.

222. Liu X, Zhang H, Lai L, Wang X, Loera S, Xue L, et al. Ribonucleotide reductase small subunit M2 serves as a prognostic biomarker and predicts poor survival of colorectal cancers. Clin Sci (Lond). 2013;124(9):567-78.

223. Yu GH, Gong XF, Peng YY, Qian J. Anti-silencing function 1B knockdown suppresses the malignant phenotype of colorectal cancer by inactivating the phosphatidylinositol 3-kinase/AKT pathway. World J Gastrointest Oncol. 2022;14(12):2353-66.

224. Li H, Wang Y, Rong SK, Li L, Chen T, Fan YY, et al. Integrin alpha1 promotes tumorigenicity and progressive capacity of colorectal cancer. Int J Biol Sci. 2020;16(5):815-26.

225. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, et al. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev. 2002;16(2):245-56.

226. Hauffe L, Picard D, Musa J, Remke M, Grunewald TGP, Rotblat B, et al. Eukaryotic translation initiation factor 4E binding protein 1 (EIF4EBP1) expression in glioblastoma is driven by ETS1- and MYBL2-dependent transcriptional activation. Cell Death Discov. 2022;8(1):91.

227. Duan R, Du W, Guo W. EZH2: a novel target for cancer treatment. J Hematol Oncol. 2020;13(1):104.

228. Liu G, Zhan W, Guo W, Hu F, Qin J, Li R, et al. MELK Accelerates the Progression of Colorectal Cancer via Activating the FAK/Src Pathway. Biochem Genet. 2020;58(5):771-82.

229. National Center for Biotechnology Information USNLoM. National Institutes of Health; 2023 [Available from: https://www.ncbi.nlm.nih.gov/gene/.

230. Nassir F, Wilson B, Han X, Gross RW, Abumrad NA. CD36 Is Important for Fatty Acid and Cholesterol Uptake by the Proximal but Not Distal Intestine*. Journal of Biological Chemistry. 2007;282(27):19493-501.

231. Lund ML, Egerod KL, Engelstoft MS, Dmytriyeva O, Theodorsson E, Patel BA, et al. Enterochromaffin 5-HT cells – A major target for GLP-1 and gut microbial metabolites. Molecular Metabolism. 2018;11:70-83.

232. Tomita H, Tanaka K, Tanaka T, Hara A. Aldehyde dehydrogenase 1A1 in stem cells and cancer. Oncotarget. 2016;7(10):11018-32.

233. Wang Y, Chen Y, Garcia-Milian R, Golla JP, Charkoftaki G, Lam TT, et al. Proteomic profiling reveals an association between ALDH and oxidative phosphorylation and DNA damage repair pathways in human colon adenocarcinoma stem cells. Chemico-Biological Interactions. 2022;368:110175.

234. Zilbauer M, James KR, Kaur M, Pott S, Li Z, Burger A, et al. A Roadmap for the Human Gut Cell Atlas. Nature Reviews Gastroenterology & Hepatology. 2023;20(9):597-614.

235. Cancedda R, Mastrogiacomo M. Transit Amplifying Cells (TACs): a still not fully understood cell population. Frontiers in Bioengineering and Biotechnology. 2023;11.

236. Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, et al. A single-cell survey of the small intestinal epithelium. Nature. 2017;551(7680):333-9.

237. Fujii M, Matano M, Toshimitsu K, Takano A, Mikami Y, Nishikori S, et al. Human Intestinal Organoids Maintain Self-Renewal Capacity and Cellular Diversity in Niche-Inspired Culture Condition. Cell Stem Cell. 2018;23(6):787-93.e6.

238. Tian Y, Denda-Nagai K, Tsukui T, Ishii-Schrade KB, Okada K, Nishizono Y, et al. Mucin 21 confers resistance to apoptosis in an O-glycosylation-dependent manner. Cell Death Discovery. 2022;8(1):194.

239. Yoshimoto T, Matsubara D, Soda M, Ueno T, Amano Y, Kihara A, et al. Mucin 21 is a key molecule involved in the incohesive growth pattern in lung adenocarcinoma. Cancer Sci. 2019;110(9):3006-11.

240. Shvartsur A, Bonavida B. Trop2 and its overexpression in cancers: regulation and clinical/therapeutic implications. Genes Cancer. 2015;6(3-4):84-105.

241. Kumar D, Vetrivel U, Parameswaran S, Subramanian KK. Structural insights on druggable hotspots in CD147: A bull's eye view. Life Sci. 2019;224:76-87.

242. Feng W, Cui G, Tang CW, Zhang XL, Dai C, Xu YQ, et al. Role of glucose metabolism related gene GLUT1 in the occurrence and prognosis of colorectal cancer. Oncotarget. 2017;8(34):56850-7.

243. Wang X, Shen X, Yan Y, Li H. Pyruvate dehydrogenase kinases (PDKs): an overview toward clinical applications. Biosci Rep. 2021;41(4).

244. Xu B, Chen L, Zhan Y, Marquez KNS, Zhuo L, Qi S, et al. The Biological Functions and Regulatory Mechanisms of Fatty Acid Binding Protein 5 in Various Diseases. Front Cell Dev Biol. 2022;10:857919.

245. Ramos FS, Serino LT, Carvalho CM, Lima RS, Urban CA, Cavalli IJ, et al. PDIA3 and PDIA6 gene expression as an aggressiveness marker in primary ductal breast cancer. Genet Mol Res. 2015;14(2):6960-7.

246. Yang Z, Liu J, Shi Q, Chao Y, Di Y, Sun J, et al. Expression of protein disulfide isomerase A3 precursor in colorectal cancer. Onco Targets Ther. 2018;11:4159-66.

247. Han Z, Wang Y, Han L, Yang C. RPN2 in cancer: An overview. Gene. 2023;857:147168.

248. Wallace L, Mehrabi S, Bacanamwo M, Yao X, Aikhionbare FO. Expression of mitochondrial genes MT-ND1, MT-ND6, MT-CYB, MT-COI, MT-ATP6, and 12S/MT-RNR1 in colorectal adenopolyps. Tumour Biol. 2016;37(9):12465-75.

249. Li T, Forbes ME, Fuller GN, Li J, Yang X, Zhang W. IGFBP2: integrative hub of developmental and oncogenic signaling network. Oncogene. 2020;39(11):2243-57.

250. Qiao Q, Bai R, Song W, Gao H, Zhang M, Lu J, et al. Human α-defensin 5 suppressed colon cancer growth by targeting PI3K pathway. Experimental Cell Research. 2021;407(2):112809.

251. Wang Y, Kang X, Kang X, Yang F. S100A6: molecular function and biomarker role. Biomarker Research. 2023;11(1):78.

252. Xia C, Yin S, To KKW, Fu L. CD39/CD73/A2AR pathway and cancer immunotherapy. Molecular Cancer. 2023;22(1):44.

253. Westrich JA, Vermeer DW, Colbert PL, Spanos WC, Pyeon D. The multifarious roles of the chemokine CXCL14 in cancer progression and immune responses. Mol Carcinog. 2020;59(7):794-806.

254. Carabet LA, Leblanc E, Lallous N, Morin H, Ghaidi F, Lee J, et al. Computer-Aided Discovery of Small Molecules Targeting the RNA Splicing Activity of hnRNP A1 in Castration-Resistant Prostate Cancer. Molecules. 2019;24(4):763.

255. Hong Z, Xu C, Zheng S, Wang X, Tao Y, Tan Y, et al. Nucleophosmin 1 cooperates with BRD4 to facilitate c-Myc transcription to promote prostate cancer progression. Cell Death Discovery. 2023;9(1):392.

256. Abbas W, Kumar A, Herbein G. The eEF1A Proteins: At the Crossroads of Oncogenesis, Apoptosis, and Viral Infections. Frontiers in Oncology. 2015;5.

257. Wei LF, Weng XF, Huang XC, Peng YH, Guo HP, Xu YW. IGFBP2 in cancer: Pathological role and clinical significance (Review). Oncol Rep. 2021;45(2):427-38.

258. Li JJ, Xie D. RACK1, a versatile hub in cancer. Oncogene. 2015;34(15):1890-8.

259. Krossa I, Strub T, Martel A, Nahon-Esteve S, Lassalle S, Hofman P, et al. Recent advances in understanding the role of HES6 in cancers. Theranostics. 2022;12(9):4374-85.

260. Muhammad S, Fan T, Hai Y, Gao Y, He J. Reigniting hope in cancer treatment: the promise and pitfalls of IL-2 and IL-2R targeting strategies. Molecular Cancer. 2023;22(1):121.

261. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746.

262. Pascual G, Avgustinova A, Mejetta S, Martín M, Castellanos A, Attolini CS-O, et al. Targeting metastasis-initiating cells through the fatty acid receptor CD36. Nature. 2017;541(7635):41-5.

263. Wuensch T, Wizenty J, Quint J, Spitz W, Bosma M, Becker O, et al. Expression Analysis of Fibronectin Type III Domain-Containing (FNDC) Genes in Inflammatory Bowel Disease and Colorectal Cancer. Gastroenterol Res Pract. 2019;2019:3784172.

264. Drury J, Rychahou PG, Kelson CO, Geisen ME, Wu Y, He D, et al. Upregulation of CD36, a Fatty Acid Translocase, Promotes Colorectal Cancer Metastasis by Increasing MMP28 and Decreasing E-Cadherin Expression. Cancers (Basel). 2022;14(1).

265. Long S, Wang J, Weng F, Pei Z, Zhou S, Sun G, et al. ECM1 regulates the resistance of colorectal cancer to 5-FU treatment by modulating apoptotic cell death and epithelial-mesenchymal transition induction. Front Pharmacol. 2022;13:1005915.

266. Wang J, Day R, Dong Y, Weintraub SJ, Michel L. Identification of Trop-2 as an oncogene and an attractive therapeutic target in colon cancers. Molecular Cancer Therapeutics. 2008;7(2):280-5.

267. Fang YJ, Lu ZH, Wang GQ, Pan ZZ, Zhou ZW, Yun JP, et al. Elevated expressions of MMP7, TROP2, and survivin are associated with survival, disease recurrence, and liver metastasis of colon cancer. International Journal of Colorectal Disease. 2009;24(8):875-84.

268. Wen Y, Ouyang D, Zou Q, Chen Q, Luo N, He H, et al. A literature review of the promising future of TROP2: a potential drug therapy target. Ann Transl Med. 2022;10(24):1403.

269. Xie D, Pei Q, Li J, Wan X, Ye T. Emerging Role of E2F Family in Cancer Stem Cells. Front Oncol. 2021;11:723137.

270. Ebrahimi N, Afshinpour M, Fakhr SS, Kalkhoran PG, Shadman-Manesh V, Adelian S, et al. Cancer stem cells in colorectal cancer: Signaling pathways involved in stemness and therapy resistance. Critical Reviews in Oncology/Hematology. 2023;182:103920.

271. Fang Y, Yu H, Liang X, Xu J, Cai X. Chk1-induced CCNB1 overexpression promotes cell proliferation and tumor growth in human colorectal cancer. Cancer Biol Ther. 2014;15(9):1268-79.

272. Imaoka H, Toiyama Y, Saigusa S, Kawamura M, Kawamoto A, Okugawa Y, et al. RacGAP1 expression, increasing tumor malignant potential, as a predictive biomarker for lymph node metastasis and poor prognosis in colorectal cancer. Carcinogenesis. 2015;36(3):346-54.

273. Chen JF, Luo X, Xiang LS, Li HT, Zha L, Li N, et al. EZH2 promotes colorectal cancer stem-like cell expansion by activating p21cip1-Wnt/beta-catenin signaling. Oncotarget. 2016;7(27):41540-58.

274. Chen Y, Wang J, Fan H, Xie J, Xu L, Zhou B. Phosphorylated 4E-BP1 is associated with tumor progression and adverse prognosis in colorectal cancer. Neoplasma. 2017;64(5):787-94.

275. Giuliano CJ, Lin A, Smith JC, Palladino AC, Sheltzer JM. MELK expression correlates with tumor mitotic activity but is not required for cancer growth. eLife. 2018;7:e32838.

276. Gan Y, Li Y, Li T, Shu G, Yin G. CCNA2 acts as a novel biomarker in regulating the growth and apoptosis of colorectal cancer. Cancer Manag Res. 2018;10:5113-24.

277. Akabane S, Oue N, Sekino Y, Asai R, Thang PQ, Taniyama D, et al. KIFC1 regulates ZWINT to promote tumor progression and spheroid formation in colorectal cancer. Pathol Int. 2021;71(7):441-52.

278. Choi S, Ku J-L. Resistance of colorectal cancer cells to radiation and 5-FU is associated with MELK expression. Biochemical and Biophysical Research Communications. 2011;412(2):207-13.

279. Wang Y, Lee Y-M, Baitsch L, Huang A, Xiang Y, Tong H, et al. MELK is an oncogenic kinase essential for mitotic progression in basal-like breast cancer cells. eLife. 2014;3:e01763.

280. Lai X, Li Q, Wu F, Lin J, Chen J, Zheng H, et al. Epithelial-Mesenchymal Transition and Metabolic Switching in Cancer: Lessons From Somatic Cell Reprogramming. Front Cell Dev Biol. 2020;8:760.

