Investigating Somatic Mutational Drivers of Immunogenicity in Colon Cancer



Jack Stephen Flower

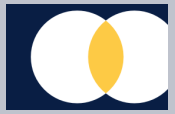
Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2025





Acknowledgments

Firstly, I extend deepest thanks to my supervisor Simon Buczacki who has provided unwavering support throughout my four-year DPhil and especially during writing this thesis. He is an exceptional scientist and mentor and I feel lucky to have been part of his lab, from organoid-flavoured lab meetings to family BBQs at his home – it's meant such a great deal. I look forward to continuing to work with him beyond this DPhil and to staying in touch for many years to come. I would like to thank all the Buczacki lab group members, particularly Yi, James, Kirsty and Amit who have provided considerable help across many aspects of my work and continue to provide a fun and warm lab environment. I also thank many colleagues in the department and across the University who have provided a supportive community and have helped my science in countless ways.

This work was funded by Cancer Research UK without whom it would not have been possible. I am deeply appreciative of charitable donations that enable students like me to contribute to cancer research while developing our careers. A sincere thanks to the DPhil Cancer Science and DTC staff Kea, Emily, Antoinette and Louise for their outstanding support throughout my time at Oxford.

I am deeply grateful to my close friends Dedrick, May and Aik, whose steady support, both personally and scientifically, has been fundamental to my DPhil journey. They have broadened my horizons well beyond the lab, from lively debates about the "diversity" of British cuisine to generous lessons in Singaporean and Malaysian life. I also thank Bea, Matt, Alice and Ben, who began their DPhil journeys alongside mine. Starting just a few desks away behind the -80°C freezers, we've shared the highs and lows from day one. I'm also thankful to my close home and university friends who have stayed by my side over the years, they mean the world.

My time at Linacre College gave me many lasting memories, especially during my time at Stoke House and my service as CR President and it forged many friendships I'll carry

forward. Heartfelt thanks to Ivy, Ting and Charlotte and to the Linacre staff especially Toni and Melissa for their essential help and support.

Special thanks to Ueshima, Grind and Lavazza for keeping me caffeinated – I estimate I drank about 7300 coffees over the four years (their beans powered this DPhil).

I extend indebted thanks to my family who have been an incredible support system throughout the eight years across three different universities. Particular thanks to the DeNobrigas for invaluable help with applications, interviews and wise counsel. To my grandmas for flapjacks and apple crumble, and to my grandad for his amazing recipes and home-grown vegetables, their nutritional support and care hold a special place in my heart. Lastly, I cannot thank my parents enough. They've been a rock for me throughout my life and academic career and their constant support (alongside Oscar and his zoomies) kept me sane during difficult times. I really wouldn't be where I am today without my family's support and I love them all so much.

Not forgetting a heartfelt thank you to my boyfriend, Kieran. I can't quite put into words how much I appreciate him and everything he does for me. Whether it's taking out the bins, introducing me to a whole world of beach volleyball, or being there when I'm stressing over a failed experiment, I truly couldn't imagine life without him.

I also want to give credit to my younger self. Over the past eight years I've achieved more than I ever thought possible coming from little old Leek in the Peak District, not only academically, but personally too. Finding the strength to accept who I am was the hardest thing I've ever done (and that's considering I've done a western blot for ZFHX4). Oxford provided me with a beautiful next chapter of my life, and I'm eternally thankful to my family and friends for their acceptance, love and understanding. And to future me: I hope you remember how proud you were at this moment, submitting this thesis.

Dedication

I would like to dedicate this thesis to many of my family members who have suffered or have passed away from Colorectal Cancer, including my late Grandad Linden who sadly passed away in 2015, one of the reasons for my passion and drive in cancer research.

Preface

The work described in this thesis was carried out in the Department of Oncology and Nuffield Department of Surgical Sciences, University of Oxford between October 2021 and 2025. It was funded by Cancer Research UK as part of the DPhil in Cancer Science Programme within the Oxford Cancer Centre.

This thesis is the result of my own work. Excluding supplementary data, tables, figures and references, this thesis does not exceed 50 000 words.

Table of contents

Acknowledgments i

Dedication iii

Preface iv

List of key abbreviations x

List of figures xii

List of tables xvi

Abstract xix

Chapter One: Introduction 1

1.1 An introduction to colorectal and colon cancer 1

1.2 Colon tumourigenesis 2

1.3 Colon cancer immunology 6

1.3.1 Foundations of tumour immunology 6

1.3.2 Immune checkpoint modulators (ICMs) for colon cancer immunotherapy 7

1.3.3 The tumour-immune microenvironment in colon cancer 11

1.4 Defining tumour immunogenicity and its determinants 13

1.5 Expanding the clinical impact of immunotherapy in colon cancer 15

1.6 Investigating somatic mutational drivers of immunogenicity 19

1.7 A brief thesis overview 20

Chapter Two: Materials and methods 22

2.1 Cell culture 22

2.1.1 Cell lines 22

2.1.2 Human tissue collection 23

2.1.3 Organoid culture 23

2.1.4 Primary T-cell culture 24

2.2 Immunohistochemistry (IHC) 25

2.2.1 Tissue preparation, embedding and sectioning 25

2.2.2 Tissue staining 26
2.3 Transcriptomic profiling 27
2.3.1 RNA extraction 27
2.3.2 RNA sequencing (RNAseq) and analysis 27
2.4 Functional profiling 28
2.4.1 Cell line-based 28
2.4.1.1 Viability and proliferation 28
2.4.1.2 Cell cycle dynamics 28
2.4.1.3 Cytokine arrays 29
2.4.1.4 Conditioned medium generation 29
2.4.1.5 Migration and invasion 30
2.4.2 Angiogenesis-based 31
2.4.3 Immune-based 31
2.5 CUT&RUN 32
2.6 CRISPR/Cas9 33
2.6.1 Generating a KO: RNP electroporation 33
2.6.2 Validating a KO: Sanger sequencing 34
2.7 siRNA knockdown 35
2.8 Western blotting 36
2.8.1 Protein extraction, quantification and preparation 36
2.8.2 SDS-PAGE and protein transfer 36
2.8.3 Blotting 37
2.8.4 Nuclear fractionation 38
2.9 Flow Cytometry 39
2.10 Bioinformatics 41
2.10.1 Programming language and software 41
2.10.2 Dataset acquisition and processing 42
2.10.3 Multiomic clustering on TCGA-COAD data 45
2.11 Data visualisation and statistical analyses 45

Chapter Three: Interrogating tumour immunogenicity in publicly available datasets 50

3.1 Introduction 50

3.2 Hypothesis 51

3.3 Experimental Design 52

3.3.1 Overview 52

3.3.2 Defining mutational load (ML) phenotype 52

3.3.3 Comparing mutations between Imm^{high} and Imm^{low} patients categorised by five immunogenicity parameters 53

3.3.4 Comparing mutations between mDOI-defined subgroups characterised as Immunogenicity classification of Colon cancer Subtypes (ICS) 54

3.4 Results 55

3.4.1 Extrapolating immune-related phenotype in transcriptomic data reveals heterogeneity within COAD tumours irrespective of mutational load phenotype 55

3.4.2 Mutation comparison between Imm^{low} and Imm^{high} groups reveals *ZFHX4* as a potential mutational driver of immunogenicity 63

3.4.3 Clustering patients into Immunogenicity Classification of Colon Cancer Subtypes (ICS1-5) offers an alternative method to identify somatic mutations in distinct immune-biological subgroups 66

3.4.4 Mutation comparison between ICS3 and ICS5 groups revealed *ZFHX4* as a potential mutational driver of immunogenicity 75

3.5 Discussion 77

3.5.1 A specific focus on the Imm^{group} approach 77

3.5.2 A specific focus on the ICS clustering approach 78

3.5.3 Evaluating both bioinformatic approaches 91

3.5.4 The association of *ZFHX4* mutations with enhanced immunogenicity 93

3.5.5 Conclusion 96

Chapter Four: ZFHX4 as a potential regulator of colon tumourigenicity 97

4.1 Introduction 97

4.1.1 ZFHX4 classification and structure 97

4.1.2 ZFHX4 in normal physiology 100

4.1.3 ZFHX4 in disease and disorders 102

4.1.4 ZFHX4 in cancer 104

4.2 Hypothesis 107

4.3 Experimental design 107

4.4 Results 108

4.4.1 ZFHX4 expression in normal and cancer intestinal tissues 108

4.4.2 Generation of *ZFHX4*$^{-/-}$ cell line and organoid models 127

4.4.3 Transcriptomic profiling of *ZFHX4*$^{-/-}$ organoids 132

4.4.4 Genomic, transcriptomic and proteomic validation of *ZFHX4*$^{-/-}$ cell lines 142

4.4.5 Transcriptomic profiling of *ZFHX4*$^{-/-}$ cell lines 162

4.4.6 Validating ZFHX4 transcriptional targets with CUT&RUN 169

4.4.7 Functional profiling of *ZFHX4*$^{-/-}$ cell lines 175

4.4.8 Integrative analysis of *ZFHX4* in clinical cohorts 198

4.5 Discussion 215

4.5.1 Organoid and cell line models 216

4.5.2 gDNA, mRNA and protein profiling of ZFHX4 221

4.5.3 Other methods and techniques 225

4.5.4 Mechanistic insights of ZFHX4 226

4.5.5 Conclusion 233

Chapter Five: ZFHX4 as a potential regulator of immunogenicity 234

5.1 Introduction 234

5.2 Hypothesis 236

5.3 Experimental Design 236

5.4 Results 237

5.4.1 *ZFHX4* mutations in cancer and COAD 237

5.4.2 Obtaining a pan-T population and testing artificial activation technique 241

5.4.3 Effects of *ZFHX4*$^{-/-}$ conditioned medium on T-cell activation and exhaustion 245

5.4.4 Optimisation of T-cell migration Transwell assays 259

5.4.5 Effects of *ZFHX4*$^{-/-}$ conditioned medium on T-cell migration 262

5.5 Discussion 265

5.5.1 *ZFHX4*$^{-/-}$ doesn't model clinically relevant missense mutations 265

5.5.2 Healthy donor T-cells are a minimalist approach toward immunogenicity 267

5.5.3 Predicted function of *ZFHX4* missense mutations 272

5.5.4 Future directions 279

5.5.5 Conclusion 281

Chapter Six: Discussion 282

References 290

Supplementary materials .....

List of key abbreviations

A-cells	activated T-cells
aa	amino acid
APCs	antigen-presenting cells
BME	basement membrane extract
bp	base pair
CDG	cancer driver gene
CM	conditioned medium
CMS	consensus molecular subtype
COAD	colon adenocarcinoma
CRC	colorectal cancer
CRISPR	clustered regularly interspaced short palindromic repeats
CTG	cell titre glo
CUT&RUN	cleavage under targets & release using nuclease
dH20	distilled water
DSS	disease-specific survival
ECL	electrochemiluminescence
ECM	extracellular matrix
EMT	epithelial to mesenchymal transition
ES	enrichment score
FDR	false discovery rate
FMO	fluorescence minus one
FSC	forward scatter
GoF	gain of function
GSEA	gene set enrichment analysis
GSVA	gene set variation analysis
HR	hazard ratio
ICE	inference of CRISPR edits
ICM	immune checkpoint modulator
ICS	immunogenicity classification of colon cancer subtypes
IDG	immunogenicity driver gene
IHC	immunohistochemistry
IO	immuno-oncology
IPMS	immunoprecipitation mass spectrometry
IVTFAs	in vitro tube formation assays
kb	kilobase
KD	knockdown
KM	kaplan-meir
KO	knockout
LoF	loss of function
LR	log-rank
Mb	megabase
mCC	metastatic colon cancer
mDOI	multiple determinants of immunogenicity
MFI	median fluorescence intensity
mgRNA	multi-guide RNA

MHC	major histocompatibility complex
MMRd	mismatch repair deficient
MMRp	mismatch repair proficient
monoKO	monoclonal knockout
MSI	microsatellite instability/unstable
MSS	microsatellite stability/stable
NA-cells	non-activated T-cells
NES	normalised enrichment score
NFW	nuclease-free water
NK	natural killer
NOM	nominal (p-value)
NuRD	nucleosome remodelling and histone deacetylase
ONT	overnight
OS	overall survival
PCA	principal component analysis
PCR	polymerase chain reaction
PFS	progression-free survival
polyKO	polyclonal knockout
PPd	polymerase proofreading deficiency
RNAseq	RNA sequencing
RTP	room temperature
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
SEM	standard error of the mean
sgRNA	single-guide RNA
SNV	single nucleotide variant
SSC	side scatter
TAM	tumour associated macrophage
TCGA	the cancer genome atlas
TCR	T-cell receptor
TF	transcription factor
TI	tumour immunogenicity
TIDE	tumour immune dysfunction and exclusion
TIL	tumour infiltrating lymphocyte
TIME	tumour immune microenvironment
TMBH	tumour mutational burden high
TMBL	tumour mutational burden low
TME	tumour microenvironment
TNM	tumour-node-metastasis
TSG	tumour suppressor gene
VCN	viable cell number
VUS	variant of unknown significance
WB	western blotting
WES	whole exome sequencing
WT	wild-type
ZF	zinc finger
ZFHX4	zinc finger homeobox 4

List of figures

---Chapter One---

Figure 1.1 *Colon cancer anatomical origin, with cancer staging criteria and their respective therapies used clinically* 1

Figure 1.2 *The mechanism of action of approved immunotherapy in colon cancer highlighted with respective clinical trials* 9

Figure 1.3 *Tumour immunogenicity (TI) and its multiple determinants* 14

Figure 1.4 *Durable clinical benefit for more colon cancer (CC) patients* 16

---Chapter Three---

Figure 3.1 *Defining mutational load (ML) phenotype in TCGA-COAD (n=341)* 57

Figure 3.2 *Stratifying TCGA-COAD samples by Immune Score irrespective of ML phenotype uncovers a spectrum of immunogenicity that remains hidden when stratifying by TMB and ML phenotype alone (n=341)* 60

Figure 3.3 *Deriving a five-parameter based immunogenicity classification for assigning TCGA-COAD samples into Imm^{low} and Imm^{high} groups* 62

Figure 3.4 *Internal validation of cluster number selection and consequent identification of TCGA-COAD subgroups following k-means clustering on multiple determinants of immunogenicity (mDOI)* 67

Figure 3.5 *Characterisation of TCGA-COAD clusters as Immunogenicity classification of Colon cancer Subtypes (ICS1-5)* 69

Figure 3.6 *Kaplan-Meier (KM) curves depicting Overall (OS; A), Progression-Free (PFS; B) and Disease-Specific Survival (DSS; C) of TCGA-COAD patients stratified by ICS group* 71

---Chapter Four---

Figure 4.1 *Domain-based classification of ZFHX4 and its variation in structure at the genomic, transcriptomic and proteomic level* 98

Figure 4.2 *Numbers and proportions of ZFHX4-expressing cell types in healthy intestine scRNAseq datasets* 110

Figure 4.3 *ZFHX4 mRNA expression across multiple cancers and their prognosis in the TCGA-PanCancerAtlas dataset* 112

Figure 4.4 *Stratifying COAD patients by ZFHX4 expression confers significant prognoses across multiple cohorts* 115

Figure 4.5 *ZFHX4 expression increases with colon cancer stage* 117

Figure 4.6	*ZFHX4 expression at the single cell level in COAD and matched normal tissue*	119
Figure 4.7	*ZFHX4 protein expression across matched normal colon, small intestine and COAD tissue from two patients*	123
Figure 4.8	*Model selection for studying ZFHX4 in cell lines and organoids*	129
Figure 4.9	*CRISPR/Cas9 strategy employed to generate ZFHX4$^{-/-}$ models and additional pipelines in cell lines to obtain monoclonal populations*	131
Figure 4.10	*Expansion and CRISPR-mediated KO of ZFHX4 in a metastatic colon organoid line*	133
Figure 4.11	*RNAseq and further sanger sequencing results of ZFHX4$^{-/-}$ metastatic organoids*	135
Figure 4.12	*Differential gene and pathway analysis of ZFHX4$^{-/-}$ metastatic organoids*	137
Figure 4.13	*Scatter plots depicting expression of top genes positively or negatively correlated with ZFHX4 editing*	140
Figure 4.14	*Gene set enrichment analysis (GSEA) results of senescence and apoptosis in ZFHX4$^{-/-}$ organoids and nosgRNA controls*	141
Figure 4.15	*Genomic validation of monoclonal ZFHX4$^{-/-}$ cell lines by sanger sequencing and their morphology*	143
Figure 4.16	*Genomic validation of polyclonal and monoclonal ZFHX4$^{-/-}$ cell lines by sanger sequencing over the first three months of culture*	146
Figure 4.17	*RNAseq data demonstrating ZFHX4 expression in WT, polyKO and monoKO cell lines*	149
Figure 4.18	*Transcript-level ZFHX4 expression in WT, polyKO and monoKO cell lines referring to all isoforms or protein-coding isoforms only*	152
Figure 4.19	*Initial WB optimisation for ZFHX4 protein*	154
Figure 4.20	*Characterisation of ZFHX4 protein expression in WT, polyKO and monoKO lines using the HPA antibody*	156
Figure 4.21	*Further ZFHX4 WB optimisation comprising nuclear fractionation and sample preparation adjustment experiments*	157
Figure 4.22	*Binding regions of all commercially available antibodies to human ZFHX4, with reference to the Q86UP3-5 isoform*	158
Figure 4.23	*Quantification of ZFHX4 protein expression across all cell lines and replicates, including a newly-optimised WB approach*	161
Figure 4.24	*PCA plots representing RNAseq expression profiles of WT, polyKO and monoKO cell lines*	163
Figure 4.25	*Differential gene and pathway analysis of SKNMC ZFHX4$^{-/-}$ cells*	164
Figure 4.26	*Differential gene and pathway analysis of NCIH716 (A-B) and COLO320DM (C-D) ZFHX4$^{-/-}$ cells*	166

Figure 4.27	*CUT&RUN of H3K4me3 peaks on SKNMC, NCIH716 and COLO320DM cell lines*	171
Figure 4.28	*Overlap of CUT&RUN H3K4me3-centric cell line data with matched RNAseq and ZFHX4-centric CUT&RUN data from the literature*	173
Figure 4.29	*Linear growth curves of various ZFHX4 WT and ZFHX4$^{-/-}$ cell lines, plotting normalised viable cell number (VCN) to day 0 over time*	176
Figure 4.30	*Quantification of proliferation in SKNMC, COLO320DM and NCIH716 WT and ZFHX4$^{-/-}$ lines, defined using doubling time calculations*	177
Figure 4.31	*siRNA knock-down of ZFHX4 in nosgRNA COLO320DM cells and resulting proliferation differences observed after 3 days*	179
Figure 4.32	*Cell cycle gating strategy, using CWT and controls alongside confirmation of consistency with NWT*	181
Figure 4.33	*Cell-cycle phase quantification in ZFHX4 WT and KO cells*	183
Figure 4.34	*Quantification of cytokine arrays measuring secreted cytokines in ZFHX4-WT and ZFHX4$^{-/-}$ conditioned medium*	186
Figure 4.35	*HUVEC cells utilised for in-vitro tube formation assays (IVTFAs) and analysis of angiogenic biology using the Angiogenesis Analyser macro*	189
Figure 4.36	*The Angiogenic Landscape of HUVECs following multiple IVTFA experiments*	192
Figure 4.37	*Angiogenic potential of ZFHX4 WT and ZFHX4$^{-/-}$ cell lines*	195
Figure 4.38	*Angiogenic potential of pro-angiogenic factors and angiogenesis inhibitors in IVTFAs*	196
Figure 4.39	*ZFHX4 expression and mutation across four colon cancer cohorts, with schematics demonstrating method of differential analysis*	199
Figure 4.40	*Concordance and correlation of differential gene and pathway analysis of ZFHX4^{high} vs ZFHX4^{low} expression across four independent colon cancer cohorts*	202
Figure 4.41	*Meta-analysis of differential gene expression and pathway enrichment between ZFHX4^{high} and ZFHX4^{low} patient groups across four independent cohorts*	206
Figure 4.42	*Correlation plots of differentially expressed genes and pathways between all four clinical cohorts and 8 experimental datasets*	210
Figure 4.43	*Bubble plots of genes up- or down-regulated in differential gene expression comparisons amongst clinical and experimental datasets*	213
Figure 4.44	*The translation from epithelial to mesenchymal cell states through a multi-layer regulatory network, known as EMT*	228

---Chapter Five---

Figure 5.1	*Mutational spectra of ZFHX4 pan-cancer and in four independent COAD datasets additionally segregated into TMBL groups*	238
Figure 5.2	*Healthy donor pan-T-cells and artificial activation by Dynabeads in vitro*	242
Figure 5.3	*Investigating activation marker expression following Dynabead incubation*	244
Figure 5.4	*QC of instrument-based compensation using A cells from D1 as an example*	246
Figure 5.5	*Flow cytometry gating strategies to identify populations of live $CD4^+$ and $CD8^+$ T-cells that express a variety of activation markers*	248
Figure 5.6	*PCA of the flow cytometry dataset spanning samples from different donors, cell activation states and conditioned media conditions*	251
Figure 5.7	*Viability, CD4:CD8 compartment and activation metric differences between NA and A cells cultured in control media across all donors*	253
Figure 5.8	*T-cell activation changes in $ZFHX4^{-/-}$ CM relative to ZFHX4-WT CM*	256
Figure 5.9	*T-cell IFNγ changes in $ZFHX4^{-/-}$ CM relative to ZFHX4-WT CM*	258
Figure 5.10	*T-cell migration assay optimisation using 24-well transwells*	260
Figure 5.11	*Migration of activated healthy donor T-cells towards controls and $ZFHX4^{-/-}$ conditioned medium*	263

---Chapter Six---

Figure 6.1	*ZFHX4 as a potential regulator of tumourigenicity and immunogenicity in colon cancer*	284

List of tables

---Chapter One---

Table 1.1	*Fundamental classification systems and respective subtypes of colon cancer*	4
Table 1.2	*Extrinsic immunosensitisation of MSS/MMRp colon cancers*	17

---Chapter Two---

Table 2.1	*Immortalised human cell line derivations with their respective media bases*	22
Table 2.2	*ENRAS supplements added to basal media to culture metastatic tumour organoids*	24
Table 2.3	*Primary and secondary antibodies used for IHC*	26
Table 2.4	*Optimisation of cytokine array method*	29
Table 2.5	*Positive control, negative control and ZFHX4 antibodies used for CUT&RUN*	32
Table 2.6	*Reagents and electroporation settings for CRISPR/Cas9-mediated gene KO for cell lines and organoids*	33
Table 2.7	*Primers used for PCR amplification of the CRISPR-targeting region of ZFHX4*	34
Table 2.8	*siRNA sequences and targeted regions of ZFHX4*	35
Table 2.9	*SDS-PAGE and protein transfer parameters*	37
Table 2.10	*Primary and secondary antibodies used for western blotting*	38
Table 2.11	*Flow cytometry antibodies and dyes used in various flow panels*	39
Table 2.12	*R packages used throughout the project (mostly installed with Bioconductor v3.16)*	41
Table 2.13	*Summary of supporting colon cancer datasets used throughout the thesis*	43
Table 2.14	*RNAseq processing of cell line and organoid ZFHX4$^{-/-}$ experiments*	44
Table 2.15	*Parameters employed in the ICS clustering pipeline, each representing a determinant of tumour immunogenicity (DOI)*	46

---Chapter Three---

Table 3.1	*Key parameters of immunogenicity that can be mined from TCGA-COAD*	53
Table 3.2	*Outlier ML groups and their ML-descriptor values*	58
Table 3.3	*Mutation comparison between Imm^{high} and Imm^{low} groups of all and nHM-only TCGA-COAD samples, with reference to GSEA results of mt vs wt hot tumour signatures as well as an independent immunotherapy-treated cohort of pan-cancer samples*	65
Table 3.4	*Summary of biological themes predominantly contributing to PC1, PC2 and PC3 in the PCA of scaled clustering matrix*	72
Table 3.5	*Proposed characterisation of ICS groups in TCGA-COAD, reflecting differences in multiple determinants of immunogenicity following clustering*	74
Table 3.6	*Comparison of ICS3 (n=82) vs ICS5 (n=41) mutational profiles*	75
Table 3.7	*An exhaustive chronological selection of previously published immune subtypes of colon cancer employing a clustering-style approach*	80

---Chapter Four---

Table 4.1	*Publications reporting on ZFHX4 in cancer and tumourigenicity through in vitro (IV) or in silico (IS) means*	104
Table 4.2	*Summary of scRNAseq datasets*	108
Table 4.3	*Staining profiles of ZFHX4 on small intestine by antibody, irrespective of concentration*	120
Table 4.4	*Top 20 positively and negatively correlated genes with % editing efficiency, potentially highlighting genes regulated by ZFHX4 in a dose-dependent manner*	138
Table 4.5	*Predictions of KO scores and top 3 indel types by ICE analysis of polyKO and monoKO clones at month 0*	145
Table 4.6	*Predictions of KO scores and top 3 indel types by ICE analysis of polyKO and monoKO clones over time*	147
Table 4.7	*Number of called peaks with SEACR, using the negative IgG control as background*	169
Table 4.8	*TRANSFAC motif enrichment in the set of all overlapping genes (n=68) performed by gProfiler*	174
Table 4.9	*Selecting the top and bottom significant (adj.p<0.05) 200 genes by t-statistic per cohort and overlapping across cohorts identifies ZFHX4-associated genes identified in at least 3 cohorts*	203
Table 4.10	*Ranked by mean log_2FC, the top 50 positively and negatively associated genes with ZFHX4 expression across all four cohorts*	207

---Chapter Five---

Table 5.1 *Top 10 commonly altered residues in ZFHX4 by mutations in all four COAD cohorts* 239

Table 5.2 *Experimental setup of two T-cell migration Transwell-based experiments* 262

Table 5.3 *A brief summary of epithelial-T-cell co-cultures spanning mostly colon contexts* 268

Abstract

Immunotherapy has transformed the management of metastatic colon cancer, with immune checkpoint modulators now approved as first-line therapy. However, fewer than 5% of metastatic cases meet the mismatch repair deficiency (MMRd) and microsatellite instability (MSI) criteria, and only 30-50% of these patients achieve durable responses. Immunogenicity, defined as the capacity of a tumour to elicit a pre-existing anti-tumour immune response conducive to immunotherapy, involves far more than mutational load alone. This project set out to identify and functionally investigate genomic regulators of immunogenicity, termed somatic mutational drivers of immunogenicity, through an integrated bioinformatic and experimental approach. Analyses of TCGA-COAD and independent colon cancer datasets revealed ZFHX4 missense mutations were significantly enriched in tumour mutational burden low yet immunogenic tumours. ZFHX4 is a poorly characterised transcription factor whose expression is strongly prognostic and among the most frequently mutated genes in colon cancer. In this thesis, ZFHX4 expression was observed in epithelial, stromal and neural compartments in colon tissue and matched cancer specimens. Using CRISPR-generated organoid and cell line models, tumourigenic and immunogenic phenotypes were explored in parallel. Although experimental results exhibited variable trends with mixed statistical significance, integration with multiomic datasets and published literature enabled the development of a unifying model of ZFHX4 function in the colon. Under physiological conditions, ZFHX4 appears to cooperate with the nucleosome remodelling and deacetylase (NuRD) complex to regulate stemness, Wnt signalling and epithelial-mesenchymal transition (EMT). In colon cancer, high ZFHX4 expression promotes tumourigenic EMT-related programmes and associates with poor prognosis, whereas missense mutations may disrupt NuRD recruitment and reorient transcriptional activity toward immune-related pathways. These findings propose ZFHX4 as a potential regulator of both tumourigenicity and immunogenicity in colon cancer, warranting further mechanistic validation.