281. Parlani M, Jorgez C, Friedl P. Plasticity of cancer invasion and energy metabolism. Trends Cell Biol. 2023;33(5):388-402.

282. Bhattacharya R, Blankenheim Z, Scott PM, Cormier RT. CFTR and Gastrointestinal Cancers: An Update. J Pers Med. 2022;12(6).

283. Chen G, Gong T, Wang Z, Wang Z, Lin X, Chen S, et al. Colorectal cancer organoid models uncover oxaliplatin-resistant mechanisms at single cell resolution. Cell Oncol (Dordr). 2022;45(6):1155-67.

284. Chen KY, Srinivasan T, Lin C, Tung KL, Gao Z, Hsu DS, et al. Single-Cell Transcriptomics Reveals Heterogeneity and Drug Response of Human Colorectal Cancer Organoids. Annu Int Conf IEEE Eng Med Biol Soc. 2018;2018:2378-81.

285. Biostars. What's the disadvantage of removing cell-cylcle gene compared with regressing out: Biostars; 2022 [Available from: https://www.biostars.org/p/9533437/.

286. Enane FO, Saunthararajah Y, Korc M. Differentiation therapy and the mechanisms that terminate cancer cell proliferation without harming normal cells. Cell Death Dis. 2018;9(9):912.

287. Dean M, Fojo T, Bates S. Tumour stem cells and drug resistance. Nat Rev Cancer. 2005;5(4):275-84.

288. Visvader JE, Lindeman GJ. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat Rev Cancer. 2008;8(10):755-68.

289. Alison MR, Lim SM, Nicholson LJ. Cancer stem cells: problems for therapy? J Pathol. 2011;223(2):147-61.

290. Phi LTH, Sari IN, Yang YG, Lee SH, Jun N, Kim KS, et al. Cancer Stem Cells (CSCs) in Drug Resistance and their Therapeutic Implications in Cancer Treatment. Stem Cells Int. 2018;2018:5416923.

291. Lopez-Lazaro M. The migration ability of stem cells can explain the existence of cancer of unknown primary site. Rethinking metastasis. Oncoscience. 2015;2(5):467-75.

292. Nouri M, Caradec J, Lubik AA, Li N, Hollier BG, Takhar M, et al. Therapy-induced developmental reprogramming of prostate cancer cells and acquired therapy resistance. Oncotarget. 2017;8(12):18949-67.

293. Cojoc M, Mabert K, Muders MH, Dubrovska A. A role for cancer stem cells in therapy resistance: cellular and molecular mechanisms. Semin Cancer Biol. 2015;31:16-27.

294. Kiyohara MH, Dillard C, Tsui J, Kim SR, Lu J, Sachdev D, et al. EMP2 is a novel therapeutic target for endometrial cancer stem cells. Oncogene. 2017;36(42):5793-807.

295. Fang D, Kitamura H. Cancer stem cells and epithelial-mesenchymal transition in urothelial carcinoma: Possible pathways and potential therapeutic approaches. Int J Urol. 2018;25(1):7-17.

296. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell. 2008;133(4):704-15.

297. Singh A, Settleman J. EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. Oncogene. 2010;29(34):4741-51.

298. Garcia-Hernandez V, Quiros M, Nusrat A. Intestinal epithelial claudins: expression and regulation in homeostasis and inflammation. Ann N Y Acad Sci. 2017;1397(1):66-79.

299. Holmes JL, Van Itallie CM, Rasmussen JE, Anderson JM. Claudin profiling in the mouse during postnatal intestinal development and along the gastrointestinal tract reveals complex expression patterns. Gene Expression Patterns. 2006;6(6):581-8.

300. Zhai Y, Lu Q, Lou T, Cao G, Wang S, Zhang Z. MUC16 affects the biological functions of ovarian cancer cells and induces an antitumor immune response by activating dendritic cells. Ann Transl Med. 2020;8(22):1494.

301. Sun L, Zhang Y, Li W, Zhang J, Zhang Y. Mucin Glycans: A Target for Cancer Therapy. Molecules [Internet]. 2023; 28(20).

302. Xiong L, Edwards CK, 3rd, Zhou L. The biological function and clinical utilization of CD147 in human diseases: a review of the current scientific literature. Int J Mol Sci. 2014;15(10):17411-41.

303. Guo C, Liu S, Wang J, Sun M-Z, Greenaway FT. ACTB in cancer. Clinica Chimica Acta. 2013;417:39-44.

304. Gemoll T, Strohkamp S, Schillo K, Thorns C, Habermann JK. MALDI-imaging reveals thymosin beta-4 as an independent prognostic marker for colorectal cancer. Oncotarget; Vol 6, No 41. 2015.

305. Sousa-Squiavinato ACM, Rocha MR, Barcellos-de-Souza P, de Souza WF, Morgado-Diaz JA. Cofilin-1 signaling mediates epithelial-mesenchymal transition by promoting actin cytoskeleton reorganization and cell-cell adhesion regulation in colorectal cancer cells. Biochim Biophys Acta Mol Cell Res. 2019;1866(3):418-29.

306. Phadngam S, Castiglioni A, Ferraresi A, Morani F, Follo C, Isidoro C. PTEN dephosphorylates AKT to prevent the expression of GLUT1 on plasmamembrane and to limit glucose consumption in cancer cells. Oncotarget. 2016;7(51):84999-5020.

307. Lu Y, Jiang Z, Wang K, Yu S, Hao C, Ma Z, et al. Blockade of the amino acid transporter SLC6A14 suppresses tumor growth in colorectal Cancer. BMC Cancer. 2022;22(1):833.

308. Qian X, Xu W, Xu J, Shi Q, Li J, Weng Y, et al. Enolase 1 stimulates glycolysis to promote chemoresistance in gastric cancer. Oncotarget. 2017;8(29):47691-708.

309. Ghanem N, El-Baba C, Araji K, El-Khoury R, Usta J, Darwiche N. The Pentose Phosphate Pathway in Cancer: Regulation and Therapeutic Opportunities. Chemotherapy. 2021;66(5-6):179-91.

310. Redza-Dutordoir M, Averill-Bates DA. Activation of apoptosis signalling pathways by reactive oxygen species. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2016;1863(12):2977-92.

311. Alfarouk KO, Ahmed SBM, Elliott RL, Benoit A, Alqahtani SS, Ibrahim ME, et al. The Pentose Phosphate Pathway Dynamics in Cancer and Its Dependency on Intracellular pH. Metabolites. 2020;10(7).

312. Riganti C, Gazzano E, Polimeni M, Aldieri E, Ghigo D. The pentose phosphate pathway: An antioxidant defense and a crossroad in tumor cell fate. Free Radical Biology and Medicine. 2012;53(3):421-36.

313. Nenkov M, Ma Y, Gassler N, Chen Y. Metabolic Reprogramming of Colorectal Cancer Cells and the Microenvironment: Implication for Therapy. Int J Mol Sci. 2021;22(12).

314. Dong Y, Zheng M, Wang X, Yu C, Qin T, Shen X. High expression of CDKN2A is associated with poor prognosis in colorectal cancer and may guide PD-1-mediated immunotherapy. BMC Cancer. 2023;23(1):1097.

315. Dahlmann M, Werner R, Kortum B, Kobelt D, Walther W, Stein U. Restoring Treatment Response in Colorectal Cancer Cells by Targeting MACC1-Dependent ABCB1 Expression in Combination Therapy. Front Oncol. 2020;10:599.

316. Vanhove K, Graulus GJ, Mesotten L, Thomeer M, Derveaux E, Noben JP, et al. The Metabolic Landscape of Lung Cancer: New Insights in a Disturbed Glucose Metabolism. Front Oncol. 2019;9:1215.

317. Xiong X, Wang S, Gao Z, Ye Y. C6orf15 acts as a potential novel marker of adverse pathological features and prognosis for colon cancer. Pathology - Research and Practice. 2023;245:154426.

318. Shivakumar BM, Chakrabarty S, Rotti H, Seenappa V, Rao L, Geetha V, et al. Comparative analysis of copy number variations in ulcerative colitis associated and sporadic colorectal neoplasia. BMC Cancer. 2016;16:271.

319. Ried T, Meijer GA, Harrison DJ, Grech G, Franch-Expósito S, Briffa R, et al. The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome. Molecular Aspects of Medicine. 2019;69:48-61.

320. Liu F, Ai FY, Zhang DC, Tian L, Yang ZY, Liu SJ. LncRNA NEAT1 knockdown attenuates autophagy to elevate 5-FU sensitivity in colorectal cancer via targeting miR-34a. Cancer Med. 2020;9(3):1079-91.

321. Liu J, Xiao Q, Xiao J, Niu C, Li Y, Zhang X, et al. Wnt/β-catenin signalling: function, biological mechanisms, and therapeutic opportunities. Signal Transduction and Targeted Therapy. 2022;7(1):3.

322. Grassilli E, Cerrito MG. Emerging actionable targets to treat therapy-resistant colorectal cancers. Cancer Drug Resist. 2022;5(1):36-63.

323. Fong L, Hotson A, Powderly JD, Sznol M, Heist RS, Choueiri TK, et al. Adenosine 2A Receptor Blockade as an Immunotherapy for Treatment-Refractory Renal Cell Cancer. Cancer Discovery. 2020;10(1):40-53.

324. Baxter RC. Signaling Pathways of the Insulin-like Growth Factor Binding Proteins. Endocr Rev. 2023;44(5):753-78.

325. Hermanto U, Zong CS, Li W, Wang LH. RACK1, an insulin-like growth factor I (IGF-I) receptor-interacting protein, modulates IGF-I-dependent integrin signaling and promotes cell spreading and contact with extracellular matrix. Mol Cell Biol. 2002;22(7):2345-65.

326. Liu Y, Nelson MV, Bailey C, Zhang P, Zheng P, Dome JS, et al. Targeting the HIF-1alpha-IGFBP2 axis therapeutically reduces IGF1-AKT signaling and blocks the growth and metastasis of relapsed anaplastic Wilms tumor. Oncogene. 2021;40(29):4809-19.

327. Schmidt S, Denk S, Wiegering A. Targeting Protein Synthesis in Colorectal Cancer. Cancers (Basel). 2020;12(5).

328. Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics. 1992;13(3):718-25.

329. Barbaux S, Poirier O, Cambien F. Use of degenerate oligonucleotide primed PCR (DOP-PCR) for the genotyping of low-concentration DNA samples. J Mol Med (Berl). 2001;79(5-6):329-32.

330. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A. 2002;99(8):5261-6.

331. Arneson N, Hughes S, Houlston R, Done S. Whole-Genome Amplification by Degenerate Oligonucleotide Primed PCR (DOP-PCR). CSH Protoc. 2008;2008:pdb prot4919.

332. Hou Y, Wu K, Shi X, Li F, Song L, Wu H, et al. Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. Gigascience. 2015;4:37.

333. Kalef-Ezra E, Turan ZG, Perez-Rodriguez D, Bomann I, Behera S, Morley C, et al. Single-cell somatic copy number variants in brain using different amplification methods and reference genomes. bioRxiv. 2023.

334. Burbulis IE, Wierman MB, Wolpert M, Haakenson M, Lopes MB, Schiff D, et al. Improved molecular karyotyping in glioblastoma. Mutat Res. 2018;811:16-26.

335. Gonzalez-Pena V, Natarajan S, Xia Y, Klein D, Carter R, Pang Y, et al. Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci U S A. 2021;118(24).

336. Borgstrom E, Paterlini M, Mold JE, Frisen J, Lundeberg J. Comparison of whole genome amplification techniques for human single cell exome sequencing. PLoS One. 2017;12(2):e0171566.

337. Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol. 2020;21(1):208.

338. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.

339. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

340. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297-303.

341. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.

342. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, et al. Interactive analysis and assessment of single-cell copy-number variations. Nature Methods. 2015;12(11):1058-60.

343. Broad Institute. Broad Institute. Picard: Broad Institute; 2023 [Available from: https://broadinstitute.github.io/picard/.

344. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

345. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14(2):144-61.

346. Caetano-Anolles D. (How to) Map and clean up short read sequence data efficiently: Broad Institute; 2023 [Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360039568932#step1.

347. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11 0 1- 0 33.

348. Caetano-Anolles D. Data pre-processing for variant discovery: Broad Institute; 2023 [Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035535912-Data-pre-processing-for-variant-discovery.

349. Broad Institute. GATK resource bundle: Broad Institute; 2023 [Available from: https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0;tab=objects?pli=1&prefix=&forceOnObjectsSortingFiltering=false.

350. UCSC. UCSC Golden Path liftOver site: UCSC; 2023 [Available from: https://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/.

351. Chaudhary S. Why "1.5" in IQR Method of Outlier Detection? : Towards Data Science; 2019 [Available from: https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097.

352. UCSC. UCSC hg19 genome annotation database UCSC; 2023 [Available from: https://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/.

353. Broad Institute. GATK - Genome Analysis Toolkit DownsampleSam (Picard): Broad Institute; 2023 [Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360037226232-DownsampleSam-Picard-.

354. nf-core. Analysis pipeline to detect germline or somatic variants: Github; 2023 [Available from: https://github.com/nf-core/sarek.

355. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. 2015;26(1):64-70.