Chapter One: Introduction

1.1 An introduction to colorectal and colon cancer

Colorectal cancer (CRC) is the third most diagnosed malignancy and the second leading cause of cancer-related death worldwide (Bray et al., 2024). Originating in intestinal epithelium, the disease can be split into colon and rectal cancers depending on anatomical derivation (Figure 1.1A) which are managed somewhat differently clinically (Minsky, 2011). This thesis focuses specifically on colon cancer, which accounts for roughly 70% of CRC cases (Hoadley et al., 2018) and aligns with the Buczacki group's core expertise and ready access to resected colon tissue for *in vitro* study.

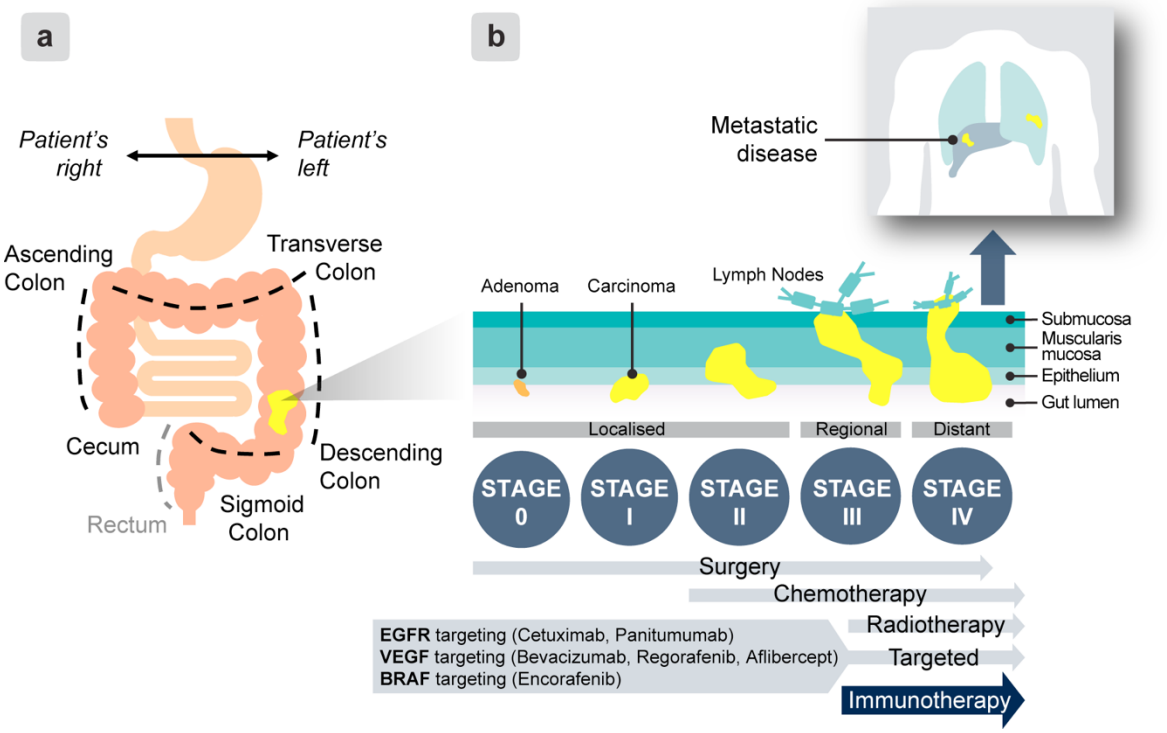


Figure 1.1 | Colon cancer anatomical origin, with cancer staging criteria and their respective therapies used clinically. (A) Schematic depicting the gastrointestinal tract and the key areas of the colon where tumours may occur. **(B)** Tumour-node-metastasis (TNM) staging of colon cancer depends on the tumour's invasion through the gastrointestinal wall, lymph nodes and eventually distant metastatic sites. Respective clinical therapies annotated from (NICE, 2025) are shown in arrows.

The tumour-node-metastasis (TNM) system stages colon cancer by disease severity according to tumour size, depth of invasion, lymph nodding involvement and distant metastasis (Figure 1.1B; (Cancer Research UK, 2025b)). Localised stage I-II colon cancers are usually curable through surgery alone and carry a ~90% 5-year survival, though stage III-IV cancers demonstrate variable responses to chemotherapy and targeted agents and present with high rates of recurrence and/or unresectable disease (Cancer Research UK, 2025a, NIH National Cancer Institute, 2025). This is reflected by poor 10% 5-year survival of stage IV metastatic colon cancer (mCC) where malignant spread through the intestinal wall commonly leads to liver and lung metastasis (Figure 1.1B; (Ilyas, 2024)). Treatment options for stage IV colon cancer are broad but commonly fail to ultimately cure, leading to patient death (Boire et al., 2024). There is therefore a strong impetus to detect colon cancer early through population screening by faecal immunochemical testing (FIT) and colonoscopy. Colonoscopy works by accurately detecting and removing polyploid precursors of invasive colon cancer but is expensive and not without morbidity (NHS England, 2025). Screening programmes have been pivotal to shift diagnosis to earlier stages, reducing colon cancer incidence and mortality (Bretthauer et al., 2022, Atkin et al., 2017, NHS England, 2025). In parallel, improving outcomes for metastatic disease remains a major focus, where improving *current* and providing *more* treatment options for the metastatic setting are preeminent. Immunotherapy exemplifies this progress and forms a central focus of this thesis.

1.2 Colon tumourigenesis

Tumourigenesis is a complex multistep biological process through which normal cells acquire genetic and epigenetic alterations that drive uncontrolled proliferation, survival and eventual dissemination into surrounding tissues (Fearon and Vogelstein, 1990, Stratton et al., 2009, Hanahan and Weinberg, 2000). Central to this process, such alterations involve activation of oncogenes and inactivation of tumour suppressor genes (TSGs), with the latter largely consisting of (1) caretaker genes responsible for maintaining genomic fidelity, (2)

gatekeeper genes responsible for controlling cellular proliferation and (3) landscaper genes capable of modulating the tumour microenvironment (Kinzler and Vogelstein, 1997, Ashworth et al., 2011). Consequent dysregulation of key homeostatic cellular pathways and acquisition of novel capabilities during tumourigenesis have been succinctly summarised as cancer hallmarks, enabled through genomic instability, tumour-promoting inflammation, epigenetic reprogramming and the tissue microbiome (Hanahan, 2022). Eight core principles crucial for neoplasia include sustaining proliferative signalling, evading growth suppressors, avoiding immune destruction, enabling replicative immortality, activating invasion and metastasis, inducing or accessing vasculature, resisting cell death and deregulating cellular metabolism (Hanahan and Weinberg, 2011). Therefore, most research into tumourigenesis has centred on defining and interrogating these neoplastic attributes with a wide range of *in vitro* and *in vivo* models established to assess them.

Colon tumourigenesis results from the contribution of sporadic and heritable genetic and environmental factors imposed upon colonic epithelium. Whilst recent work highlights the impact of the environment (Diaz-Gay et al., 2025, Murphy et al., 2012, Aleksandrova et al., 2014), the majority of recent research has been focused on its genetic elements. Originally proposed in the early 90s in the classic adenoma-carcinoma sequence, the molecular pathogenesis of sporadic colon cancer describes gatekeeper *APC* mutations as key initiator events to metaplasia, followed by *TP53* and *SMAD4* TSG inactivation and *KRAS* oncogene activation (Fearon and Vogelstein, 1990, Powell et al., 1992, Markowitz and Bertagnolli, 2009). These events principally drive proliferation in intestinal stem cells, for example through upregulated Wnt signalling, and prompt aberrant crypts, benign polyp adenomas and eventually carcinoma – a sequence that can unfold up to 15 years (Barker et al., 2009, Fearon and Vogelstein, 1990, Dekker et al., 2019).

However, this traditional view of colon cancer initiation and progression is now considered fairly simplistic. Not all cancers follow this classical sequence (Chen et al., 2021) and distinct mutational processes can coexist within a single tumour, giving rise to intratumoural

heterogeneity governed by clonal selection – an important contributor to both disease progression and therapy response (McGranahan and Swanton, 2017). Beyond standard histological grading and TNM staging, there has been significant progress in the past 50 years to subclassify colon cancer based on molecular and clinical features. These classification systems range from broad histopathological patterns to precise molecular subtypes, reflecting the complexity and heterogeneity of the disease (Table 1.1). While many of these classifications overlap, the table provides a high-level overview of how different groups have approached subtyping, reinforcing the notion that there is no single pathway to colon tumourigenesis.

Table 1.1 | Fundamental classification systems and respective subtypes of colon cancer

Classification system	Subtype	Phenotype
Genetic aetiology	Sporadic[1,2]	Somatic mutation aetiology (not inherited). Represent >70% colon cancer cases and have low lifetime risk (3-5%) with old age onset.
	Familial[1,3]	Non-syndromic germline mutation aetiology (inherited but poorly penetrant). Represent 25-30% colon cancer cases with causes largely unknown, but colon cancer risk increases with number of affected first-degree relatives.
	Hereditary[1,4-6]	Syndromic germline mutation aetiology (inherited and highly penetrant). Represent <5% colon cancer cases, with high lifetime risk (>80%) and young age onset. Well known examples include HNPCC-Lynch (MMR mutations) and FAP (APC mutations)
Histo-pathology of pre-cursor lesions[7-10]	Conventional adenomatous	High cancer progression risk. Represent majority of polyps overall and present with dysplasia. Can be tubular, villous or tubulovillous. Associated with classic adenoma-carcinoma model and MSS/MMRp/TMBL
	Hyperplastic	Represent most serrated polyps and are mostly of distal origin. Non-neoplastic, with low cancer risk. Further subtypes of these exist
	Traditional serrated adenoma (TSA)	Represent ~5% serrated polyps and characterised by KRAS and/or BRAF mutations. Mostly of distal origin. Associated with cIMP, MSI, MGMT hypermethylation
	Sessile serrated lesion (SSL)	Represent ~25% serrated polyps and characterised by BRAF mutations. Usually of proximal origin. May or may not present with dysplasia. Associated with cIMP, MSI, MLH1 hypermethylation
	Harmartom-atous	Rare – specific to certain syndromes like Peutz-Jeghers

Histo-pathology of adeno-carcinoma[11-13]	Conventional	Desmoplastic stroma. Associated with the classic adenoma-carcinoma model and represent most colon cancers
	Mucinous	Abundant with extracellular mucin, accounting for <50% tumour volume. Usually proximal and advanced stage
	Signet-ring	Characteristic of signet-ring cells, often have poor prognosis and advanced stage
	Medullary	Very rare, poorly or undifferentiated. Commonly with excessive lymphocyte infiltration and associated with MSI
Anatomical sidedness[14,15]	Left	Hindgut derived, greater incidence in men, CMS2 and CMS4 predominance
	Right	Midgut derived, greater incidence in women, associated with CMS1/CSM3, MSI, SSLs and mucinous tumours
Genomic (instability)[16-18]	CIN	Associated with the classical adenoma-carcinoma sequence and describes sequential accumulation of chromosomal alterations
	MSI	Results from faulty DNA MMR pathways and describes expansion of repetitive DNA sequences called microsatellites. Leads to TMBH
	cIMP	Overlaps with CIN and MSI. Describes high frequency of methylation at CpG sites and associated with BRAF muations, MLH1 hypermethylation and MSI
Mutational load[19]	Non-hypermutated	MMRp and subsequently MSS. Tumours are TMBL and have <10mt/Mb
	Hypermutated	MMRd and subsequently MSI. This leads to TMBH (>10mt/Mb)
	Ultra-hypermutated	Displays MSS features but POLE/POLD1 mutations generate excessive mutational load (>100mt/Mb)
Transcriptomic (CMS)[20]	CMS1	Roughly 15% all colon cancers, associated with MSI, cIMP, TMBH, immune infiltration
	CMS2	Roughly 37% all colon cancers, associated with canonical APC mutations and Wnt activation
	CMS3	Metabolic deregulation, KRAS mutations
	CMS4	Stromal infiltration, angiogenesis, TGFβ activation, worst relapse-free and overall survival
Transcriptomic (other)	Other transcriptomic classifications have been reported, including iCMS[21], PDS[22], CRIS[23] and CRPS[24]	

[1](Kuipers et al., 2015), [2](Lichtenstein et al., 2000), [3](Schoen et al., 2015), [4](Vasen et al., 1999), [5](Groden et al., 1991), [6](Monahan et al., 2020), [7](Bettington et al., 2017), [8](De Palma et al., 2019), [9](Crockett and Nagtegaal, 2019), [10](Chen et al., 2021), [11](Wu et al., 2019), [12](Maung et al., 2024) [13](Yan et al., 2021), [14](Loree et al., 2018), [15](Dekker et al., 2019), [16](Grady and Carethers, 2008), [17](Weisenberger et al., 2006), [18](Markowitz and Bertagnolli, 2009), [19](Campbell et al., 2017), [20](Guinney et al., 2015). [21](Joanito et al., 2022), [22](Malla et al., 2024), [23](Isella et al., 2017), [24](Nunes et al., 2024) [CIN=chromosomal instability, MSI=microsatellite instability, MSS=microsatellite stable, cIMP=CpG island methylator phenotype, MMRp/d=mismatch repair proficient/deficient, CMS=consensus molecular subtype, HNPCC=hereditary nonpolyposis colorectal cancer]

The above classification systems primarily reflect an epithelial perspective, however research into the tumour microenvironment (TME) is equally critical given its well-established influence on cancer pathogenesis, prognosis and treatment (Hanahan, 2022). The TME is a cancer-supporting ecosystem describing symbiotic signalling between malignancy and surrounding cells deep within an extracellular matrix (ECM) and is vital across all steps of tumour progression including metastasis (Hanahan, 2022). Targeting the TME has therefore emerged as an important therapeutic strategy. Over the past decade, expanding insight into the role of infiltrating and surrounding immune cells, and colon cancer immunology more broadly, has deepened our understanding of tumourigenesis and facilitated the emergence of alternative treatment modalities, particularly immunotherapy.

1.3 Colon cancer immunology

1.3.1 Foundations of tumour immunology

The role of the immune system in cancer was first proposed more than fifty years ago through the immuno-surveillance theory which suggested immune cells actively monitor, recognise and eliminate tumour cells, supported by the concept of tumours as ‘wounds that don’t heal’ (Burnet, 1957, Dvorak, 1986, Shankaran et al., 2001). Since then, the field of immuno-oncology (IO) has rapidly expanded across all cancer types, giving rise to several landmark concepts including:

(1) **the cancer-immunoediting model**, which describes dynamic tumour shaping (immunoediting) by immune cells, consisting of elimination, equilibrium and escape phases, and the ability of the immune system to both suppress and promote tumourigenesis on an immune cell-specific basis (Dunn et al., 2002, Dunn et al., 2006, Schreiber et al., 2011, Mittal et al., 2014);

(2) **immune contexture**, which characterises the local immune response by the distribution, density and functional state of immune cells within and surrounding the

tumour. This can be quantified through immunoscoring (*in situ* T-cell infiltrate) which is prognostic (Galon et al., 2006, Galon et al., 2012, Pages et al., 2018);

(3) **the cancer-immunity cycle and cancer-immune set-point**, which describes seven stepwise events from the release of cancer-derived antigen to cancer cell killing. Stimulatory factors must outweigh inhibitory factors to overcome the cancer-immune set point threshold to achieve effective cancer immunity (Chen and Mellman, 2017, Chen and Mellman, 2013);

(4) **immune-phenotype continuum**, which describes the classification of tumours into immune-inflamed, immune-excluded and immune-desert groups based on $CD8^+$ T-cell infiltration, cytokine expression and PD-L1/major histocompatibility complex (MHC) expression. These groups correlate with immunotherapy response and alter in prevalence across and within tumour types (Hegde and Chen, 2020, Gajewski, 2015, Hegde et al., 2016)

These concepts align with recognised cancer hallmarks, including tumour-promoting inflammation and immune evasion, highlighting the central role of cancer-immune interactions in tumour development (Hanahan, 2022). The work by Galon, Angell and Pages have specifically highlighted the impact of infiltrating immune cells in colon cancer. Quantifying immune contexture by measuring $CD3^+$ (all T-cell populations) and $CD8^+$ (cytotoxic T-cell populations) cells in the tumour core and invasive margins showed increased infiltration was significantly prognostic, even out-performing TNM staging alone (Pages et al., 2018). Given the importance of immunology in oncology, immune-based cancer therapies have rapidly advanced, and their full potential in colon cancer is only beginning to be realised.

1.3.2 Immune checkpoint modulators (ICMs) for colon cancer immunotherapy

Harnessing the immune system to potentiate anti-tumour activity, immunotherapy is a newly established pillar of precision cancer medicine and has revolutionised treatment of many

tumour types. There are multiple modalities but ultimately all converge on their action directly on cancer cells or indirectly by targeting immune cells (Waldman et al., 2020). Chimeric antigen receptor T-cells, cancer vaccines, oncolytic viruses and adoptive tumour-infiltrating lymphocyte (TIL) transfer are some key approved cancer immunotherapeutics, but immune checkpoint modulators (ICMs) have gained specific traction in colon cancer.

The original pioneering treatment that led the field, ICMs epitomise the unprecedented success of immunotherapy to impede tumourigenesis and still represent the dominant modality (Upadhaya et al., 2020, Saez-Ibanez et al., 2023). From potent tumour killing in mice (Leach et al., 1996) to improving survival in metastatic melanoma in a phase III trial (Hodi et al., 2010), the anti-CTLA-4 ICM ipilimumab remains mainstay first-line therapy for metastatic melanoma and has strikingly improved survival from 7 months to more than 7 years (Wolchok et al., 2025, Larkin et al., 2019). ICMs (otherwise known as immune checkpoint blockade) function to remove inhibitory signals of activation on T-cells, considered the ‘foot soldiers of the immune system’, and enhance anti-tumour T-cell response mediated through IFNγ, granzyme B and perforin release (Figure 1.2A; (Sharma and Allison, 2015, Zhang and Bevan, 2011)).

In colon cancer, two key ICM clinical trials prompted UK-NICE recommendation in 2021 of anti-PD-1 (nivolumab and pembrolizumab) and anti-CTLA4 (ipilimumab) agents alone or in combination (Figure 1.2B). Over the past couple of years, more clinical trials have led to further approvals depending on disease stage and previous treatment (Figure 1.2C). For example, nivolumab and ipilimumab combination ICM-therapy recently demonstrated superior progression-free survival (PFS) compared to nivolumab alone in CHECKMATE-8HW (Andre et al., 2025), reemphasising how trials continue to redefine treatment strategies in advanced colon cancer. However currently in the UK, immunotherapy is only approved for advanced unresectable or metastatic disease and offered to patients explicitly possessing mismatch repair-deficient (MMRd) and/or microsatellite instability-high (MSI) tumours, representing <5% mCC cases (NICE, 2025).

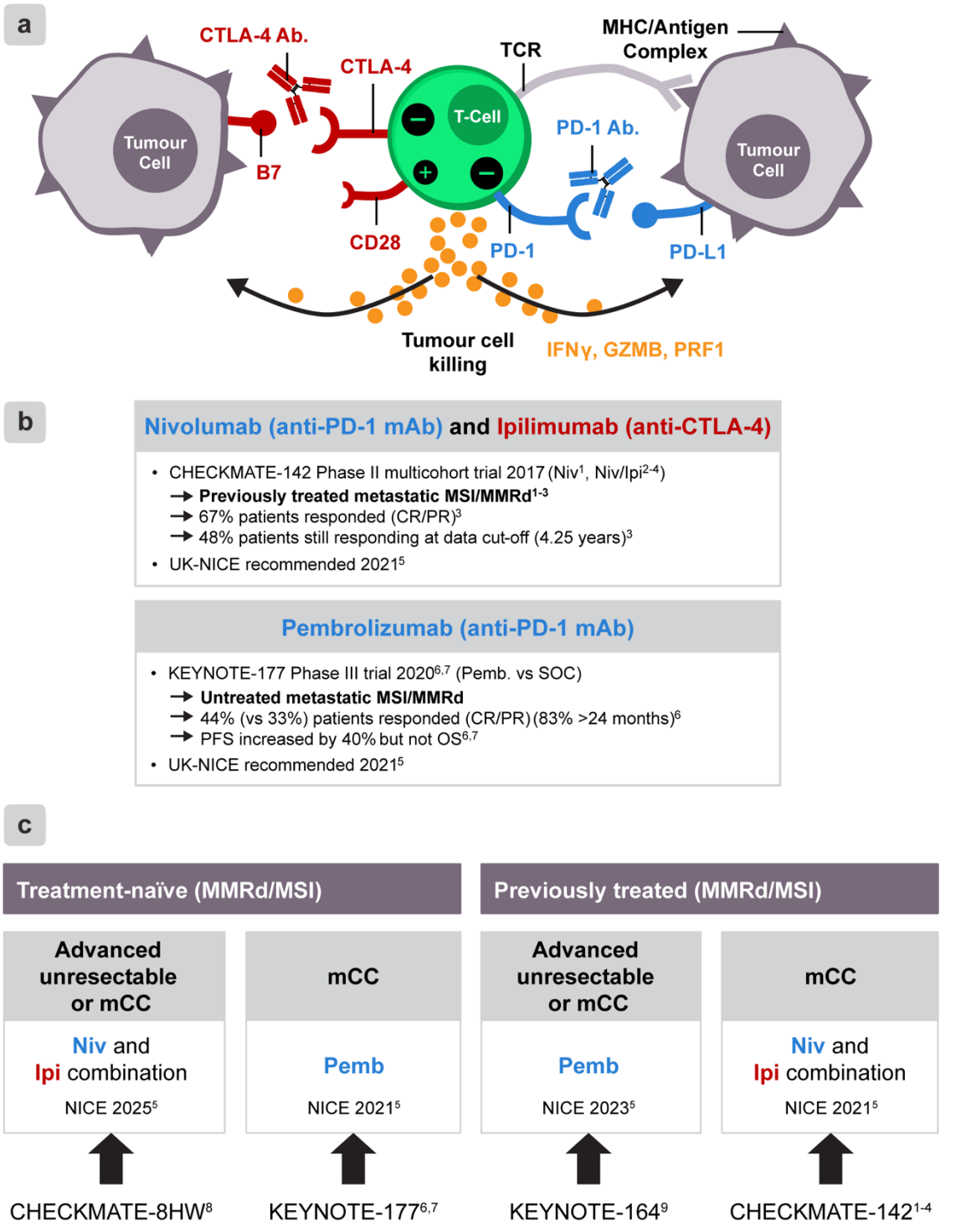


Figure 1.2 | The mechanism of action of approved immunotherapy in colon cancer highlighted with respective clinical trials. Nivolumab, Ipilimumab and Pembrolizumab are immune checkpoint modulators (ICMs) that act to remove inhibitory pathways of T-cell activation to upregulate anti-tumour immunity. These inhibitory pathways are mediated through CTLA4 and PD-1 receptors. *[1](Overman et al., 2017), [2](Overman et al., 2018), [3](Andre et al., 2022), [4](Lenz et al., 2022), [5](NICE, 2025), [6](Andre et al., 2020), [7](Diaz et al., 2022), [8](Andre et al., 2025), [9](Le et al., 2023). [OS=overall survival, PFS=progression-free survival, CR=complete response, PR=partial response, SOC=standard of care, mCC=metastatic colon cancer]*

Maintaining genomic fidelity, MMR proteins repair base-base mismatch and indel loop mutations formed following DNA replication (Bellacosa, 2001, Li, 2008). Malfunctioning following somatic or hereditary aberrations of key MMR machinery, MMRd drives spontaneous mutation rates 100 to 1000-fold and repetitive genomic regions including microsatellites expand in length – a process termed MSI (Hewish et al., 2010, Pecina-Slaus et al., 2020, Bonneville et al., 2017). MMRd is diagnosed clinically following confirmed loss of ≥1 MMR protein (MLH1/MSH2/MSH6/PMS2) by immunohistochemistry (IHC) staining on tumour biopsies (Bonneville et al., 2017). MMRd confirmation is sometimes followed or replaced with MSI testing by pentaplex polymerase chain rection (PCR), confirmed if ≥2 canonical microsatellite regions have expanded in size (Bonneville et al., 2017, Suraweera et al., 2002). Both MMRd and MSI largely overlap and inevitably describe hypermutated tumours of tumour mutational burden high (TMBH), with >10 mutations per megabase (Mb) (Campbell et al., 2017). Though not all TMBH tumours are MMRd/MSI, as POLE/POLD1 mutant tumours describe an ultra-mutated phenotype (>100 mutations per Mb) but remain microsatellite stable (MSS) (Campbell et al., 2017, Palles et al., 2013). Increased mutational load through such mechanisms is relevant for immunotherapy because it promotes enhanced mutated protein (neoantigen) production and surface expression via MHC (Le et al., 2017, Schumacher et al., 2019, Ma et al., 2018). This increases the likelihood of adaptive immunologic recognition, infiltration of cytotoxic immune cells and tumour cell lysis – termed the neoantigen hypothesis (Le et al., 2017). Immunotherapies such as ICMs rely on a pre-existing immune response, characterised by infiltrating T-cells and PD-L1 expression on tumour cells. Consequently, most drug approvals have been limited to MMRd, MSI or TMBH tumours where these biomarkers predict improved post-treatment survival across cancer types (Le et al., 2017, Le et al., 2015).