356. Chela J. Mutectplatypus analysis pipeline: Github; 2023 [Available from: https://github.com/chelauk/nf-core-mutectplatypus.

357. Chela J. Mutectplatypus analysis pipeline, analyse_cn_sequenza.R.: Github; 2023 [Available from: https://github.com/chelauk/nf-core-mutectplatypus/blob/master/bin/analyse_cn_sequenza.R.

358. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19(9):1586-92.

359. Sepulveda N, Campino SG, Assefa SA, Sutherland CJ, Pain A, Clark TG. A Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. BMC Genomics. 2013;14:128.

360. Toure AY, Dossou-Gbete S, Kokonendji CC. Asymptotic normality of the test statistics for the unified relative dispersion and relative variation indexes. J Appl Stat. 2020;47(13-15):2479-91.

361. National Institute of Standards and Technology. Index of dispersion: National Institute of Standards and Technology; 2023 [updated 30 June 2017. Available from: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/ind_disp.htm#:~:text=The%20index%20of%20dispersion%20is%20sometimes%20used%20for%20count%20data,should%20be%20greater%20than%201.

362. Zrelak PA. Use of the Poisson Distribution Is a Helpful Tool That Is Underused in Nursing Practice. J Nurs Care Qual. 2022;37(3):E54-E7.

363. Harris T, Yang Z, Hardin JW. Modeling Underdispersed Count Data with Generalized Poisson Regression. The Stata Journal. 2012;12(4):736-47.

364. Shimizu Y. Multiple Desirable Methods in Outlier Detection of Univariate Data With R Source Codes. Front Psychol. 2021;12:819854.

365. Koen T, Sebastiaan V, Florian R, Daniel B, Michiel Van Der H, Oskar M-B, et al. Single-cell Genome-and-Transcriptome sequencing without upfront whole-genome amplification reveals cell state plasticity of melanoma subclones. bioRxiv. 2023:2023.01.13.521174.

366. Chandramohan R, Reuther J, Gandhi I, Voicu H, Alvarez KR, Plon SE, et al. A Validation Framework for Somatic Copy Number Detection in Targeted Sequencing Panels. J Mol Diagn. 2022;24(7):760-74.

367. Rohrback S, April C, Kaper F, Rivera RR, Liu CS, Siddoway B, et al. Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. Proc Natl Acad Sci U S A. 2018;115(42):10804-9.

368. Gusnanto A, Taylor CC, Nafisah I, Wood HM, Rabbitts P, Berri S. Estimating optimal window size for analysis of low-coverage next-generation sequence data. Bioinformatics. 2014;30(13):1823-9.

369. Ravindran A, Krieger KL, Kaushik AK, Hovington H, Mehdi S, Piyarathna DWB, et al. Uridine Diphosphate Glucuronosyl Transferase 2B28 (UGT2B28) Promotes Tumor Progression and Is Elevated in African American Prostate Cancer Patients. Cells. 2022;11(15).

370. Biezuner T, Raz O, Amir S, Milo L, Adar R, Fried Y, et al. Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. Sci Rep. 2021;11(1):17171.

371. Erfanian N, Heydari AA, Feriz AM, Ianez P, Derakhshani A, Ghasemigol M, et al. Deep learning applications in single-cell genomics and transcriptomics data analysis. Biomed Pharmacother. 2023;165:115077.

372. Kabel J, Henriksen TV, Demuth C, Frydendahl A, Rasmussen MH, Nors J, et al. Impact of Whole Genome Doubling on Detection of Circulating Tumor DNA in Colorectal Cancer. Cancers (Basel). 2023;15(4).

373. Pino MS, Chung DC. The chromosomal instability pathway in colon cancer. Gastroenterology. 2010;138(6):2059-72.

374. Wang X, Zhang H, Chen X. Drug resistance and combating drug resistance in cancer. Cancer Drug Resist. 2019;2(2):141-60.

375. Andel D, Viergever BJ, Peters NA, Elisabeth Raats DA, Schenning-van Schelven SJ, Willem Intven MP, et al. Pre-existing subclones determine radioresistance in rectal cancer organoids. Cell Rep. 2024;43(2):113735.

376. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012;481(7382):506-10.

377. Park DJ, Kwon A, Cho BS, Kim HJ, Hwang KA, Kim M, et al. Characteristics of DNMT3A mutations in acute myeloid leukemia. Blood Res. 2020;55(1):17-26.

378. Zhou F, Chen B. Acute myeloid leukemia carrying ETV6 mutations: biologic and clinical features. Hematology. 2018;23(9):608-12.

379. Liu R, Chen Y, Liu G, Li C, Song Y, Cao Z, et al. PI3K/AKT pathway as a key link modulates the multidrug resistance of cancers. Cell Death Dis. 2020;11(9):797.

380. Della Corte CM, Viscardi G, Di Liello R, Fasano M, Martinelli E, Troiani T, et al. Role and targeting of anaplastic lymphoma kinase in cancer. Molecular Cancer. 2018;17(1):30.

381. Bavi P, Jehan Z, Bu R, Prabhakaran S, Al-Sanea N, Al-Dayel F, et al. ALK gene amplification is associated with poor prognosis in colorectal carcinoma. Br J Cancer. 2013;109(10):2735-43.

382. Shen W, Zhou Q, Peng C, Li J, Yuan Q, Zhu H, et al. FBXW7 and the Hallmarks of Cancer: Underlying Mechanisms and Prospective Strategies. Front Oncol. 2022;12:880077.

383. Lan H, Sun Y. FBXW7 E3 ubiquitin ligase: degrading, not degrading, or being degraded. Protein Cell. 2019;10(12):861-3.

384. Iwatsuki M, Mimori K, Ishii H, Yokobori T, Takatsuno Y, Sato T, et al. Loss of FBXW7, a cell cycle regulating gene, in colorectal cancer: clinical significance. Int J Cancer. 2010;126(8):1828-37.

385. Mikubo M, Inoue Y, Liu G, Tsao MS. Mechanism of Drug Tolerant Persister Cancer Cells: The Landscape and Clinical Implication for Therapy. J Thorac Oncol. 2021;16(11):1798-809.

386. Zhu X, Li S, Xu B, Luo H. Cancer evolution: A means by which tumors evade treatment. Biomed Pharmacother. 2021;133:111016.

387. Bravo-Estupinan DM, Aguilar-Guerrero K, Quiros S, Acon MS, Marin-Muller C, Ibanez-Hernandez M, et al. Gene dosage compensation: Origins, criteria to identify compensated genes, and mechanisms including sensor loops as an emerging systems-level property in cancer. Cancer Med. 2023;12(24):22130-55.

388. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res. 2023;51(D1):D933-D41.

389. Tickle T, Tirosh I, Georgescu C, Brown M, Haas B. inferCNV of the Trinity CTAT Project. 2019 [Available from: https://github.com/broadinstitute/inferCNV.

390. Broad Institute. InferCNV: Inferring copy number alterations from tumor single cell RNA-Seq data: Github; 2023 [Available from: https://github.com/broadinstitute/inferCNV/wiki.

391. Broad Institute. Trinity CNV repository Github; 2022 [Available from: https://data.broadinstitute.org/Trinity/CTAT/cnv/.

392. UCSC. UCSC hg38 genome annotation database: UCSC; 2022 [Available from: https://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/.

393. Dehner C, Moon CI, Zhang X, Zhou Z, Miller C, Xu H, et al. Chromosome 8 gain is associated with high-grade transformation in MPNST. JCI Insight. 2021;6(6).

394. Zhang X, Hirbe A. Inference of copy number variations and clonality analysis: Bio-protocol 2022 [Available from: https://bio-protocol.org/exchange/preprintdetail?id=1885&type=3.

395. Ensembl. Ensembl BioMart: Ensembl; 2024 [Available from: https://www.ensembl.org/biomart/martview/d11be1c7082d29d2a9de1d498a881ded.

396. UCSC. Lift Genome Annotations: UCSC; 2023 [Available from: https://genome.ucsc.edu/cgi-bin/hgLiftOver.

397. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396-401.

398. Merid SK, Goranskaya D, Alexeyenko A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC Bioinformatics. 2014;15(1):308.

399. Casadesús J, Noyer-Weidner M. Epigenetics. In: Maloy S, Hughes K, editors. Brenner's Encyclopedia of Genetics (Second Edition). San Diego: Academic Press; 2013. p. 500-3.

400. Mattioli K, Oliveros W, Gerhardinger C, Andergassen D, Maass PG, Rinn JL, et al. Cis and trans effects differentially contribute to the evolution of promoters and enhancers. Genome Biol. 2020;21(1):210.

401. Bian C, Liu Z, Li D, Zhen L. PI3K/AKT inhibition induces compensatory activation of the MET/STAT3 pathway in non-small cell lung cancer. Oncol Lett. 2018;15(6):9655-62.

402. Gowhari Shabgah A, Haleem Al-Qaim Z, Markov A, Valerievich Yumashev A, Ezzatifar F, Ahmadi M, et al. Chemokine CXCL14; a double-edged sword in cancer development. Int Immunopharmacol. 2021;97:107681.

403. Ozawa S, Kato Y, Ito S, Komori R, Shiiki N, Tsukinoki K, et al. Restoration of BRAK / CXCL14 gene expression by gefitinib is associated with antitumor efficacy of the drug in head and neck squamous cell carcinoma. Cancer Sci. 2009;100(11):2202-9.

404. Gallo S, Ricciardi S, Manfrini N, Pesce E, Oliveto S, Calamita P, et al. RACK1 Specifically Regulates Translation through Its Binding to Ribosomes. Mol Cell Biol. 2018;38(23).

405. Huang T. Copy Number Variations in Tumors. In: Boffetta P, Hainaut P, editors. Encyclopedia of Cancer (Third Edition). Oxford: Academic Press; 2019. p. 444-51.

406. Santos GC, Zielenska M, Prasad M, Squire JA. Chromosome 6p amplification and cancer progression. J Clin Pathol. 2007;60(1):1-7.

407. Mohanty V, Wang F, Mills GB, Network CTDR, Chen K. Uncoupling of gene expression from copy number presents therapeutic opportunities in aneuploid cancers. Cell Rep Med. 2021;2(7):100349.

408. Mohanty V, Akmamedova O, Komurov K. Selective DNA methylation in cancers controls collateral damage induced by large structural variations. Oncotarget. 2017;8(42):71385-92.

409. Casas-Mollano JA, Zinselmeier MH, Erickson SE, Smanski MJ. CRISPR-Cas Activators for Engineering Gene Expression in Higher Eukaryotes. CRISPR J. 2020;3(5):350-64.

410. Nemudryi AA, Valetdinova KR, Medvedev SP, Zakian SM. TALEN and CRISPR/Cas Genome Editing Systems: Tools of Discovery. Acta Naturae. 2014;6(3):19-40.

411. Rao DD, Vorhies JS, Senzer N, Nemunaitis J. siRNA vs. shRNA: similarities and differences. Adv Drug Deliv Rev. 2009;61(9):746-59.

412. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013;23(12):2126-35.

413. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11(8):817-20.

414. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669-80.

415. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015;109:21 9 1- 9 9.

416. Xu W, Wen Y, Liang Y, Xu Q, Wang X, Jin W, et al. A plate-based single-cell ATAC-seq workflow for fast and robust profiling of chromatin accessibility. Nat Protoc. 2021;16(8):4084-107.

417. Shao X, Lv N, Liao J, Long J, Xue R, Ai N, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med Genet. 2019;20(1):175.

418. Bhattacharya A, Bense RD, Urzúa-Traslaviña CG, de Vries EGE, van Vugt MATM, Fehrmann RSN. Transcriptional effects of copy number alterations in a large set of human cancers. Nature Communications. 2020;11(1):715.

419. Harmanci AS, Harmanci AO, X Z. Inference of Clonal Copy Number Alterations from RNASequencing Data. J Cancer Immunol. 2020;2(3):66-8.

420. Piovesan A, Caracausi M, Antonaros F, Pelleri MC, Vitale L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. Database (Oxford). 2016;2016.

421. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17(3):175-88.

422. Blanco L, Bernad A, Lazaro JM, Martin G, Garmendia C, Salas M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. J Biol Chem. 1989;264(15):8935-40.

423. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, et al. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. Nucleic Acids Res. 2004;32(9):e71.

424. de Bourcy CF, De Vlaminck I, Kanbar JN, Wang J, Gawad C, Quake SR. A quantitative comparison of single-cell whole genome amplification methods. PLoS One. 2014;9(8):e105585.

425. Hård J, Mold JE, Eisfeldt J, Tellgren-Roth C, Häggqvist S, Bunikis I, et al. Long-read whole-genome analysis of human single cells. Nature Communications. 2023;14(1):5164.

426. Selleckchem. MK-2206 2HCl: Selleckchem; 2024 [Available from: https://www.selleckchem.com/products/MK-2206.html.

427. Selleckchem. Capivasertib (AZD5363): Selleckchem; 2024 [Available from: https://www.selleckchem.com/products/azd5363.html.

428. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168(4):613-28.

429. Gupta PB, Pastushenko I, Skibinski A, Blanpain C, Kuperwasser C. Phenotypic Plasticity: Driver of Cancer Initiation, Progression, and Therapy Resistance. Cell Stem Cell. 2019;24(1):65-78.

430. Qi L, Toyoda H, Xu DQ, Zhou Y, Sakurai N, Amano K, et al. PDK1-mTOR signaling pathway inhibitors reduce cell proliferation in MK2206 resistant neuroblastoma cells. Cancer Cell Int. 2015;15:91.

431. Stottrup C, Tsang T, Chin YR. Upregulation of AKT3 Confers Resistance to the AKT Inhibitor MK2206 in Breast Cancer. Mol Cancer Ther. 2016;15(8):1964-74.

432. Stratikopoulos EE, Dendy M, Szabolcs M, Khaykin AJ, Lefebvre C, Zhou MM, et al. Kinase and BET Inhibitors Together Clamp Inhibition of PI3K Signaling and Overcome Resistance to Therapy. Cancer Cell. 2015;27(6):837-51.

433. Tsang T, He Q, Cohen EB, Stottrup C, Lien EC, Zhang H, et al. Upregulation of Receptor Tyrosine Kinase Activity and Stemness as Resistance Mechanisms to Akt Inhibitors in Breast Cancer. Cancers (Basel). 2022;14(20).

434. Junankar S, Baker LA, Roden DL, Nair R, Elsworth B, Gallego-Ortega D, et al. ID4 controls mammary stem cells and marks breast cancers with a stem cell-like phenotype. Nat Commun. 2015;6:6548.

435. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009;324(5930):1029-33.

436. Barba I, Carrillo-Bosch L, Seoane J. Targeting the Warburg Effect in Cancer: Where Do We Stand? Int J Mol Sci. 2024;25(6).

437. de la Cruz-Lopez KG, Castro-Munoz LJ, Reyes-Hernandez DO, Garcia-Carranca A, Manzo-Merino J. Lactate in the Regulation of Tumor Microenvironment and Therapeutic Approaches. Front Oncol. 2019;9:1143.

438. Perez-Tomas R, Perez-Guillen I. Lactate in the Tumor Microenvironment: An Essential Molecule in Cancer Progression and Treatment. Cancers (Basel). 2020;12(11).

439. Walenta S, Chau TV, Schroeder T, Lehr HA, Kunz-Schughart LA, Fuerst A, et al. Metabolic classification of human rectal adenocarcinomas: a novel guideline for clinical oncologists? J Cancer Res Clin Oncol. 2003;129(6):321-6.

440. San-Millan I, Julian CG, Matarazzo C, Martinez J, Brooks GA. Is Lactate an Oncometabolite? Evidence Supporting a Role for Lactate in the Regulation of Transcriptional Activity of Cancer-Related Genes in MCF7 Breast Cancer Cells. Front Oncol. 2019;9:1536.

441. Liao M, Yao D, Wu L, Luo C, Wang Z, Zhang J, et al. Targeting the Warburg effect: A revisited perspective from molecular mechanisms to traditional and innovative therapeutic strategies in cancer. Acta Pharm Sin B. 2024;14(3):953-1008.

442. Lytle NK, Barber AG, Reya T. Stem cell fate in cancer growth, progression and therapy resistance. Nat Rev Cancer. 2018;18(11):669-80.

443. Emmink BL, Verheem A, Van Houdt WJ, Steller EJ, Govaert KM, Pham TV, et al. The secretome of colon cancer stem cells contains drug-metabolizing enzymes. J Proteomics. 2013;91:84-96.

444. Liu PP, Liao J, Tang ZJ, Wu WJ, Yang J, Zeng ZL, et al. Metabolic regulation of cancer cell side population by glucose through activation of the Akt pathway. Cell Death Differ. 2014;21(1):124-35.

445. Zhong X, He X, Wang Y, Hu Z, Huang H, Zhao S, et al. Warburg effect in colorectal cancer: the emerging roles in tumor microenvironment and therapeutic implications. J Hematol Oncol. 2022;15(1):160.

446. Gris-Oliver A, Palafox M, Monserrat L, Braso-Maristany F, Odena A, Sanchez-Guixe M, et al. Genetic Alterations in the PI3K/AKT Pathway and Baseline AKT Activity Define AKT Inhibitor Sensitivity in Breast Cancer Patient-derived Xenografts. Clin Cancer Res. 2020;26(14):3720-31.

447. Dunn S, Eberlein C, Yu J, Gris-Oliver A, Ong SH, Yelland U, et al. AKT-mTORC1 reactivation is the dominant resistance driver for PI3Kβ/AKT inhibitors in PTEN-null breast cancer and can be overcome by combining with Mcl-1 inhibitors. Oncogene. 2022;41(46):5046-60.

448. Jakubowski JM. Investigating Mechanisms of Acquired Resistance to the AKT Inhibitor Capivasertib (AZD5363): University of Kent; 2019.

449. Zhang Y, Rajput A, Jin N, Wang J. Mechanisms of Immunosuppression in Colorectal Cancer. Cancers (Basel). 2020;12(12).

450. Wang HB, Yao H, Li CS, Liang LX, Zhang Y, Chen YX, et al. Rise of PD-L1 expression during metastasis of colorectal cancer: Implications for immunotherapy. J Dig Dis. 2017;18(10):574-81.

451. Liu J, Zhang N, Li Q, Zhang W, Ke F, Leng Q, et al. Tumor-associated macrophages recruit CCR6+ regulatory T cells and promote the development of colorectal cancer via enhancing CCL20 production in mice. PLoS One. 2011;6(4):e19495.

452. Rodriguez PC, Quiceno DG, Zabaleta J, Ortiz B, Zea AH, Piazuelo MB, et al. Arginase I production in the tumor microenvironment by mature myeloid cells inhibits T-cell receptor expression and antigen-specific T-cell responses. Cancer Res. 2004;64(16):5839-49.

453. Brandacher G, Perathoner A, Ladurner R, Schneeberger S, Obrist P, Winkler C, et al. Prognostic value of indoleamine 2,3-dioxygenase expression in colorectal cancer: effect on tumor-infiltrating T cells. Clin Cancer Res. 2006;12(4):1144-51.

454. Uyttenhove C, Pilotte L, Theate I, Stroobant V, Colau D, Parmentier N, et al. Evidence for a tumoral immune resistance mechanism based on tryptophan degradation by indoleamine 2,3-dioxygenase. Nat Med. 2003;9(10):1269-74.

455. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med. 2015;372(26):2509-20.

456. Chen EX, Jonker DJ, Kennecke HF, Berry SR, Couture F, Ahmad CE, et al. CCTG CO.26 trial: A phase II randomized study of durvalumab (D) plus tremelimumab (T) and best supportive care (BSC) versus BSC alone in patients (pts) with advanced refractory colorectal carcinoma (rCRC). Journal of Clinical Oncology. 2019;37(4_suppl):481-.

457. Argilés G, Saro J, Segal NH, Melero I, Ros W, Marabelle A, et al. Novel carcinoembryonic antigen T-cell bispecific (CEA-TCB) antibody: Preliminary clinical data as a single agent and in combination with atezolizumab in patients with metastatic colorectal cancer (mCRC). Annals of Oncology. 2017;28:iii151.

458. Zhen YH, Liu XH, Yang Y, Li B, Tang JL, Zeng QX, et al. Phase I/II study of adjuvant immunotherapy with sentinel lymph node T lymphocytes in patients with colorectal cancer. Cancer Immunol Immunother. 2015;64(9):1083-93.

459. Whelan SA. Generation and investigation of resistance mechanisms to AZD5363 in breast cancer: University of Kent; 2017.

460. Zhu Z, Wang W, Lin F, Jordan T, Li G, Silverman S, et al. Genome profiles of pathologist-defined cell clusters by multiregional LCM and G&T-seq in one triple-negative breast cancer patient. Cell Rep Med. 2021;2(10):100404.

Trends in Genetics

CellPress OPEN ACCESS

Review

Into the multiverse: advances in single-cell multiomic profiling

Silvia Ogbeide,[1,3] Francesca Giannese,[2,3] Laura Mincarelli,[1] and Iain C. Macaulay[1,*]

Single-cell transcriptomic approaches have revolutionised the study of complex biological systems, with the routine measurement of gene expression in thousands of cells enabling construction of whole-organism cell atlases. However, the transcriptome is just one layer amongst many that coordinate to define cell type and state and, ultimately, function. In parallel with the widespread uptake of single-cell RNA-seq (scRNA-seq), there has been a rapid emergence of methods that enable multiomic profiling of individual cells, enabling parallel measurement of intercellular heterogeneity in the genome, epigenome, transcriptome, and proteomes. Linking measurements from each of these layers has the potential to reveal regulatory and functional mechanisms underlying cell behaviour in healthy development and disease.

Highlights

To understand intercellular heterogeneity within an organism, it is essential to make coordinated measurements linking the genome and its epigenetic regulation to gene and protein expression at the single-cell level.

Rapid advances in single-cell multiomics approaches have enabled analysis of multiple molecular modalities from the same single cell.

Methods incorporating several modalities now exist, although several challenges remain with regard to resolution, data integration, and scale.

Further developments in multiomics approaches will provide unique insights into the regulatory processes governing how individual cells function collectively to produce whole-organism phenotypes in development, health, and disease.

The many sources of cellular heterogeneity

As fundamental biological units, cells within a multicellular organism are capable of remarkable diversity in form and function throughout development and disease. Individual cells have typically been classified to particular 'types' or 'states' by phenotypic measurement, such as marker gene expression, morphology, or function. scRNA-seq has been instrumental in revealing a broader scheme for cell type classification through simultaneous measurement of the expression of thousands of genes in thousands – even millions – of cells, and therefore more detailed classification of cell types, subtypes, and states in dynamic and complex developmental systems [1–7]. These rapid advances in scRNA-seq technologies have made whole-organism single-cell profiling a reality, underpinning the efforts of major consortia aiming to produce a comprehensive map of cell types in the human body [8].

However, a cell's transcriptome is just one aspect of its phenotype – an incomplete representation of cellular identity, reflecting both the regulatory status of the genome and implied protein production. Cell-type-specific mRNA expression is governed by epigenetic mechanisms and, in general, only has functional potential when translated into protein. Thus, **molecular cellular identity** (see Glossary) is a product of the interplay between many different modalities within the cell (Figure 1, Key figure), all of which can vary as a result of intrinsic and extrinsic factors. To truly understand how individual cells within a multicellular organism can demonstrate such remarkable heterogeneity, it is essential to be able to make coordinated measurements linking the genome and its epigenetic regulation to gene products (transcripts and proteins).

In parallel with the rapid and widespread adoption of scRNA-seq, there has been an adaptive radiation of single-cell multiomics approaches for the simultaneous analysis of multiple molecular modalities from the same single cell (Figure 2). These powerful approaches enable associations to be made between genome sequence, structure, and regulatory state and the transcriptional and proteomic phenotype of the cell. While a cell can be classified on the basis of any one of these measurements, a cell's identity can only be understood through the integration of these different

[1]Earlham Institute, Norwich Research Park, Norwich, UK
[2]Center for Omics Sciences, IRCCS San Raffaele Institute, Milano, Italy
[3]These authors contributed equally.

*Correspondence:
iain.macaulay@earlham.ac.uk
(I.C. Macaulay).

layers. This kind of analysis can not only enhance our ability to classify cell identity but brings us closer to being able to perform mechanistic, functional genomic studies of individual cells within a population. This has a particular impact on the study of development, ageing, and disease, where heterogeneity at multiple levels can contribute to cellular phenotypes which have profound impact on the organismal phenotype.

Linking somatic variation and gene expression

Within the lifetime of an organism, genomic diversification between cells – known as **somatic variation** – can occur as a result of programmed and spontaneous mechanisms. Thus, the genomes of individual cells within a multicellular organism can have substantial and significant deviations from the 'prime' genome – that of the fertilised zygote. For example, programmed somatic variation occurs in B and T lymphocytes to produce diversity in specificity of antibody and T cell receptors. Spontaneous somatic variation, where individual cells acquire genomic diversity – from **single-nucleotide variants (SNVs)** to whole-chromosome **copy-number variants (CNVs)** – is common in normal mammalian development and ageing [9,10]. This phenomenon can become pathogenic when a particular variant (or set of variants) acquired in a single cell confers a competitive advantage to the cell and its subsequent progeny. This cellular evolutionary process, where genotypic changes create competitive phenotypic heterogeneity, can lead to clonal expansion and the formation of malignant or cancerous clones through the acquisition of further mutations and genomic rearrangements [11].

Changes in the genome itself have limited impact unless they modify the sequence of genes or their regulatory elements, thereby modifying gene expression and the overall phenotype of the cell. Therefore, linking somatic variation to gene expression in the same cell is critical to understand the functional consequences of acquired mutations and how these can introduce functional cellular heterogeneity. Early single-cell multiomics methods, such as DR-seq (gDNA and mRNA-sequencing) [12] and G&T-seq (genome and transcriptome-sequencing) [13], performed parallel analysis of genomes and transcriptomes of individual cells, typically isolated manually or by **FACS**. In DR-seq, combined amplification of a single cell's genome and transcriptome is performed in a single reaction, while in G&T-seq, mRNA is physically separated from genomic DNA before parallel amplification of both (Figure 3A). Both **plate-based** methods enabled links to be made between genomic variation – from chromosomal copy number down to single-nucleotide resolution – and gene expression. They also demonstrated, for the first time, the direct impact of chromosomal copy number on gene expression in the same cell, with a clear correlation between copy number and gene expression. In the case of G&T-seq it was possible to demonstrate this correlation immediately after the cell cycle in which reciprocal chromosomal gains or losses occurred. Additionally, the combination of full-length RNA-seq and whole-genome sequencing in G&T-seq enabled parallel detection of a fusion transcript and the causative genomic rearrangement in the same cell of a breast cancer cell line. Both of these early methods demonstrated potential for single-cell multiomic studies in cancer (Box 1) in which the transcriptional phenotype of the cell can be associated with evolutionary events recorded in the genome.