1.3.3 The tumour-immune microenvironment in colon cancer

The tumour-immune microenvironment (TIME) is highly complex and dynamic throughout colon cancer progression, reflecting continuous exposure to genetic and environmental insults and resulting chronic inflammation (Kennel and Greten, 2025). The immunosuppressive compartment includes myeloid-derived suppressor cells, regulatory T-cells, tumour-associated macrophages (TAMs) and cancer-associated fibroblasts, which are often organised around central TGFβ signalling (Ganesh et al., 2019, Kennel and Greten, 2025). Conversely, anti-tumour immune cells such as $CD8^+$ tumour-infiltrating lymphocytes (TILs), $CD4^+$ T-helper 1 cells, natural-killer (NK) cells and innate-like γδ-T-cells accompany core IFNγ signalling (Ganesh et al., 2019, Kennel and Greten, 2025). The multifaceted interplay between these cell populations ultimately determines the overall anti-tumour response, consistent with previously described IO models. As expected, single-cell RNA sequencing (scRNAseq) and flow cytometric studies have revealed compositional and spatial differences in the TIME between MMRp and MMRd colon cancers, namely an abundance of TILs, γδ-T-cells and NK cells in MMRd tumours which display gene expression programs characteristic of both T-cell activation and T-cell exhaustion (Becker et al., 2022, Pelka et al., 2021, de Vries et al., 2020).

However, this adaptive immune-rich microenvironment in MMRd tumours is not uniformly predictive of immunotherapeutic response. Similar techniques performed on MMRd colon tissue pre- and post-ICM immunotherapy reveals substantial heterogeneity within MMRd/TMBH subgroups in both TIME composition/functional states and response. For example, one study found systemic rather than intratumoural immunity correlated with treatment response, driven by circulating MHC class-II-related $CD8^+$ T-cells (Chen et al., 2024). Another study reported that responders exhibited intratumoural antigen-presenting macrophages engaged in direct interactions with $CD8^+$ T-cells (Bortolomeazzi et al., 2021). Interestingly, a spatial transcriptomics study found IFNγ signalling and T-cell induced CD74 expression on tumour cells and TAMs determined response to ICMs, and this was also

identified in some MMRp ICM-responders from the AtezoTRIBE cohort (Acha-Sagredo et al., 2025). This underscores that true tumour immunogenicity is neither solely dictated by mutational burden nor by the presence of the TIME alone.

These findings align with clinical observations showing substantial variability in immunotherapy response among TMBH colon cancers, with 50-60% patients deriving no objective response or survival benefit (Lenz et al., 2022, Diaz et al., 2022, Le et al., 2023). Although less common than in TMBH tumours, some TMBL patients have also demonstrated response in immunotherapy clinical trials (Chalabi et al., 2020). Whilst biomarkers such as TMB and $CD8^+$ T-cell infiltration have pioneered IO and the development of immunotherapies, it's increasingly clear that tumour immunogenicity is shaped by a broader and more complex set of determinants.

1.4 Defining tumour immunogenicity and its determinants

‘Tumour immunogenicity’ (TI) or ‘immunogenicity’ is the capacity of a tumour to elicit a pre-existing anti-tumour immune response that’s conducive to immunotherapy (Blankenstein et al., 2012, Medler et al., 2021). Tumours with high TI are considered immunogenic and typically more responsive to immunotherapy. Fundamentally, true TI is functionally realised when a patient derives clinical benefit from immunotherapy, reflecting the presence of an underlying, albeit supressed, immune response. This aligns with the core principle of immunotherapy: to reinvigorate an existing immune reaction rather than initiate one *de novo*, making the presence of a baseline immune response a key requirement of high TI. It’s important to note most literature inadvertently equates immunogenicity to antigenicity, a tumours ability to induce an immune response based on antigen recognition by cognate T-cell receptors (Wang et al., 2019). As discussed above, antigenicity is important but is not the sole contributor to TI.

Here, I define a new holistic perspective on TI. My review of the literature revealed that research on TI and immunotherapy can be distilled into a unifying conceptual framework (Figure 1.3). Many tumour intrinsic and extrinsic factors (Figure 1.3A), most of which can be seen in the literature presenting as biomarkers of immunotherapy response, can differentially contribute to five different abilities or multiple determinants of immunogenicity (mDOI) (Figure 1.3B). This includes the ability to (1) express antigen and (2) process and present antigen, forming what is traditionally known as antigenicity. Others comprise the ability to (3) attract an immune-destructive environment, (4) exclude an immune-suppressive environment and (5) succumb to immune-mediated killing, forming an overarching ‘immune infiltration and signalling’ descriptor. On a single-cell level, the degree of TI reflects the combined functional strength of these five abilities, such that increasing each contributes cumulatively, or even synergistically, to a heightened anti-tumour immune response and thus response to immunotherapy (Figure 1.3C).

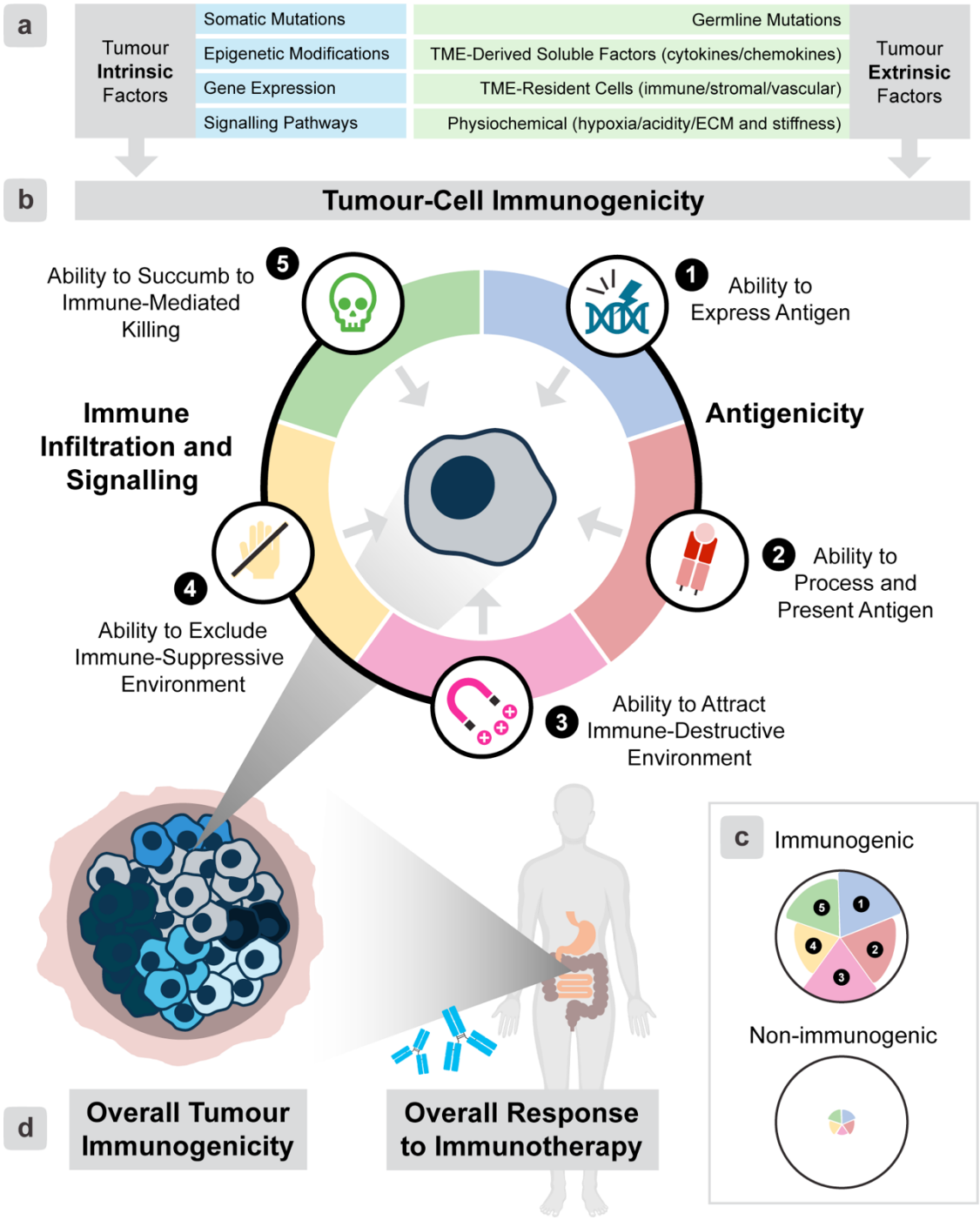


Figure 1.3 | Tumour immunogenicity (TI) and its multiple determinants. Many tumour intrinsic and extrinsic factors **(A)** contribute to each determinant of TI **(B)**, described by five abilities related to antigenicity and immune infiltration and signalling. TI is a result of the cumulative combination of each determinant **(C)**, describing either immunogenic or non-immunogenic tumours. Overall response to immunotherapy is determined by overall TI – the aggregate of all cellular TI **(D)**.

Paralleling intratumoural heterogeneity, overall TI can be conceptualised as the aggregate of immunogenic features expressed at the single-cell level across the tumour (Figure 1.3D). This concept can be applied pan-cancer and fits with the notions of immune escape/evasion and inherited or acquired resistance to immunotherapy. For example, somatic and/or germline mutations in MMR genes predominantly contribute towards a cells ability to express antigen, generating high TMB and thus likelihood of immune recognition as by the neoantigen hypothesis. As discussed earlier, mutational load has been a dominant focus in IO for the past decade. However, a cell also needs the ability to process and present this neoantigen. B2-microglobulin (B2M) is a key component of MHC class I (MHC-I) and crucial for antigen presentation, and its truncating mutations resulting in loss of MHC-I expression have been reported in intrinsic (Gurjao et al., 2019) and acquired (Zaretsky et al., 2016) resistance to ICM immunotherapy. Despite exhibiting high TMB, these patients didn't respond to immunotherapy, underscoring the importance of adopting a holistic view of TI in IO and highlighting its inherent complexity.

1.5 Expanding the clinical impact of immunotherapy in colon cancer

It's clear that for colon cancer, two critical clinical challenges remain profound. Firstly, current immunotherapy targeting is not optimal. This is highlighted by fewer than 50% MMRd/MSI/TMBH patients currently responding to ICM therapy (Lenz et al., 2022, Diaz et al., 2022, Le et al., 2023). Secondly, patient applicability remains limited, with <5% mCC exhibiting high mutational load and underscoring the small proportion of patients currently eligible for immunotherapy (NICE, 2025). To provide durable clinical benefit for more colon cancer patients, I propose two solutions which are becoming evident in the field (Figure 1.4). Firstly, *more* and *better* biomarkers delineating non-immunogenic and immunogenic tumours are required that would both avoid treating the wrong patient and treating patients which may respond but otherwise were ineligible. The future of IO will likely rely on a

composite biomarker approach, integrating multiple parameters rather than a single predictive marker (Sharma and Allison, 2015). Secondly, sensitising a tumour to become immunogenic prior to immunotherapy, a notion I term immunosensitisation, would enhance eligibility for immunotherapy even further.

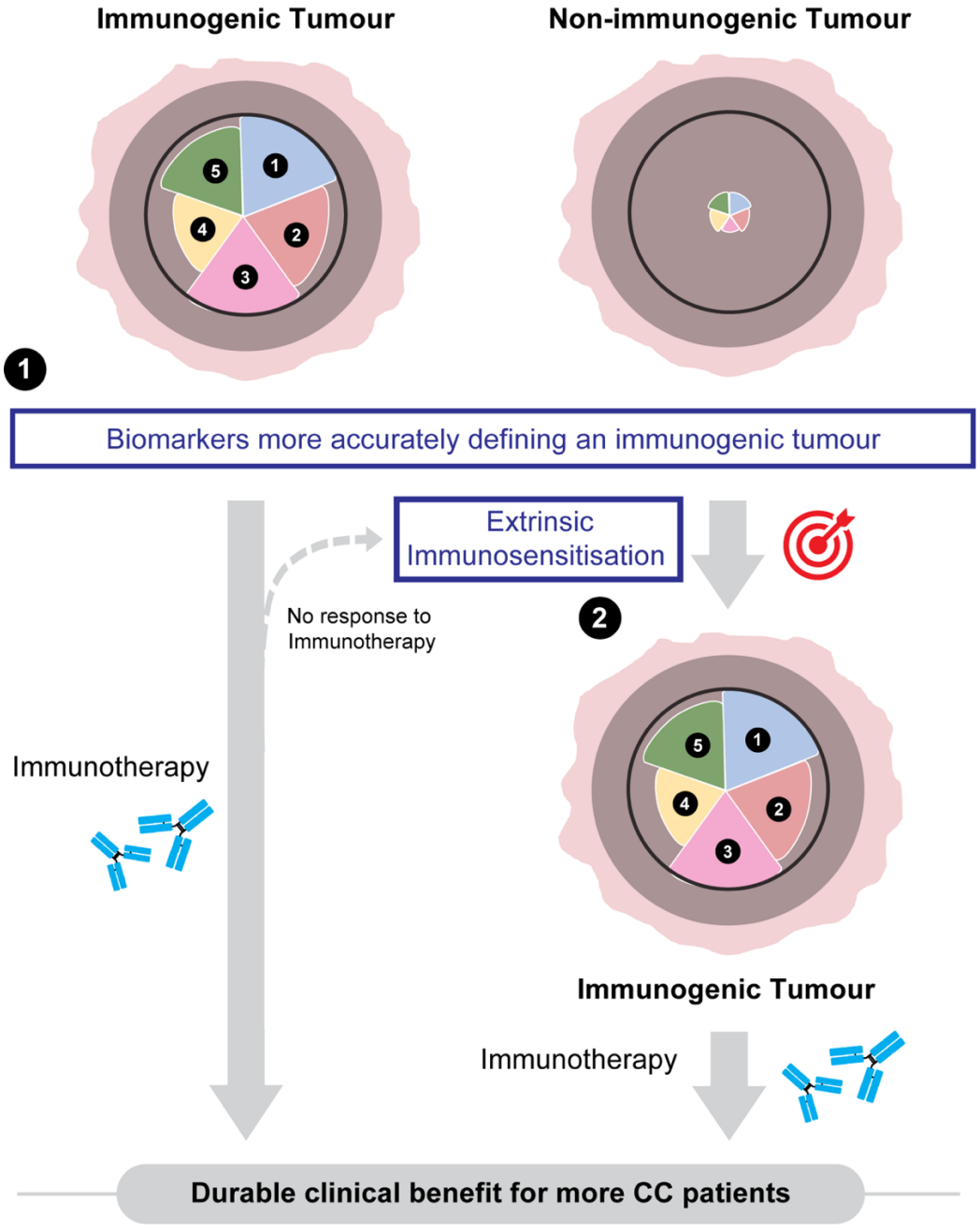


Figure 1.4 | Durable clinical benefit for more colon cancer (CC) patients. This could be achieved through (1) identification of biomarkers more accurately defining an immunogenic tumour and (2) extrinsic immunosensitisation where poorly-immunogenic tumours can be sensitised to immunotherapy. The multiple determinants of immunogenicity schematic has been included to represent immunogenic and non-immunogenic tumours.

Eloquently proposed by Sharma and Allison over a decade ago, the future of immunotherapy lies with immunosensitisation (Sharma and Allison, 2015). This is a concept that's largely described in the literature as 'turning a cold tumour hot', however this is rather simplistic considering our understanding that $CD8^+$ T-cell infiltrate is not solely responsible for TI (Galon and Bruni, 2019). Immunosensitisation describes the modulation of TI occurring intrinsically or extrinsically respective to neoplasia and is achieved through the presence or use of immunosensitisers. Extrinsic immunosensitisation currently dominates the field, with numerous combination immunotherapy trials aiming to sensitise MSS or MMRp tumours (Table 1.2). Once considered 'non-responders', it's now evident MSS/MMRp patients can achieve meaningful responses when treated with combination immunotherapy.

Table 1.2 | Extrinsic immunosensitisation of MSS/MMRp colon cancers

Extrinsic immunosensitiser	Rationale	ICM	Patients	Response
Regorafenib	Multi-kinase inhibitor to target VEGF: vasculature normalisation and enhancement of immune infiltration, decreased TAMs	Niv	Previously treated advanced/ unresectable mCC	OR 7%[1]
		Niv		OR 33%[2]
		Sintilimab		OR 21.4%[3]
		Ipi + Niv		OR 36.4%[4]
Botensilimab	A novel CTLA4 ICM, boosts NK cells and depletes Tregs	Balstilimab		OR 17%[5]
				OR 60%[6]
Chemotherapy and Bevacizumab	Induction of immunogenic cell death and antigen release, VEGF targeting as above	Atezolizumab	Previously untreated advanced/ unresectable mCC	30% reduced risk of progression[7]
Celecoxib	COX-2 inhibitor: inhibition of PGE_2 synthesis and tumour-promoting inflammation	Ipi + Niv	Previously untreated, Stage I-III	OR 27% with Ipi + Niv only (2 patients responded to celecoxib)[8]

Encorafenib and Cetuximab	Encorafenib inhibits $BRAF^{V600E}$, Cetuximab targets EGFR feedback, less pro-tumour environment	Niv	$BRAF^{V600E}$, previously treated advanced/ unresectable mCC	OR 50%[9]
Cobimetinib and Regorafenib	Cobimetinib is a MEK1/2 (MAPK pathway) inhibitor to enhance $CD8^+$ infiltration, VEGF targeting with Regorafenib as above	Atezolizumab	Previously treated advanced/ unresectable mCC	OR <3%[10]

[1](Fakih et al., 2023), [2](Fukuoka et al., 2020), [3](Liu et al., 2025), [4](Xiao et al., 2024), [5](Bullock et al., 2024), [6](Hissong et al., 2025), [7](Antoniotti et al., 2022), [8](Chalabi et al., 2020), [9](Morris et al., 2025), [10](Eng et al., 2019) [OR=objective response where tumour shrinks or disappears i.e.>50% regression, Ipi=ipilimumab, Niv=nivolumab, OS=overall survival, ICM=immune checkpoint modulator, mCC=metastatic colon cancer, TAM=tumour-associated macrophage]

Although immunotherapy was initially developed for late-stage (IV) metastatic disease, reflecting limited therapeutic options for these patients, recent advances have shifted toward early-stage (I-III) non-metastatic indications and emphasise the strong association between tumour stage and immunogenicity. Notably, early-stage MMRd tumours demonstrate exceptional sensitivity to ICM-immunotherapy, with complete response rates 50-100% in colon (Jiang et al., 2024) and rectal (Cercek et al., 2022) cancer cohorts. Striking responses have also been observed with early-stage MMRp tumours following neoadjuvant immunotherapy-combination treatment in colon (Chalabi et al., 2020) and rectal (Hanna et al., 2021) patients. The latter PRIME-RT trial demonstrated 67% complete response at 6 months following immunotherapy-combination treatment with short-course radiotherapy, emphasising the potential for organ-preserving treatment strategies (ESTRO, 2025, Hanna et al., 2021). Together, this exemplifies clear accumulating evidence that early-stage tumours possess heightened baseline immunogenicity, further amplified in tumours with high TMB, comparative to late-stage tumours. Recent evolutionary work in both lung (Hessey et al., 2025) and colon adenocarcinoma (Lakatos et al., 2025) suggest immunogenicity-related processes such as immune evasion are acquired early during tumour development, preceding invasion and metastatic dissemination. Furthermore, $CD8^+$

TILs are less frequent in late-stage colon tumours (Galon et al., 2006, Mlecnik et al., 2011) and display signatures of reduced proliferation and activation (Saleh et al., 2020). Late-stage colon tumours exhibit diminished immune-related adaptive and cytotoxic transcriptional signatures comparative to early-stage colon tumours, consistent with an inverse relationship between tumourigenicity and immunogenicity (Mlecnik et al., 2016). This suggests fundamental stage-dependent differences in tumour immunogenicity exist, require further investigation and should be considered when investigating immunogenicity. Collectively, these advances highlight a hopeful trajectory toward broadening immunotherapy eligibility across colorectal cancer subtypes and disease stages.

1.6 Investigating somatic mutational drivers of immunogenicity

Central to both proposed solutions enabling durable benefit for more colon cancer patients is a deeper understanding of the mDOI. Investigating factors that shape these mDOI, or more broadly uncovering drivers of immunogenicity, represents a major focus in the field. In colon cancer, identification of such drivers has primarily relied on *in silico* analyses which are subsequently validated through *in vitro* and *in vivo* experimentation (Acha-Sagredo et al., 2025, Wang et al., 2022, Li et al., 2025). Most of these studies are transcriptomic biomarker-centric, however recently there has been interesting work looking at mutation quality over quantity, and how this could influence TI and immunotherapy response heterogeneity. For example, in MMRd/TMBH colon cancers, some mutations have been associated with enhanced anti-tumour immune response, whereas others not (Wang et al., 2023). This suggests a mutation may be associated with anti-tumour immunity and immunotherapy response through its alteration of gene function, rather than generation of neoantigen (Porta-Pardo and Godzik, 2016). Such findings underscore the need to identify and characterise mutations that reshape immune-related pathways, expanding the concept of immunogenicity beyond the neoantigen paradigm.

Previous studies have mechanistically explored how prevalent mutations in cancer driver genes associate with TI, such as KRAS gain-of-function (Liao et al., 2019, Zhou et al., 2024) and APC loss-of-function (Wu et al., 2024) mutations. However, a more systematic and exploratory approach that considers the mDOI may uncover novel insights including mutations not previously linked to TI before. Importantly, a gene may not contribute to tumourigenesis yet still play a critical role in shaping tumour immunogenicity. An investigation particularly in the MMRp/MSS/TMBL setting would not only eliminate the confounding of high mutational load in such exploratory analyses, but be of clinical relevance, due to the substantial evidence that immunogenic subsets exist within. I hypothesise that specific mutations in colon cancer can modulate tumour immunogenicity by altering gene function, thereby conferring clinically relevant immunogenicity even in traditionally 'poorly immunogenic' TMBL tumours.

1.7 A brief thesis overview

Chapter 2 is set out as a non-chronological summary of the materials and methods used throughout this thesis. Each following results chapter begins with individual hypotheses and aims in respective 'Experimental design' subsections for clarity.

To address the knowledge gap outlined in this introduction, **Chapter 3** presents a bioinformatic investigation to identify novel somatic mutational drivers of tumour immunogenicity. Two complementary multiomic analyses of the Cancer Genome Atlas (TCGA) were employed to (1) infer immunogenic subgroups and (2) assess mutational enrichment between these groups. This approach identified the transcription factor zinc finger homeobox 4 (*ZFHX4*) as a strong candidate for further investigation, based not only on its mutational enrichment in TMBL immunogenic tumours but its reported additional roles in tumourigenesis and a notable absence of prior study in colon biology. Permitting a parallel

investigation of tumourigenicity and immunogenicity was an appropriate measure considering the limited tools available to investigate immunogenicity *in vitro*.

Building on this rationale, **Chapter 4** explores ZFHX4's role in colon tumourigenesis, hypothesising a pro-tumourigenic biology as its high expression confers poor prognosis. Following a comprehensive review of its structure, regulation and reported biological roles in healthy, disease and cancer contexts, *ZFHX4* expression was characterised across normal and malignant tissues using bulk and single-cell RNA sequencing alongside immunohistochemistry. CRISPR-based *ZFHX4*$^{-/-}$ knockout organoid and cell line models were generated and extensively validated at all central dogma levels. This enabled functional assessment across key cancer hallmarks including proliferation, cell cycle dynamics, cytokine secretion, angiogenesis and migration and invasion. While phenotypic heterogeneity limited definitive conclusions, integrated analysis of additional multiomic datasets supported a model in which ZFHX4 interacts with the epigenetic nucleosome remodelling and deacetylase (NuRD) complex to regulate pro-tumourigenic attributes including epithelial-mesenchymal transition, angiogenesis, stemness and Wnt signalling.

To explore a role in immunogenicity, **Chapter 5** first begins with an overview of *ZFHX4* (missense) mutations in the same four multiomic cohorts as Chapter 4. Employing the same previously generated *ZFHX4*$^{-/-}$ models as surrogates for missense mutations, conditioned medium was respectively derived and cultured with healthy donor pan-T-cells. Flow cytometry and Transwell migration assays enabled assessment of T-cell activation and migration in response to this medium, and *ZFHX4*$^{-/-}$ was found to enhance an IFN-related phenotype in $CD8^+$ T-cells.

Finally, **Chapter 6** integrates these findings, proposing that ZFHX4 functions as a non-canonical oncogene exhibiting dosage-sensitive, context-dependent and potentially haploinsufficient behaviour, with dichotomous roles in colon tumourigenicity and immunogenicity.

Chapter Two: Materials and methods

2.1 Cell culture

2.1.1 Cell lines

Cell lines (Table 2.1) were kindly provided by respective University of Oxford research groups and cultured in medium with supplements accordingly. Cells were subcultured when 80-90% confluency was reached, dissociating adherent cells with a Dulbecco's Phosphate Buffered Saline (DPBS; Sigma, #D8662) wash and 5 minute 0.05%Trypsin-EDTA (Gibco, #25300-054) incubation. For suspension cells, at least a 10-fold dilution was performed. Washing and spinning of cells were performed with Heraeus Megafuge 16R Centrifuges and Eppendorf 5425R Microcentrifuges (300g, 5 minutes).