These approaches were not without limitations, suffering from sequence errors introduced in the **whole-genome amplification** processes, as well as allelic and locus dropout that is inherent in single-cell genome amplification. Gaining high coverage data from the entire genomes of single cells, in parallel with rich transcriptomic data, is also expensive, which limits reasonable throughput to 100s or 1000s of cells.

More recently, Target-seq [14] was developed to enable parallel mRNA-seq and targeted genotyping, rather than whole-genome sequencing, of the same single cell. The plate-based

Glossary

ATAC-seq: an 'assay for transposase-accessible chromatin with high-throughput sequencing' in which DNA from accessible chromatin is selectively sequenced. This gives an overview of the 'openness' of the chromatin across the genome, probed by hyperactive Tn5 transposase.
Chromatin velocity: a trajectory of cell lineage commitment based on the measurement of changes in euchromatin and heterochromatin in thousands of cells, as measured by the GET-seq assay.
Combinatorial indexing: methods which use serial barcoding of pools of nuclei or cells to generate highly complex combinations of barcodes attached to individual molecules (DNA or RNA), and thus increasing throughput without the need for dedicated microfluidics platforms. These methods are often appropriate for experiments where large number of cells (>1000s) undergo parallel analysis and classification.
Copy-number variant (CNV): an increase or decrease in the number of copies of a region of the genome, ranging from increased numbers of short tandem repeats through to whole chromosome gains and losses.
CUT&Tag: 'cleavage under targets and tagmentation', a method which uses antibody-tethered transposases to target specific DNA–protein interactions for sequencing, including histone modifications and transcription factors.
DNA methylation: in mammals, this is an epigenetic mechanism involving the transfer of a methyl group onto cytosine bases in the genome which can have a regulatory impact on gene expression. This is typically measured using bisulfite sequencing, in which unmethylated cytosines are converted to uracil – which will appear as a thymine base in sequencing data – while methylated cytosines remain unchanged.
Epigenetic plasticity: variability in epigenetic regulation that permits cells to undergo cell fate transitions due to stochastic activation of gene expression.
Euchromatin: loosely packed or 'open' chromatin, which is often the site of active gene expression.
FACS: fluorescence-activated cell sorting, a method for the sorting of single cells based on phenotypic measurements, including size, granularity, and protein/antigen expression.

protocol features an optimised version of the Smart-seq2 mRNA amplification, after which the sample is split and primers targeting regions of interest within the transcriptome and/or genome are used to generate targeted amplicon sequencing libraries. By focussing on known mutations, this approach significantly increases the sensitivity and reduces the cost of mutation detection. A related, microfluidic targeted genome sequencing approach has been commercialised by Mission Bio, enabling high-throughput genotyping of single cells, but linking with protein expression information rather than transcriptomic data. These targeted methods are highly relevant for studies where a known repertoire of mutations is prevalent, such as studies of intratumoural heterogeneity and cancer evolution, where recurrent mutations are common. However, in complex mutational backgrounds, or where mutation discovery is important, bulk or single-cell whole-genome sequencing may still be more applicable.

Methods involving physical separation of the nucleus and cytoplasm of a cell have also been demonstrated (Figure 3B). 'Simultaneous isolation of genomic DNA and total RNA' (SIDR) [15] was the first such method. This approach has seen massive increases in throughput in direct nuclear tagmentation and RNA sequencing (DNTR-seq) [16], which relies on nuclear/cytoplasmic separation, followed by full-length mRNA amplification from the cytoplasmic fraction, and direct tagmentation-based genomic library preparation from the nuclear DNA, obviating the need for a traditional whole-genome amplification step. This represents a significant cost reduction and contributes to the increased scale at which the method can operate. However, like other methods which require physical separation of nucleus and cytoplasm, it is unclear how they are affected by disassembly of the nuclear envelope during the mitotic cell cycle.

Linking the epigenome and gene expression

Although intercellular diversity in genome sequence and structure is common, the phenotypic heterogeneity of cells is a hallmark of multicellular organisms and emerges from the regulation of gene expression through epigenetic modification of the genome. Starting from the same genetic background, cells can acquire highly specialised functions during development and are able to dynamically modify their phenotype in response to environmental stimuli. Many epigenomic approaches have been adapted to make measurements in single cells, but only assays for **DNA methylation** and chromatin accessibility have been incorporated into multiomic assays. These assays, by linking genome regulation and gene expression in the same cell, can shed light on lineage determination, developmental dynamics (Box 2), and mechanisms of disease development.

The first methods that attempted to link epigenetic diversity with transcriptional heterogeneity in single cells focussed on the association between DNA methylation at CpG sites and gene expression. To achieve this, single-cell bisulfite sequencing methods – either post-bisulfite adaptor tagging (PBAT) [17], which measures DNA methylation across the genome, or reduced representation bisulfite sequencing (RRBS) [18], which enriches for regions with high CpG content – have been combined with transcriptomic analysis of individual cells. scM&T-seq (single-cell methylome and transcriptome sequencing) built upon the G&T-seq method (Figure 3A), but instead uses the purified genomic DNA for a modified PBAT protocol, generating genome-wide methylation data, while the transcriptome is again sequenced using a modified Smart-seq2 protocol [19]. The method was first applied to mouse embryonic stem cells to discover novel correlations between heterogeneity at DNA methylation of distal regulatory elements and expression of hundreds of genes, including key pluripotency genes. The G&T-seq approach was also adapted for Smart-RRBS, which enables joint profiling of DNA methylation (by RRBS) and transcriptome analysis [20]. Other approaches involving physical separation of the nucleus and the cytoplasm have been used to obtain gene expression and

Heterochromatin: tightly packed or 'closed' chromatin, which is less accessible for transcription.
Hi-C: a chromosome conformation capture assay which enables the genome-wide measurement of long-range interactions between genomic loci.
Microfluidic assays: in this case referring to assays which isolate individual cells in microfluidic droplets in the presence of barcoded oligonucleotide-coated beads to enable the capture and barcoding of molecules of multiple classes (DNA, RNA, and protein) from single cells. These methods are often appropriate for experiments in which a large number of cells (>1000s) undergo parallel analysis and classification.
Molecular cellular identity: the amalgamation of molecular events that make a cell belong to a particular type or state.
Plate-based assay: in this case refers to a single-cell multiomic approach for which cells are isolated into 96- or 384-well plates for processing (typically) using liquid handling robotics. These methods are often appropriate for experiments where small numbers (100s–1000s) of cells undergo a detailed analysis.
Single-nucleotide variant (SNV): a single base change in the genome.
Somatic variation: genetic diversity occurring between cells within the same organism, arising from mutations occurring after conception.
Whole-genome amplification: describes several possible methods for genome-wide amplification of cellular DNA, in this case to enable single-cell genome sequencing.

Key figure

Multiomic exploration of molecular cell identity

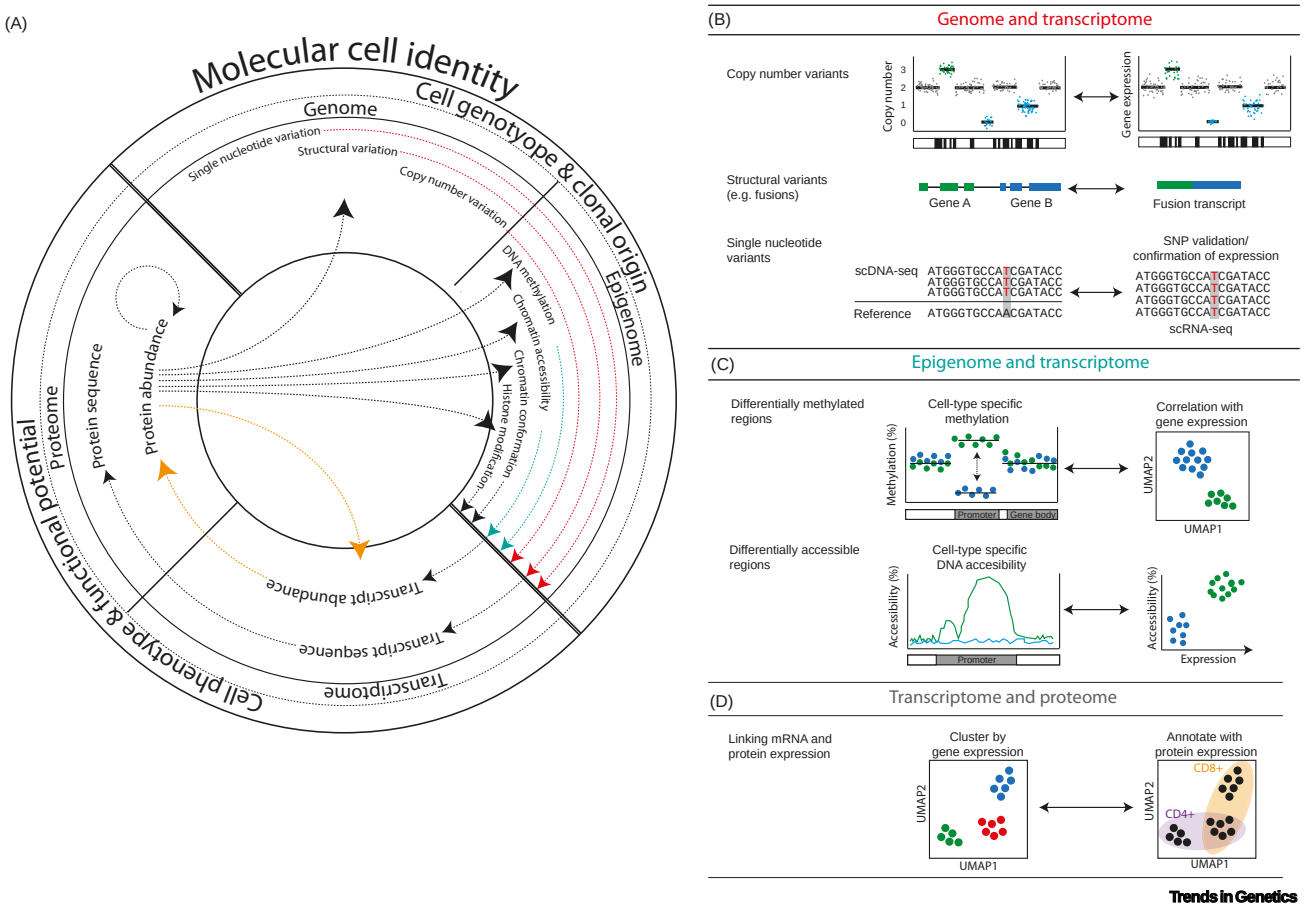


Figure 1. (A) Molecular cell identity comprises the interaction of many different molecular layers within the cell. Genomic and epigenomic variation influence the sequence and abundance of transcripts and proteins, which in turn can influence each molecular layer within the cell. Areas in which single-cell multiomic analysis has made significant advances are highlighted in (B–D).

DNA methylation data from the same single cell, including MT-seq [21] and scTRIO-seq [22] (Figure 3B).