Table 2.1 | Immortalised human cell line derivations with their respective media bases

Cell Line and RRID	Disease	Morphology	Form	Research Group	Media Base	Supp.
COLO320DM (CVCL_0219)	Colon Carcinoma (MSS)	Epithelial	Ad	Bodmer	DMEM/F-12 (Gibco, #11320033)	10% FBS 1% PS
NCIH716 (CVCL_1581)			Susp			
SKNMC (CVCL_0530)	Ewing Sarcoma	Mesenchymal	Ad	Hassan		
Jurkat (Clone E6-1, CVCL_0367)	Acute Lympho-blastic Leukaemia	T lymphoblast	Susp	Parkes	RPMI-1640 (Gibco, 11875093)	10% FBS 1% PS 50μM βm
HUVEC (pooled donors, #C2519A)	Normal (umbilical)	Endothelial	Ad	Buckley	EBM-2 (Lonza, #CC-3156)	EGM-2 (Lonza, #CC-4176)

Supp.=supplements, ad=adherent, susp=suspension, FBS=heat-inactivated foetal bovine serum (FBS) (Gibco, #16140-071), PS=penicillin-streptomycin (Gibco, #15140-122), βm=2-β-mercaptoethanol (Gibco, #31350-010)

All lines were maintained in a 37°C and 5% CO_2 humidified incubator (Binder C160, #16-03944) and routinely tested negative for Mycoplasma with the LookOut Mycoplasma Detection Kit (Sigma, #MP0035) and Jumpstart Taqman (Sigma, #D9307) as by manufacturer's instructions. STR profiling was performed with NorthGene (Deeside, UK),

achieving >95% concordance with online databases. Maintaining stock, cell lines were frozen in Liquid Nitrogen in 10% DMSO-supplemented (Santa Cruz, #sc-358801) media. Microscope images were taken on the EVOS-M5000 (ThermoFisher, #AMF5000). Live counts were recorded in duplicate and averaged using Trypan Blue exclusion (Gibco, #15250061) and the LUNA-II cell counter (Logos Biosystems, #L12001,40002).

2.1.2 Human tissue collection

All gastrointestinal tissue samples were obtained from adult patients undergoing resective surgery for Colorectal Cancer (CRC) at the Churchill Hospital, Oxford University Hospitals NHS Foundation Trust. Written consent was obtained from all patients and ethics were approved under the University of Oxford, Translational Gastroenterology Unit (21/YH/0206). ‘Normal colon’ refers to tissue taken at least 5cm distant from the tumour and appearing macroscopically normal, ‘normal small intestine’ refers to healthy tissue taken from the small intestine and ‘tumour’ refers to resected tumour tissue. Occasionally these were provided from the same patient during right hemicolectomies to obtain matched specimens. Metastatic tumour samples were obtained from patients undergoing liver resections for CRC liver metastases. Omental or mesenteric fat was removed from all samples before washing thrice in DPBS to remove faecal debris and contaminants.

2.1.3 Organoid culture

Organoid derivation and culture were followed as described by a published protocol (Chan et al., 2023). Previously established metastatic tumour lines were kindly provided by Dr Scott Collins (Buczacki Group, University of Oxford) and these were maintained in ENRAS media (Table 2.2). R-spondin conditioned medium (CM) was generated by the Buczacki Group from 293T-HA-Rspol-Fc cells following previously published protocol (Chan et al., 2023). Basal media constituted Advanced DMEM/f-12 (Gibco, #12634010) with 1% GlutaMAX (Gibco, #35050061) and 1% HEPES (Gibco, 15630049).

Table 2.2 | ENRAS supplements added to basal media to culture metastatic tumour organoids

Ref.	Reagent	Stock Conc.	Final Conc. in Basal Media	Supplier and Cat. No.
R	R-spondin CM	-	20%	na
E	EGF	1μg/μl	50ng/ml	Peprotech, #AF-100-15
N	Noggin	1μg/μl	50ng/ml	Peprotech, #120-10C
A	SB431542 (ALK5i)	5mM	500nM	Bio-Techne, #1614
S	SB202190 (p38i)	10mM	3μM	Bio-Techne, #1264
	Nicotinamide	1M	10mM	Sigma, #240206
	N-acetylcysteine	0.5M	1.25mM	Sigma, #A0737
	B27-plus	50X	1X	Gibco, #A3582801
	PGE2	10μM	10nM	Bio-Techne, #2296
	Y27632 (only upon passage)	10mM	10μM	Bio-Techne, #1254

Organoids were grown in 40μl domes of Cultrex Reduced Growth Factor Basement Membrane Extract Type 2 (RGF-BME; Biotechne, #3533-005-02) and passaged roughly once a week using TrypLE Select (Gibco, #12563029) for single-cell dissociation.

2.1.4 Primary T-cell culture

Frozen lymphocytes (B/T/NK-cells) were obtained with thanks to the Seymour Group (Department of Oncology, University of Oxford) which were processed from healthy blood donations (NHS Blood and Transplant, REC 19/LO/1848). Lymphocytes were thawed 24 hours prior to pan-T-cell isolation and cultured in pre-warmed media equating to Jurkat media above without Pen-Strep in standing T25 flasks. Cells were washed in sterile-0.22μM filtered (Sarstedt, #83.3940.501) MACS buffer (500ml DPBS; Gibco, #14190094, 2mM EDTA; ThermoFisher, #15575020, 0.5% Bovine Serum Albumin (BSA); Bio-techne, #5217) and centrifuged (450g, 5 minutes). Counting was performed with a manual haemocytometer and demonstrated around $30x10^6$ cells per vial. Proportional to $1x10^7$ cells, negative pan-T isolation was performed according to Miltenyi Biotec protocol using the Pan T-Cell Isolation Kit (#130-096-535), LS Columns (#130-042-401), MACS Multistand (#130-042-303) and QuadroMACS Separator (#130-090-976). Pan-T-cells were recounted and demonstrated

around 50% of pre-isolation values before using in appropriate experiments in either a non-activated or activated format.

To enable T-cell activation by artificial receptor stimulation unless otherwise specified, CD3/CD28 Dynabeads (Gibco, #11131D) were washed in DPBS on a DynaMag-2 Magnet (Invitrogen, #12321D) and used at a 1:1 cell ratio in T-cell media. 30U/ml recombinant human IL-2 (StemCell Technologies, #78036.3) was added to aid proliferation and these cells were plated in relevant suspension plates at densities of 1 x 10^6 cells/ml per 3.2cm^2 surface area. Following a 3-day 37°C incubation, bead:T mixtures were harshly washed using a P1000 pipette and the DynaMag to remove beads. Non-activated (NA) and activated (A) T-cells were prepared per blood donor and used for respective experiments.

2.2 Immunohistochemistry (IHC)

2.2.1 Tissue preparation, embedding and sectioning

Following washing as previously described, tissues were fixed in 10% Formalin (Avantor, #11699404) for 24 hours at room temperature (RTP) in 50ml falcons. After 3 washes with DPBS, tissue was then submersed in 70% Ethanol (EtOH)/distilled water (dH_20) (JTBaker, #15568454) and kept at 4°C indefinitely. Tissue processing, wax embedding and sectioning was performed using equipment at IHC facilities within the Department of Oncology and Ludwig Institute in collaboration with Professor Xin Lu and Dr Brittany-Amber Jacobs. Tissues were processed overnight on a Thermo Scientific Excelsior AS Tissue Processor and embedded using the Thermo Scientific HistoStar in paraffin blocks. Embedding orientation was performed with help from Dr James Bundred to position epithelium and stroma in an appropriate fashion for sectioning. 6μm sections were cut with a Rotary Microtome-7010 and placed on 40°C-heated water before attaching to SuperFrost slides (Appleton Woods, #MS526) and drying overnight at 37°C ensuring secure attachment.

2.2.2 Tissue staining

Sections were dewaxed with two 5-minute washes in 100% (v/v) xylene (Sigma, #534056), followed by rehydration through a graded series of industrial methylated spirit (IMS) washes (100%, 100%, 90%, 80%, and 70% (v/v) in MilliQ water, 2 minutes each). Sections were then rinsed in MilliQ water for 5 minutes. Antigen retrieval was performed by boiling slides in 10mM citrate buffer (Sigma, #C9999) at 110°C for 3 minutes using an antigen decloaking oven (Biocare, #DC2002) and then cooling to RTP before washing 3 x 5 minutes with DPBS. Endogenous peroxidase activity was blocked with 1% H_2O_2 (Sigma, #H1009) for 10 minutes at RTP and slides were rinsed in MilliQ water for 5 minutes. Hydrophobic rings around sections were drawn with an ImmEdge Pen (VectorLabs, #H4000) and evaporation during subsequent incubations was mitigated using an enclosed humidity chamber. Permeabilisation was performed with 0.2% Triton X-100 (Sigma, #9036-19-5) in DPBS for 10 minutes, RTP. Sections were blocked with 10% normal serum (matched to the secondary antibody host species) diluted in 1% BSA in DPBS and incubated (1 hour, RTP). Primary antibodies (Table 2.3) were applied in 1% BSA and incubated overnight at 4 °C. Slides were then washed in DPBS (3 x 5 minutes). Species-specific biotinylated secondary antibodies (Table 2.3) in 1% BSA were added for 2 hours, RTP followed by additional DPBS washes.

Table 2.3 | Primary and secondary antibodies used for IHC

Ab ref.	Target*	Host Species	Clonality	Supplier and Cat. No.	Dilution
Primary Antibodies					
PA5	ZFHX4	Rabbit	Poly	Thermo Fisher, #PA5-99757	1:100
abc	ZFHX4	Rabbit	Poly	Abcam, #ab254654	1:100
HPA	ZFHX4	Rabbit	Poly	Sigma, #HPA023837	1:100
M11	ZFHX4	Mouse	Mono (2B12)	Abnova, #H00079776-M11	1:100
M02	ZFHX4	Mouse	Mono (5E6)	Abnova, #H00079776-M02	1:100
V	Villin	Mouse	Mono (3E5G11)	Abcam, #ab201989	1:100
B	ß-catenin	Rabbit	Poly	Abcam, #ab6302	1:100
Secondary Antibodies**					
Anti-mouse IgG(H+L)		Horse	unknown	VectorLabs, #BA-2000	1:400
Anti-rabbit IgG(H+L)		Goat	unknown	VectorLabs, #BA-1000	1:400

**all target species are anti-human, **all secondary antibodies are biotinylated*

Slides were incubated with VECTASTAIN® ABC reagents (VectorLabs, #PK-6100) in DPBS for 1 hour at RTP and washed again in DPBS. Peroxidase activity was visualised using DAB reagent (Agilent, #K346811-2) which was incubated for precisely 1 minute on every section and then washed in dH_2O for 10 minutes. Nuclei were counterstained with Haematoxylin (Abcam, #ab220365) for 20 seconds and rinsed in running tap water for 5 minutes. Sections were dehydrated through IMS washes (70%, 90%, 100%, 100%, 2 minutes each), cleared with two 2-minute 100% xylene washes and mounted with Entellan Rapid Mounting Medium (Merck, #107960) on SuperFrost slides. Imaging was performed on the Olympus-BX43 microscope with EP50 View Camera attachment.

2.3 Transcriptomic profiling

2.3.1 RNA extraction

For cell lines, cells were passaged and pellets were washed once in DPBS. RNA was extracted using the Qiagen AllPrep DNA/RNA/Protein Mini kit (Qiagen, #80004). For organoids, RNA extraction was performed similarly but directly in the 24-well plate on ice for complete breakdown of RGF-BME and retainment of as much sample as possible. Quantification was performed using the NanoDropOne (Thermo Fisher Scientific, #ND-ONE-W), enabling normalisation to around 2µg.

2.3.2 RNA sequencing (RNAseq) and analysis

At least 500ng and 20ng/ul RNA with A260/280 values between 1.8-2.2 OD were submitted for RNAseq to Genewiz (Abingdon, UK) using the Illumina platform and standard paired-end protocol with 150bp read length, PolyA selection for rRNA removal, 50 million reads per sample and an ERCC spike-in (Invitrogen, #4456740). Raw FASTQ files were processed by Dr Amit Mandal (Buczacki Group, University of Oxford) and then further analysed using R-programming as outlined in Section 2.10.

2.4 Functional profiling

2.4.1 Cell line-based

2.4.1.1 Viability and proliferation

CellTitre-Glo (CTG) 2.0 (Promega, #G9243) was employed to measure cell viability and proliferation and brought to 22°C before use. For quantification of viable cell number (VCN), equal volumes of CTG were added to 30-minute-rested cell suspensions in 96-well black optical plates (ThermoFisher, #165305), incubated (30 minutes, RTP), shaken (600rpm, Heidolph Platform Shaker, 5 minutes) and then rested in the dark (25 minutes, RTP). Luminescence was read on the POLARstar Omega Plate Reader (BMG Labtech) and values were calculated by subtracting an average background luminescence from a media-only control. For measurements of proliferation rates, VCN was calculated across timepoints and doubling times were calculated as described in Supplementary Methods 2.1.