There are still major limitations to the detection of DNA methylation in single cells – bisulfite treatment is destructive to the DNA, resulting in a high level of allelic and locus dropout. Similarly, the sequencing libraries generated using these approaches are typically rich in PCR duplicates, which, combined with dropouts and the expense of pursuing high genomic coverage from single cells, make the measurement of DNA methylation at single-base resolution challenging. Furthermore, the C > T substitution inherent in the approach makes the calling of genomic variants difficult, making existing approaches unsuitable for parallel methylation and SNV calling. Recently, the epi-gSCAR approach (epigenomics and genomics of single cells analysed by restriction) demonstrated the feasibility of bisulfite-free single-cell library preparation – using the methylation-sensitive restriction

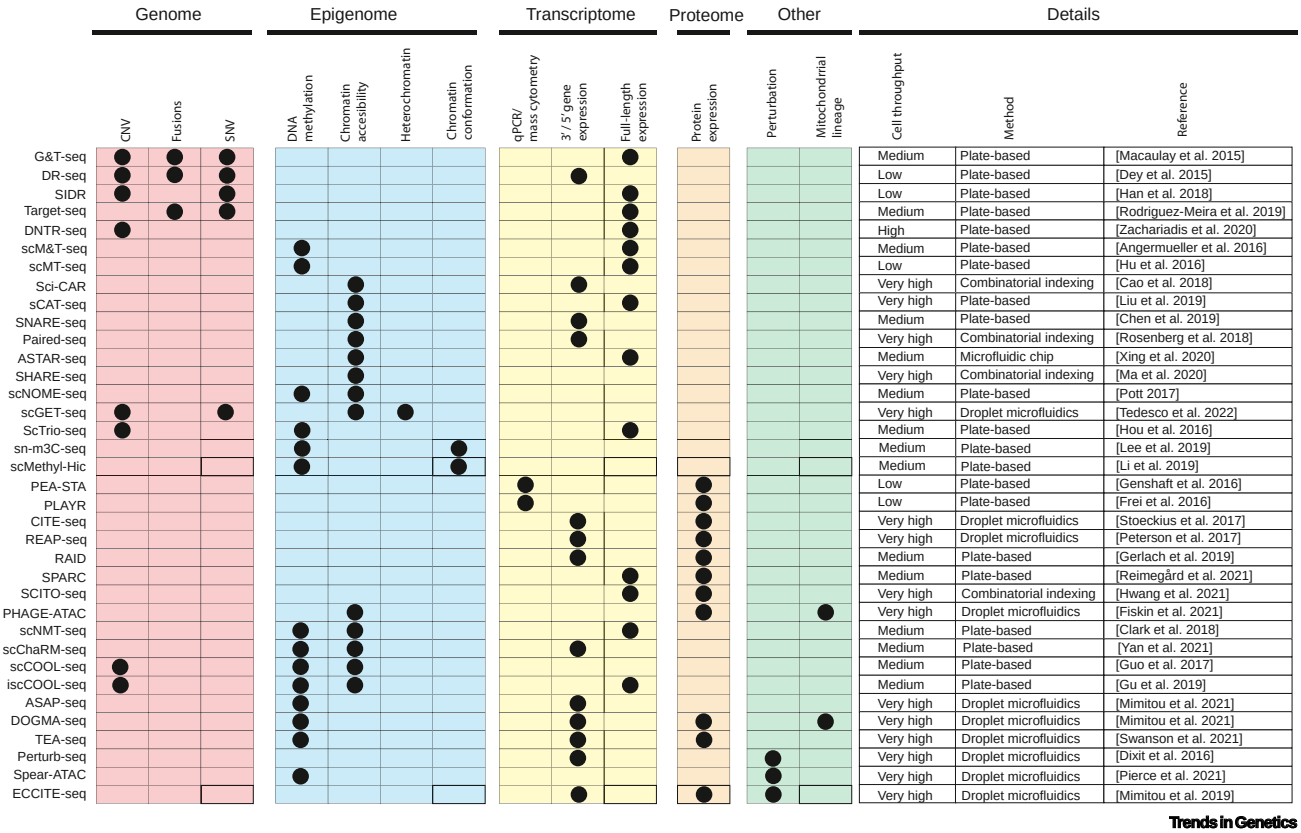


Figure 2. An overview of current single-cell multiomic approaches. See [12,13,15,16,19,21,22,24–26,28,29,36,41,45–50,53–56,58,59,61,63–65,102].

enzyme *HhaI* and quasilinear amplification – to study genome-wide methylation and genomic variation at single-nucleotide resolution in cancer cell lines [23].

The accessibility of sequences within the genome is considered to be a mark of genomic activity, representing the expression of particular genes or the openness of particular sequences, including enhancers or transcription factor binding sites. Chromatin accessibility in single cells is now routinely measured by the 'assay for transposase-accessible chromatin' using sequencing (**ATAC-seq**), in which the Tn5 transposase is used to fragment and insert sequencing adaptors into open regions of the genome (**euchromatin**). Due to the nature of the ATAC-seq method, it is compatible with considerably higher throughput than the analysis of DNA methylation. In general, these high-throughput methods rely on the tagmentation of accessible chromatin in a bulk preparation of nuclei before paired barcoding of the tagmented DNA and RNA from the same cell, either through **combinatorial indexing** or in droplet-based approaches (Figure 3C,D). For example, sciCAR-seq [24], used combinatorial indexing to process over 11 000 nuclei per experiment. Medium-throughput methods, working with intact cells rather than nuclei, have also been described (scCAT-seq, [25] and ASTAR-seq [26]) and may potentially be more applicable to experiments in which rare cells are to be profiled.

Throughput was dramatically increased in Paired-seq [27] by implementation of a ligation-based combinatorial indexing strategy which enabled processing of one million nuclei per experiment. Building on Paired-seq and a similar approach, SPLiT-seq [28] and SHARE-seq [29], further increased the sensitivity of the combinatorial indexing approach to measure the 'chromatin

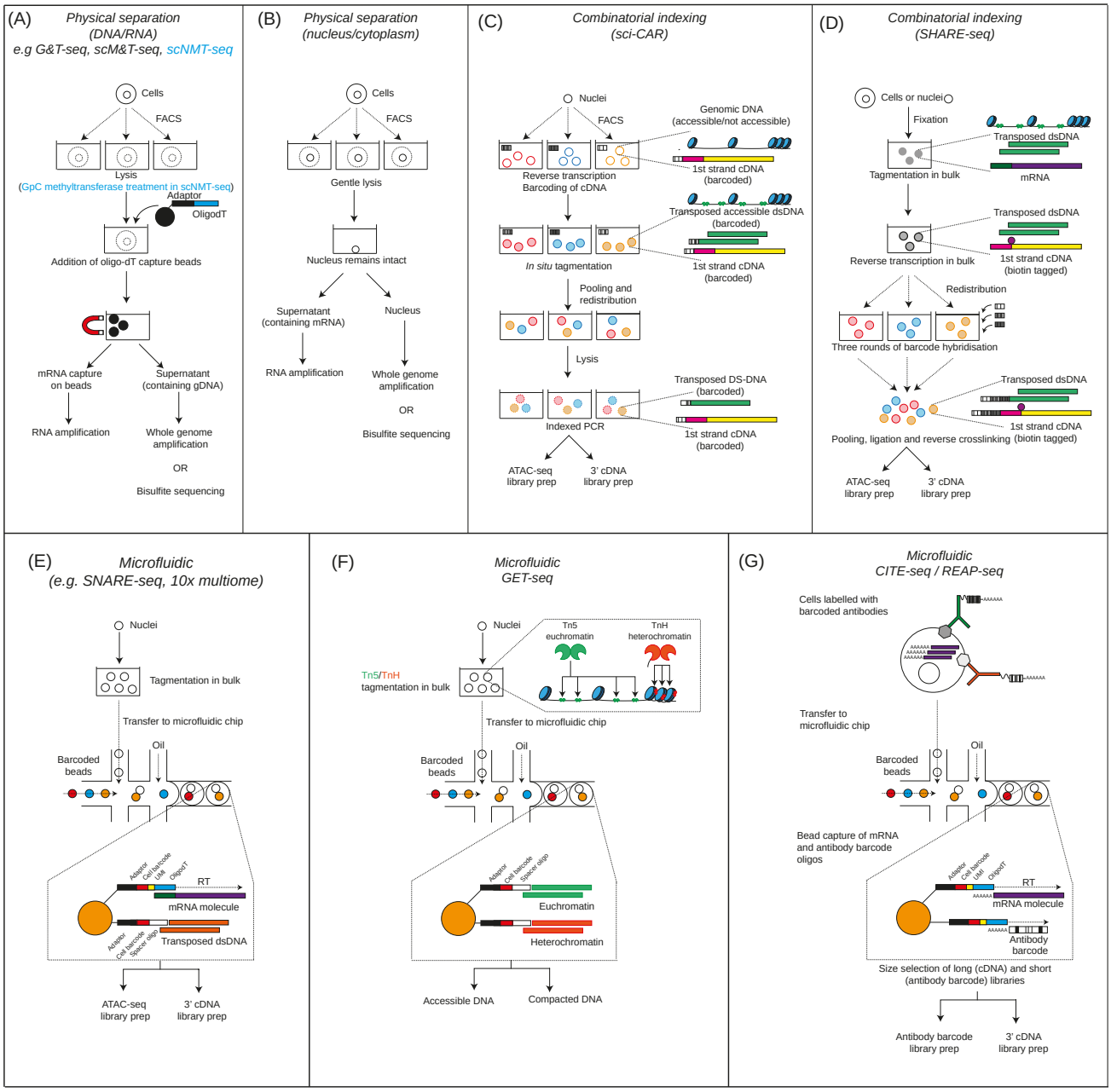


Figure 3. Capturing multiple layers of information from the same single cell. Various approaches have emerged to extract distinct layers of omic information from the same single cell. (A) G&T-seq, and those methods based on it, perform physical separation of genomic DNA and mRNA following capture on magnetic beads. (B) An alternative approach involves physical separation of the nucleus and cytoplasm of the cell. These methods allow both genome sequencing and methylation sequencing to be performed on the isolated DNA. High-throughput combinatorial indexing has been applied in (C) sci-CAR and (D) SHARE-seq to obtain linked transcriptome and chromatin accessibility from the same cell, while droplet-based microfluidic approaches (E) have enabled parallel capture of these modalities using SNARE-seq and the 10X Genomics Chromium platform. (F) Droplet microfluidics has also been used to sequence DNA from accessible and compacted chromatin using GET-seq. (G) CITE-seq and REAP-seq take advantage of polyadenylated oligonucleotide tags attached to antigen-specific antibodies to capture protein expression information in parallel with mRNA expression. Abbreviations: ATAC-seq, 'assay for transposase-accessible chromatin with high-throughput sequencing'; CITE-seq, cellular indexing of transcriptomes and epitopes by sequencing; FACS, fluorescence-activated cell sorting; G&T-seq, genome and transcriptomesequencing; GET-seq, genome and epigenome by transposases sequencing; sci-CAR, single-cell combinatorial indexing-chromatin accessibility and RNA sequencing; SNARE-seq, single-nucleus chromatin accessibility and mRNA expression sequencing.

Box 1. Single-cell multiomics in cancer evolution

Cancer development within an individual is an evolutionary process in which cells evolve by mutation and subsequent selection for clones with increased proliferative capability, fitness, and resistance to therapeutic intervention. While mutational profiling is informative in understanding tumour evolution, it is critical to link these mutations or genotypes with cellular phenotype or functional data. Single-cell multiomic approaches enable the integration of genotypic information – from single nucleotide to whole chromosome resolution – with epigenomic, transcriptomic, and increasingly, proteomic information, and several landmark studies have demonstrated the application of these approaches in cancer.

The scTrio-seq approach [22] was optimised to profile human colorectal cancer cells from paired primary tumours and lymphatic or liver metastases [72]. In one patient, this permitted the identification of 12 sublineages that originated from two different progenitors, one of which was maintained throughout tumour progression and was still present in both the final neoplasm and distant metastases. Overlaying DNA methylation data along these lineages revealed that methylation levels were homogeneous among cells within the same genetic lineage but varied among different lineages.

Based on the physical separation of genomic DNA and transcriptomes, Smart-RRBS has been applied in the study of epigenetic evolution in chronic lymphocytic leukaemia (CLL) [73]. Here, the parallel measurement of epigenetic and transcriptional changes enabled the linkage of epimutations in *SF3B1*-mutant CLL cells to a 3′ splicing phenotype and subclones of cells with epigenetic and transcriptional phenotypes that expanded following chemotherapeutic treatment. Subsequently, the same approach demonstrated that a decrease in epigenetic–transcriptional coordination in CLL could partially be explained by intercellular epigenetic diversification [74]. The Smart-RRBS approach was also recently applied in the study of primary diffuse glioma, where it enabled joint capture of transcriptional, genetic, and epigenetic data from the same single cell [75]. The scRRBS enabled CNV analysis at 20 Mb resolution for genome-wide analysis, but also 0.1 Mb resolution to reveal CNVs at the *EGFR* locus. Furthermore, it could be used to generate lineage trees from individual glioblastoma samples, with individual branches annotated with transcriptomic cell types and states.

Recently G&T-seq [13] was coupled with laser capture microdissection (LCM) to generate spatially resolved genomic and transcriptional profiles of cancer cells with the potential for lymphovascular invasion in a patient with triple-negative breast cancer [76].

While these studies are still relatively small in scale, continued development of these methods, including increases in throughput and resolution, reductions in cost, and incorporation of additional layers of data, will undoubtedly transform single-cell multiomic profiling into a mainstream tool in the study of cancer evolution.

Box 2. Single-cell multiomics analysis in developmental systems

In multicellular organisms, cells can adapt an immense array of phenotypes and states, despite having the same or highly similar genomes. During development and a healthy lifespan, as cells commit first to specific germ layers then cell types, the regulation of genome function through epigenetic modification is fundamental to the emergence of this complexity. Cell fate decisions are made by individual cells responding to intrinsic and extrinsic factors resulting in changes to epigenomic, transcriptomic, and proteomic aspects of cell identity. The integration of different omic layers of the same single cell through multiomic analysis can provide a unique perspective on the dynamics of these processes.

During early embryogenesis, global demethylation erases the epigenetic signatures of the highly specialised gametes to enable the embryonic cells to become totipotent. scCOOL-seq, which measures DNA methylation, chromatin accessibility, and copy number variation has been applied to study this epigenetic reprogramming in mouse [57] and human embryos [77], revealing the dynamics of parental genome activity in the first cell divisions after fertilisation. The iscCOOL-seq method was subsequently applied to the study of mouse oocytes, identifying dynamic associations between chromatin accessibility, methylation, and expression during oocyte maturation. Similarly, scM&T-seq was also used to explore the heterogeneity in DNA methylation of oocytes from young and aged mice, with those from aged mice showing increased molecular heterogeneity indicative of epigenetic dysregulation [78].

Combined gene expression and whole-genome methylation profiling was used to characterise the post-implantation DNA methylation landscapes in mouse embryos (from eight-cell stage to E6.5 epiblast and extraembryonic ectoderm), revealing divergent methylation patterns in the extraembryonic tissue, with methylation in these lineages mirroring the aberrant methylation of the promoters of developmental genes observed in tumorigenesis [79].

The emergence of high-throughput approaches, combining ATAC-seq and RNA-seq from the same cell, with the potential to integrate protein expression and cell-lineage tracing (e.g., DOGMA-seq), has enormous potential in unravelling the dynamic interactions underpinning molecular cell identity during early development and organogenesis as well as in cellular systems undergoing constant replenishment (e.g., blood).

potential' of individual cells, which explores the predictive power of chromatin accessibility on future gene expression changes and lineage commitment within the cell. Single-nucleus chromatin accessibility and mRNA expression sequencing (SNARE-seq) leverages the **microfluidic** Drop-seq [30] method to perform parallel chromatin accessibility and gene expression measurements on the same nuclei [31]. This approach has been adapted for the 10X Genomics Chromium platform using hydrogel beads carrying separate capture oligonucleotides which capture both the tagmented genome and mRNA. A recent alternative method for single-nucleus multiomic profiling, ISSAAC-seq [32], based on the Sequencing HEteRo RNA-DNA-hYbrid (SHERRY) approach [33], exploits a first tagmentation reaction on accessible chromatin followed by reverse transcription and then a second tagmentation round on DNA–RNA hybrids. Nuclei are loaded on microfluidic or FACS apparatus for single-cell analysis, and separate DNA and RNA libraries are produced, exploiting the different adaptor configuration for the two tagmentation steps.

Linking different aspects of the epigenome

ATAC-seq will only provide sequence information from accessible chromatin and does not capture genetic alterations and chromatin remodelling events associated with **heterochromatin**. Compacted chromatin is crucial for lineage specification [34] and genome stability [35]. Recently, chromatin accessibility profiling has been combined with heterochromatin sampling in the single-cell genome and epigenome by transposases sequencing (scGET-seq) assay [36]. This assay builds on scATAC-seq with droplet microfluidic exploiting engineered transposases to simultaneously probe H3K9me3-enriched compacted chromatin alongside accessible chromatin. This combined epigenomic and genetic characterization allowed for increased resolution in CNV calling and it was used to compute a new metric called **chromatin velocity** – based on the differential enrichment between closed and open chromatin – to reveal patterns of **epigenetic plasticity** during stem cell reprogramming and key transcription factors correlated to developmental commitment. The introduction of engineered transposases in single-cell genomics unlocks immense potential for targeted analysis of other domains within the epigenome. A similar method named scCUT&Tag2for1, a modification of standard **CUT&Tag** [37], uses antibody-guided tagmentation to simultaneously characterise accessible and silenced regulome by targeting the initiation form of RNA polymerase II (Pol2 Serine-5 phosphate) and repressive Polycomb domains (H3K27me3) [38]. A further CUT&Tag development, scCUT&Tag-pro, allows simultaneous profiling of histone modifications with protein abundances on whole cells [39].

Chromatin conformation assays, such as **Hi-C**, have revealed the extent to which three-dimensional conformation of the genome regulates gene expression in health, disease, and senescence. Two multiomic approaches, single-nucleus methyl-3C [40] and scMethyl-HiC [41], have described methods to obtain linked chromatin conformation and methylation data from the same single cell, using bisulfite conversion of crosslinked genomic DNA. These approaches reveal that chromatin conformation alone can identify cell types within heterogeneous populations and differential methylation signatures associated with cell-type-specific chromatin interactions in human brain cells.

Linking transcript and protein expression

Much of cell behaviour is determined by the functions of proteins, and it is generally accepted that mRNA expression levels offer only a weak proxy for direct measurement of protein expression [42]. The obvious biochemical differences between nucleic acids and protein constitute a challenge for developing single-cell approaches – there is no method for protein sequence amplification and so measurements are dependent on antibody-based protein detection or mass spectrometry for peptide identification.

Proximity extension assays (PEAs) have been exploited to detect protein expression using antibodies recognising different epitopes on the same protein. PEA is based on proximity ligation assay (PLA) [43] in which antibodies conjugated with single-stranded DNA oligonucleotides colocalise on the target protein, enabling ligation and generation of a sequence that is detectable, in parallel with mRNA molecules, by qPCR [44,45]. Proximity ligation assay for RNA (PLAYR) expanded the throughput of the PLA approach by detecting transcripts and proteins using mass cytometry, enabling parallel measurement of over 40 different transcripts and protein epitopes in thousands of cells [46]. More recently, the 'single-cell protein and RNA coprofiling' (SPARC) method, in which mRNA and protein lysate are physically separated, enables parallel whole transcriptome mRNA-seq and detection of extracellular and intracellular proteins using PEA [47].

Increases in throughput have been enabled by the combination of oligonucleotide-conjugated antibodies with droplet-based microfluidic (e.g., 10X Genomics) and micro-well platforms (e.g., BD Rhapsody). This approach was pioneered in RNA expression and protein sequencing (REAP-seq) [48], Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) [49]. In these methods, cells are labelled with panels of antibodies, each tagged with a specific polyadenylated barcode which can be captured in parallel with the mRNA from the same cell following lysis (Figure 3G). SCITO-seq demonstrated a combinatorial indexing approach for antibody barcoding, enabling extreme multiplexing of cells as well as multimodal profiling of more than 150 surface proteins in parallel with mRNA expression from the same cells [50].

Antibody-based methods are severely limited by the availability of antigen-specific reagents – detection requires a reliable epitope-specific antibody (or pair of antibodies for PEA-based assays) which dramatically reduces the number of proteins or epitopes that can be surveyed. To obtain a more complete overview of the cellular proteome, antibody-independent methods are essential. Single-cell mass spectrometry-based approaches, such as single cell proteomics by mass spectrometry (SCoPE-MS) [51] and SCoPE2 [52], can analyse thousands of proteins and post-translational modifications in individual cells; however, they have yet to be directly incorporated into a combined multiomics approach. Recently, the PHAGE-ATAC assay [53] demonstrated an alternative approach where antibodies are replaced with nanobody phage-display libraries. This may offer a potential route towards protein detection without the need for antibodies.

Triple and higher-order single-cell multiomics

To fully explore the causes and consequences of intercellular heterogeneity, it is important to simultaneously capture data from as many aspects of the cell as possible. As an early example, scTRIO-seq could simultaneously measure genomic copy number changes at ~10 Mb resolution, DNA methylation, and gene expression from the same cell [22]. The single-cell nucleosome, methylation and transcription sequencing (scNMT-seq) approach [54] combines 'single-cell nucleosome occupancy' and methylome sequencing (scNOMe-Seq [55]) with a modification of G&T-seq [13] in a plate-based assay. In this approach, the genomic DNA is methylated, using a GpC methyltransferase, at GpC sites that are not bound by nucleosomes. Following physical separation of DNA and mRNA, the DNA undergoes bisulfite conversion which allows parallel measurement of nucleosome positioning, DNA methylation, and the cell's transcriptome [54]. A similar approach, scChaRM-seq, was also recently described [56]. NOMe-seq approaches were further adapted for scCOOL-seq [57], which can measure various genomic aspects of the cell in parallel, including chromatin state, nucleosome positioning, DNA methylation, CNV, and ploidy. This method has been modified to incorporate transcriptomic measurements in iscCOOL-seq [58].

Building on microfluidic workflows for parallel ATAC- and CITE-seq from single cells, ASAP-seq [59] demonstrated parallel chromatin accessibility, cell-surface and intracellular protein

measurements. Furthermore, the method enabled mutational profiling of the mitochondrial genome, allowing simultaneous lineage inference from mitochondria mutations, as previously demonstrated with mtscATAC-seq [60]. This was further expanded in the same manuscript to incorporate RNA-seq measurements – thus reading four layers of information from the same cell – in a method referred to as DOGMA-seq [59]. A similar approach, TEA-seq, was also recently described [61]. In these studies, whole cells were analysed instead of nuclei, which has the advantage of allowing more comprehensive phenotypic characterization, surface-marker enrichment prior to analysis, and retention of cytoplasmic RNA in multimodal assays.

By capturing these multiple layers, the epigenetic determinants of differentiation, and their dynamics, can be dissected with unprecedented detail – variations in accessibility and methylation can be directly correlated with variation in gene and protein expression levels. This will enable the construction of genome-wide regulatory models which incorporate the cell as the unit in which genomes are regulated and genes are expressed.

Concluding remarks

The emergence of methods enabling multiomic profiling of single cells continues at a staggering pace. It is now possible to profile multiple molecular layers of thousands of individual cells, with newer methods approaching 'Omni-seq' – where multiple omic measurements can be combined with spatial and lineage-based information to determine a cell's molecular state, microenvironment, and life-history in a single readout [62]. This has significant implications for current and future studies of developmental and cancer biology, where changes in individual cells are fundamental to the progression of healthy development or disease. These methods, especially when coupled with perturbations using the CRISPR/Cas system [63–65], will have immense potential to unravel cellular (epi)genotype/phenotype associations and the mechanisms that govern the emergence of cellular heterogeneity.

However, several challenges remain. At present, the analysis of each molecular level is imperfect – single-cell measurements of any kind are prone to noise and, in particular, drop-out, where critical signals of mutation, modification, or expression may be lost. As methods scale to incorporate thousands, even millions, of cells, there is a concomitant loss of detail per cell (see Outstanding questions). While the future development of methods will undoubtedly see the incorporation of further omics measurements – including expanded proteomic and metabolomic profiling [66] – there is still a need to refine many of the existing methods to obtain high resolution, accurate measurements of base-level events in the genome, and sensitive, quantitative, measurements of both gene and isoform expression from individual cells.

Aside from the macromolecular components of the cell, there are also many metabolites that can be instrumental in the regulation of cell function, and new approaches for their measurement are emerging [67]. No cell lives in isolation – beyond molecular profiling, the life history of the cell and its spatial relationship to other cells are critical determinants of cell identity. Undoubtedly, the considerable advances in cell lineage tracing [68] and spatial transcriptomics [69] will converge with multiomic profiling to enable comprehensive analysis of cellular identity in the context of where it is (Box 3), and where it has come from, but this will come with additional – and complex – computational and data science challenges. While each layer of information added to a multiomic analysis can bring new opportunities to classify cells and their biological context, it will also bring opportunities to study the mechanistic relationships between these different modalities in individual cells. While this an extremely exciting prospect, it requires the development of robust methods for the integration of diverse data types, each with their own idiosyncrasies. Packages such as Seurat [70] and MOFA+ [71] enable data integration from single-cell multiomic experiments, with the latter designed to

Outstanding questions

What is the optimal trade-off between resolution (number of measurements per cell) and throughput (number of cells analysed)?

What are the upper and lower limits of detection required to make meaningful, comprehensive investigations of cell type and state?

By operating at ultra-high-throughput, do we risk missing key details of cellular phenotypes – for example, lowly expressed genes, base-level (epi)genomic variation?

How can key measurements, such as histone modifications, DNA protein interactions, and isoform level gene expression, be integrated into multiomic approaches?

How can the nonmacromolecular components of the cell be integrated into multiomic studies – for example, the metabolome and lipidome?

What are the optimal computational approaches for data integration in multiomics approaches which take into account the various errors and sources of noise in parallel but distinct types of measurement made from the same single cell?

What level of detail is required to build accurate predictive models linking (epi)genomic variation and the function of protein–protein interaction networks?

How can antibody-independent proteomics be integrated with existing multiomic workflows?

Can single-cell multiomics methods be combined with spatial measurements, perhaps even in real time?

Can single-cell whole-genome sequencing – to base level resolution – be enabled at high throughput in a multiomic approach?

What are the applications for wider scientific questions – for example, can single-cell multiomics be applied to assemble and annotate genomes of nonmodel single-cell organisms?

Box 3. Spatial multiomics

The organisation of cellular structures and corresponding cell–cell interactions are fundamental to the operation of any multicellular system. Understanding the spatial organisation of cells within tissues is therefore essential to link molecular cell identities with organ- or organism-level functional biology. However, single-cell methods are not able to capture the spatial context of cells as the analysed tissue must be dissociated in order to be analysed. To address this need, there have been considerable advances in spatial transcriptomics, with transcriptome-wide or targeted approaches revealing gene expression patterns with regional, cellular, and even subcellular resolution [80].

Conventional *in situ* hybridisation [81] allows transcript detection at subcellular resolution, and recent developments of this approach have increased the multiplexing capacity for this approach from tens [82–85] to hundreds and even thousands of transcripts [85–88]. Untargeted methods have also expanded imaging-based *in situ* methodology to genome-wide profiling of gene expression [89,90]. Substrate-based approaches use positionally barcoded oligo-dT microarray features to locally capture mRNA molecules from tissue sections [91], with resolutions ranging from 50 μm (e.g., the 10X Visium platform), spanning multiple cells, through to methods approaching single-cell [92–94] and subcellular (<1 μm) resolution [95]. Spatial epigenomics approaches are also emerging, firstly with sciMAP-ATAC [96], where chromatin accessibility profiles obtained from tissue micropunches were matched with tissue spatial coordinates using combinatorial indexed transposition and sci-ATAC-seq workflow.

Spatial multiomic approaches are emerging – fluorophore- and oligonucleotide-conjugated antibodies can be incorporated into both *in situ* and array-based methods to enable parallel mRNA and protein detection, which has been demonstrated for several of the spatial transcriptomics methods mentioned previously [87,97], using the Nanostring GeoMX DSP instrument [98] and also very recently demonstrated in SPOTS [99], which combines the 10X Genomics Visium Platform with CITE-seq antibody-based protein detection. Novel approaches, such as the recently described DBiT-seq, can perform spatial profiling of mRNA and protein with 10 μm resolution [100]. DBiT-seq is based on two-step microfluidic-delivery of DNA barcodes directly to the surface of a tissue slide, and this approach has also enabled the spatially resolved profiling of accessible chromatin at approximately 20 μM resolution using *in situ* Tn5 transposition combined with microfluidic spatial barcoding [101].

Although only now emerging, these methods are likely to evolve rapidly, and far beyond transcriptomic and proteomic integration. Bringing multiomic methods with single-cell resolution together with imaging approaches will eventually enable comprehensive, three-dimensional molecular profiling of the dynamics of multicellular systems in development and disease.

identify cell classification factors and regulatory dependencies in scNMT-seq data. The continued development of computational tools that go beyond cell type classification, and can infer regulatory networks across multiple layers, is essential for future single-cell multiomic studies.

The ongoing convergence of methods enabling multiomic profiling of cellular molecular identity, localisation, and life history will dramatically change how we study multicellular living systems, offering unique insights into the regulatory processes governing how individual cells function collectively to produce whole-organism phenotypes in development, health, and disease.

Acknowledgments

The authors acknowledge support from an AIRC/CRUK/FC AECC Accelerator Award 'Single Cell Cancer Evolution in the Clinic', A26815 (AIRC number program 2279), Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation, Core Capability Grant (BB/CCG1720/1) and the National Capability in Genomics and Single Cell Analysis (BBS/E/T/000PR9816) and the BBSRC Core Strategic Programme Grant (Genomes to Food Security) BB/CSP1720/1. I.C.M. is supported by a BBSRC New Investigator Grant (BB/P022073/1).

Declaration of interests

F.G. is an inventor of a pending patent application relating to the GET-seq method.

References

1. Farrell, J.A. *et al.* (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* 360, eaar3131
2. Wagner, D.E. *et al.* (2018) Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360, 981–987
3. Xiang, L. *et al.* (2020) A developmental landscape of 3D-cultured human pre-gastrulation embryos. *Nature* 577, 537–542
4. Vento-Tormo, R. *et al.* (2018) Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* 563, 347–353

5. La Manno, G. *et al.* (2016) Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* 167, 566–580.e19
6. Tiklová, K. *et al.* (2019) Single-cell RNA sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development. *Nat. Commun.* 10, 581
7. Eze, U.C. *et al.* (2021) Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.* 24, 584–594
8. Regev, A. *et al.* (2017) The human cell atlas. *eLife* 6, e27041
9. Vijg, J. and Dong, X. (2020) Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell* 182, 12–23
10. Martincorena, I. (2019) Somatic mutation and clonal expansions in human tissues. *Genome Med.* 11, 1–3
11. McGranahan, N. and Swanton, C. (2017) Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 168, 613–628
12. Dey, S.S. *et al.* (2015) Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* 33, 285–289
13. Macaulay, I.C. *et al.* (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12, 519–522
14. Rodriguez-Meira, A. *et al.* (2019) Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. *Mol. Cell* 73, 1292–1305.e8
15. Han, K.Y. *et al.* (2018) SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res.* 28, 75–87
16. Zachariadis, V. *et al.* (2020) A highly scalable method for joint whole-genome sequencing and gene-expression profiling of single cells. *Mol. Cell* 80, 541–553.e5
17. Smallwood, S.A. *et al.* (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11, 817–820
18. Guo, H. *et al.* (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* 23, 2126–2135
19. Angermueller, C. *et al.* (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13, 229–232
20. Gu, H. *et al.* (2021) Smart-RRBS for single-cell methylome and transcriptome analysis. *Nat. Protoc.* 16, 4004–4030
21. Hu, Y. *et al.* (2016) Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* 17, 88
22. Hou, Y. *et al.* (2016) Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* 26, 304–319
23. Niemöller, C. *et al.* (2021) Bisulfite-free epigenomics and genomics of single cells through methylation-sensitive restriction. *Commun. Biol.* 4, 153
24. Cao, J. *et al.* (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 361, 1380–1385
25. Liu, L. *et al.* (2019) Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.* 10, 470
26. Xing, Q.R. *et al.* (2020) Parallel bimodal single-cell sequencing of transcriptome and chromatin accessibility. *Genome Res.* 30, 1027–1039
27. Zhu, C. *et al.* (2019) An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* 26, 1063–1070
28. Rosenberg, A.B. *et al.* (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182
29. Ma, S. *et al.* (2020) Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 183, 1103–1116.e20
30. Macosko, E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214
31. Chen, S. *et al.* (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* 37, 1452–1457
32. Xu, W. *et al.* (2022) ISSAAC-seq enables sensitive and flexible multimodal profiling of chromatin accessibility and gene expression in single cells. *bioRxiv*, Published online January 17, 2022. https://doi.org/10.1101/2022.01.16.476488
33. Lu, B. *et al.* (2020) Transposase-assisted tagmentation of RNA/DNA hybrid duplexes. *eLife* 9, e54919
34. Ninova, M. *et al.* (2019) The control of gene expression and cell identity by H3K9 trimethylation. *Development* 146, dev181180
35. Peters, A.H. *et al.* (2001) Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* 107, 323–337
36. Tedesco, M. *et al.* (2022) Chromatin velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nat. Biotechnol.* 40, 235–244
37. Janssens, D.H. *et al.* (2021) Automated CUT&Tag profiling of chromatin heterogeneity in mixed-lineage leukemia. *Nat. Genet.* 53, 1586–1596
38. Janssens, D.H. *et al.* (2021) Simultaneous CUT&Tag profiling of the accessible and silenced regulome in single cells. *bioRxiv*, Published online December 21, 2021. https://doi.org/10.1101/2021.12.19.473377
39. Zhang, B. *et al.* (2022) Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat. Biotechnol.* Published online March 24, 2022. https://doi.org/10.1038/s41587-022-01250-0
40. Lee, D.-S. *et al.* (2019) Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods* 16, 999–1006
41. Li, G. *et al.* (2019) Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat. Methods* 16, 991–993
42. Vogel, C. and Marcotte, E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232
43. Fredriksson, S. *et al.* (2002) Protein detection using proximity-dependent DNA ligation assays. *Nat. Biotechnol.* 20, 473–477
44. Darmanis, S. *et al.* (2016) Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep.* 14, 380–389
45. Genshaft, A.S. *et al.* (2016) Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol.* 17, 188
46. Frei, A.P. *et al.* (2016) Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* 13, 269–275
47. Reimegård, J. *et al.* (2021) A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun. Biol.* 4, 624
48. Peterson, V.M. *et al.* (2017) Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* 35, 936–939
49. Stoeckius, M. *et al.* (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868
50. Hwang, B. *et al.* (2021) SCITO-seq: single-cell combinatorial indexed cytometry sequencing. *Nat. Methods* 18, 903–911
51. Budnik, B. *et al.* (2018) SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* 19, 161
52. Specht, H. *et al.* (2021) Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.* 22, 50
53. Fiskin, E. *et al.* (2021) Single-cell profiling of proteins and chromatin accessibility using PHAGE-ATAC. *Nat. Biotechnol.* 40, 374–381
54. Clark, S.J. *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* 9, 781
55. Pott, S. (2017) Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife* 6, e23203
56. Yan, R. *et al.* (2021) Decoding dynamic epigenetic landscapes in human oocytes using single-cell multi-omics sequencing. *Cell Stem Cell* 28, 1641–1656.e7
57. Guo, F. *et al.* (2017) Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* 27, 967–988

58. Gu, C. *et al.* (2019) Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes. *Cell Res.* 29, 110–123
59. Mimitou, E.P. *et al.* (2021) Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* 39, 1246–1258
60. Lareau, C.A. *et al.* (2021) Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* 39, 451–461
61. Swanson, E. *et al.* (2021) Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* 10, e63632
62. Macaulay, I.C. *et al.* (2017) Single-cell multiomics: multiple measurements from single cells. *Trends Genet.* 33, 155–168
63. Dixit, A. *et al.* (2016) Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866.e17
64. Pierce, S.E. *et al.* (2021) High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat. Commun.* 12, 2969
65. Mimitou, E.P. *et al.* (2019) Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* 16, 409–412
66. Su, Y. *et al.* (2020) Multi-omic single-cell snapshots reveal multiple independent trajectories to drug tolerance in a melanoma cell line. *Nat. Commun.* 11, 2345
67. Seydel, C. (2021) Single-cell metabolomics hits its stride. *Nat. Methods* 18, 1452–1456
68. Kebschull, J.M. and Zador, A.M. (2018) Cellular barcoding: lineage tracing, screening and beyond. *Nat. Methods* 15, 871–879
69. Rao, A. *et al.* (2021) Exploring tissue architecture using spatial transcriptomics. *Nature* 596, 211–220
70. Hao, Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29
71. Argelaguet, R. *et al.* (2020) MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* 21, 111
72. Bian, S. *et al.* (2018) Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* 362, 1060–1063
73. Gaiti, F. *et al.* (2019) Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* 569, 576–580
74. Pastore, A. *et al.* (2019) Corrupted coordination of epigenetic modifications leads to diverging chromatin states and transcriptional heterogeneity in CLL. *Nat. Commun.* 10, 1874
75. Chaligne, R. *et al.* (2021) Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nat. Genet.* 53, 1469–1479
76. Zhu, Z. *et al.* (2020) Genome profiles of lymphovascular breast cancer cells reveal multiple clonally differentiated outcomes with multi-regional LCM and G&T-seq. *bioRxiv*, Published online February 1, 2020. https://doi.org/10.1101/807156
77. Li, L. *et al.* (2018) Single-cell multi-omics sequencing of human early embryos. *Nat. Cell Biol.* 20, 847–858
78. Castillo-Fernandez, J. *et al.* (2020) Increased transcriptome variation and localised DNA methylation changes in oocytes from aged mice revealed by parallel single-cell analysis. *Aging Cell* 19, e13278
79. Smith, Z.D. *et al.* (2017) Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer. *Nature* 549, 543–547
80. Moses, L. and Pachter, L. (2022) Museum of spatial transcriptomics. *Nat. Methods* Published online March 10, 2022. https://doi.org/10.1038/s41592-022-01409-2
81. Femino, A.M. *et al.* (1998) Visualization of single RNA transcripts in situ. *Science* 280, 585–590
82. Codeluppi, S. *et al.* (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* 15, 932–935
83. Wang, F. *et al.* (2012) RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagn.* 14, 22–29
84. Kishi, J.Y. *et al.* (2019) SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nat. Methods* 16, 533–544
85. Eng, C.-H.L. *et al.* (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 568, 235–239
86. Wang, X. *et al.* (2018) Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361, eaat5691
87. Chen, K.H. *et al.* (2015) RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348, aaa6090
88. Chen, X. *et al.* (2018) Efficient *in situ* barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res.* 46, e22
89. Lee, J.H. *et al.* (2014) Highly multiplexed subcellular RNA sequencing *in situ*. *Science* 343, 1360–1363
90. Alon, S. *et al.* (2021) Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* 371, eaax2656
91. Ståhl, P.L. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82
92. Rodriques, S.G. *et al.* (2019) Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463–1467
93. Stickels, R.R. *et al.* (2021) Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* 39, 313–319
94. Vickovic, S. *et al.* (2019) High-definition spatial transcriptomics for *in situ* tissue profiling. *Nat. Methods* 16, 987–990
95. Cho, C.-S. *et al.* (2021) Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* 184, 3559–3572.e22
96. Thornton, C.A. *et al.* (2021) Spatially mapped single-cell chromatin accessibility. *Nat. Commun.* 12, 1274
97. Vickovic, S. *et al.* (2022) SM-Omics is an automated platform for high-throughput spatial multi-omics. *Nat. Commun.* 13, 795
98. Merritt, C.R. *et al.* (2020) Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat. Biotechnol.* 38, 586–599
99. Ben-Chetrit, N. *et al.* (2022) Integrated protein and transcriptome high-throughput spatial profiling. *bioRxiv*, Published online March 18, 2022. https://doi.org/10.1101/2022.03.15.484516
100. Liu, Y. *et al.* (2020) High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* 183, 1665–1681.e18
101. Deng, Y. *et al.* (2021) Spatial-ATAC-seq: spatially resolved chromatin accessibility profiling of tissues at genome scale and cellular level. *bioRxiv*, Published online June 7, 2021. https://doi.org/10.1101/2021.06.06.447244
102. Gerlach, J.P. *et al.* (2019) Combined quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells. *Sci. Rep.* 9, 1469