2.4.1.2 Cell cycle dynamics

To investigate cell cycle dynamics, the ClickiT-EdU AF488 Kit (Invitrogen, #C10337) and AttuneNxT Flow Cytometer (Invitrogen, #A24858) were employed as by manufacturer's instructions. Cells were seeded at 110k in 24-well plates and after 48 hours, EdU was incubated (10μM, 1 hour). FxCycle Violet (Invitrogen, #F10347) was added as the DNA stain post-fixation. Unstained and single-stained EdU controls were provided alongside samples (n=1). Flow cytometry was performed on the AttuneNxT Flow Cytometer and Autosampler (V4/B3/Y4/R3, Invitrogen, #A24858) and data was analysed using FlowJo-v10.10.0 software (BDBiosciences, Oregon, USA).

2.4.1.3 Cytokine arrays

To investigate the cell secretome, Human XL Cytokine Profiler Arrays (Bio-techne, #ARY022B) were used as by manufacturer's instructions. Generation of conditioned medium (CM) was optimised throughout (Table 2.4). Despite differences between experimental setups, within each experiment a knockout (KO) included its respective wild-type (WT) and thus fold changes could be compared across different experiments. CM was spun at 1000g for 10 minutes before incubating overnight on the arrays. Experiments which included a cell count to control for cell number differences involved scaling CM volume between 200-500μl to allow normalised loading onto the arrays.

Table 2.4 | Optimisation of cytokine array method

Exp. Code	# arrays	Plate (w)	Media vol. (μl)	Inc. time (d)	Cell density (k)	ONT rest with media replacement	Count norm.	Stronger chemi reagent
A	4	6	2000	5	300	N	Y	N
B	4	24	1000	5	50	Y	N	N
C	8	24	1500	3	200	Y	Y	N
D	8	24	750	3	200	Y	Y	Y
E	8	24	750	3	200	N	Y	Y

ONT=overnight, Y=yes, N=no

Dual membrane imaging was performed using the ChemiDocXRS+ system (Bio-Rad) and its "Chemi High Sensitivity" automated exposure program was employed to reduce background noise (positive control array spots allowed consistency between array imaging). Protein signal was quantified as by Supplementary Methods 2.2.

2.4.1.4 Conditioned medium generation

Passage A (current passage) and Passage B (thawed cells from a 5-month prior passage) cells were used as biological replicates (if not available for some clones, technical replicates of Passage A were used instead). $2x10^6$ cells were seeded into T25s containing antibiotic-free media supplemented with 10% or 5% FBS. Control blank CMs were also prepared

(media only, no cells). Following a 3-day conditioning, media was extracted, centrifuged (1000g, 10 minutes), aliquoted and frozen at -80°C. Cells were also counted in all conditions to allow normalisation depending on the experiment. Fibroblast CM was generated and kindly provided by Dr Kirsty Rooney (Buczacki Group, University of Oxford) from healthy colonic fibroblasts derived from normal colonic tissue. This was a 5-day conditioning in Fibroblast Growth Medium (PromoCell, #C-23110) and aliquots consisted of two different timepoints from the same donor line.

2.4.1.5 Migration and invasion

For the Transwell assays, 8μm Polyethylene Terephthalate (PET) Transwell Inserts (Millipore, #PCEP24H48) were used. Bottom compartments (BCs) were filled with serum-free media (SFM) or various chemoattractants such as 10%/20% FBS and Fibroblast CM. Cells were washed with DPBS, resuspended in SFM and seeded (100k cells/200μl media) per top compartment (TC). For adherent cells, TC media and cells were gently aspirated and inserts were washed once with DPBS. A cotton swab was used to remove non-migrated/invaded cells from the apical insert side. Cells were fixed (ice-cold 100% methanol (Sigma, #34860), 20 minutes) and washed with DPBS. Inserts were stained (0.5% Crystal Violet (Pro-Lab, #PL.8000), 10 minutes, RTP) before rinsing in tap water and air-drying overnight. Membranes were excised using a scalpel and mounted cell-side face-down on SuperFrost microscope slides with mounting medium and VWR 1.0 coverslips (Avantor, #631-0124). Five fields of view were imaged per membrane using the Olympus-BX43 at 10X magnification. Membrane coverage was inferred using image thresholding and pore subtraction on Fiji and macros were written separately for low and high confluent images, available on my GitHub repository (Flower, 2025). For suspension cells, cells in BCs were submitted for CTG-VCN as by Section 2.4.1.1. For the 3D Matrigel Droplet invasion assay, RGF-Matrigel (Corning, #354320) was thawed on ice and cells were seeded (50k/dome) in 100% Matrigel as by a previous protocol (Aslan et al., 2021).

2.4.2 Angiogenesis-based

HUVEC *in vitro* tube formation assays (IVTFAs) were based off a previously published protocol (Arnaoutova and Kleinman, 2010). Specialised chamber slides (ibidi, #81506) were coated with 11μl RGF-Matrigel (1 hour, 37°C) and HUVECs were passaged, washed and counted in basal EBM-2 media. HUVECs were prepared in 10k cells/50μl concentrations in different CMs and controls. Positive controls were human recombinant basic Fibroblast Growth Factor (bFGF; Peprotech, #100-18B) and Vascular Endothelial Growth Factor A ($VEGFA^{165}$; Peprotech, #100-20-10UG) at different concentrations in basal EBM-2 media. Negative controls included the pleiotropic angiogenesis inhibitor Suramin Salt (Sigma, #574625-50MG) at various concentrations in complete EGM-2 media. HUVECs were plated (10k/well) in 50μl and incubated (12 hours, 37°C). Calcein AM dye (Invitrogen, #C3100MP) reconstituted in DMSO was added for staining (3μM, 30 minutes, 37°C). Wells were imaged on the EVOS-M5000 at 1.25X magnification to capture the whole well. Fiji and the "Angiogenesis Analyzer" macro quantified angiogenesis with numerous parameters (Carpentier et al., 2020, Schindelin et al., 2012, Flower, 2025). 1.25X images were cropped and background subtracted (50-rolling radius) before running the 'Phase Contrast' angiogenesis programme with adjusted settings specified in my GitHub repository. All data points were divided by total pixel area and multiplied by $1x10^6$ to get units per million pixels.

2.4.3 Immune-based

For migration-based experiments, 3μm PET 24-well Transwell Inserts (Millipore, #PCSP24H48) or 3μm Polycarbonate 96-well Transwell Inserts (Corning, #3386) were used following Section 2.4.1.5 protocol. In some experiments, human recombinant CXCL12 (Biotechne, #6448-SD-025) and FBS were used as positive controls in BCs. Migrated T-cells were then subjected to CTG-VCN to quantify migration rate. For flow-based experiments, NA and A T-cells were incubated in different CMs in 96-well plates for various durations. Some experiments performed activation with beads during this incubation and

some were performed following standard 3-day activation. At the end of the experiment, cells were prepared for flow cytometry as described in Section 2.9.

2.5 CUT&RUN

Due to limited publications trialling CUT&RUN on ZFHX4, two different kits were used for different sets of antibodies on two separate runs (Table 2.5). Both kits were run as by manufacturer's instructions, with the incorporation of some EpiCypher antibodies in the CST run. Both run's positive (H3K4me3) and negative (IgG) controls contained a spike-in of SNAP-CUTANA K-MetStat Panel (EpiCypher, #19-1002) which was added just before primary antibody at a dilution of 1:100. Whilst both kits were run on the same 4 WT lines, the CST kit required 100k cells and EpiCypher kit required 500k cells.

Table 2.5 | Positive control, negative control and ZFHX4 antibodies used for CUT&RUN

Ab ref.	Target	Host Species	Clonality	Supplier and Cat. No.	Dilution for ONT incubation
RUN ONE with CST Kit (#86652), with yeast spike-in					
PA5	ZFHX4	Rabbit	Poly	Thermo Fisher, #PA5-99757	1:100
abc	ZFHX4	Rabbit	Poly	Abcam, #ab254654	1:100
M11	ZFHX4	Mouse	Mono (2B12)	Abnova, #H00079776-M11	1:100
M02	ZFHX4	Mouse	Mono (5E6)	Abnova, #H00079776-M02	1:100
CST-IgG	n/a	Rabbit	Mono (DA1E)	CST, #66362	1:20
Epi-H3K4me3	H3K4me3	Rabbit	Mono (2909-3D7)	EpiCypher, #13-0060	1:40
RUN TWO with EpiCypher Kit (#14-1048), with E.coli spike-in					
HPA	ZFHX4	Rabbit	Poly	Sigma, #HPA023837	1:50
Epi-IgG	n/a	Rabbit	Poly	EpiCypher, #13-0042	1:50
Epi-H3K4me3	H3K4me3	Rabbit	Mono (2909-3D7)	EpiCypher, #13-0060	1:50

DNA fragments were submitted to Genewiz (Abingdon, UK) requiring 10 million reads per sample. Sequencing data was processed by Dr Amit Mandal and peaks were called using SEACR (Meers et al., 2019). Peaks were further filtered to select only those occurring ±3kb

from the transcription start site (TSS), collapsing to one peak per gene using the maximum peak signal (maximum read depth at the peak summit position). A published ZFHX4 CUT&RUN (HPA antibody, healthy human neuronal progenitor cells) dataset was accessed (Supplementary Table 13; (Perez Baca et al., 2025)), converting HUGO symbols using gProfiler (gProfiler, 2025), removing duplicates and obtaining 119 total peak genes.

2.6 CRISPR/Cas9

2.6.1 Generating a KO: RNP electroporation

Multi-guide RNA (mgRNA) targeting exon 2 of ZFHX4 was designed and ordered through Synthego (California, USA). This constituted 3 single-guide RNAs (sgRNAs) (5'-UGCAGGGAUGGAGCCUGACA-3', 5'-GCGCAAAAGUGAAGCCUUGC-3' and 5'-CAGCAAAGGAGAUACCCUGC-3') in one 100μM Tris-EDTA-based solution. Recombinant Cas9 protein (Sigma, #C120010-250UG) was reconstituted to 30μM. CRISPR editing was performed on cells using the Neon NxT Transfection System (ThermoFisher, #MPK5000) through ribonucleoprotein (RNP) electroporation-mediated gene KO in different reaction constitutions for cell lines and organoids (Table 2.6).

Table 2.6 | Reagents and electroporation settings for CRISPR/Cas9-mediated gene KO for cell lines and organoids

		Cell Lines			Organoids		
		Stock Conc.	**Final Conc.**	**Total Volumes**	**Stock Conc.**	**Final Conc.**	**Total Volumes**
Reagents	Resus. Buffer R	-	-	7.5μl RNP complex	-	-	10μl RNP complex
	Cas9	30μM	3.6μM		30μM	3.6μM	
	mgRNA	100μM	8μM		100μM	16μM	
Electro-poration settings	Voltage (V)	1320			1350		
	Width (ms)	30			20		
	Pulse (no.)	1			2		
Sample size per reaction		1 x 10^5 cells			4 confluent domes		

RNP complexes were incubated (10 minutes, RTP) before mixing 1:1 with Resuspension Buffer T containing relevant cell suspensions. No-guide controls were performed alongside every reaction, with the volume of mgRNA replaced with Resuspension Buffer R. Standard culture followed electroporation, using antibiotic-free media.

2.6.2 Validating a KO: Sanger sequencing

gDNA was extracted from all KO and matched WT/no-guide control cells using a Quick-DNA Miniprep Kit (Zymo, #D3024). Primers (IDT; Leuven, Belgium) for polymerase chain reaction (PCR) were designed using the Primer3-Benchling tool (San Francisco, USA) to amplify the mgRNA-targeting region (Table 2.7).

Table 2.7 | Primers used for PCR amplification of the CRISPR-targeting region of ZFHX4

	Sequence (5' – 3')	T_m °C	GC %	Amplicon Size
Forward Primer	AGTTTGGTCGGTGCTGCCTGTG	61.9	59.1	899bp
Reverse Primer	GAAGGCGGGCATTAGGGCAGTG	62.2	63.6	

PCR was performed using 100ng DNA, Q5 High-Fidelity Taq technology (NEB, #M0491S) with ThermoFisher deoxynucleotide triphosphates (#18427013) and a T100-Thermocycler employing a primer-specific 72°C annealing temperature calculated using NEB software (NEB, 2025). Amplified products were purified using QIAquick PCR Purification (Qiagen, #28104) and DNA was quantified using the NanoDropOne before submitting for Sanger Sequencing to Genewiz (Abingdon, UK). Trace results were analysed using Synthego's inference of CRISPR edits (ICE) analysis tool (v3.0; (Synthego, 2019)). This was performed for all polyclonal KO cell lines and organoids, with further validation of monoclonal populations obtained through single-cell cloning by limiting dilution in flat-bottom 96-well plates (Corning, #10695951).

2.7 siRNA knockdown

siRNA knockdown (KD) was performed with Dharmacon protocol and reagents (Horizon; Cambridge, UK). A set of 4 ZFHX4-targeting siRNAs (Table 2.8, #LQ-018835-00-0005) and a non-targeting control pool siRNA (#D-001810-10-05) were reconstituted in siRNA buffer (#B-002000-UB-100) to 100μM. Quality control was checked with the NanoDropOne to ensure all concentrations were around 1300ng/ul following reconstitution.

Table 2.8 | siRNA sequences and targeted regions of ZFHX4

siRNA ref.	Sequence (5' – 3')	Targeted Region
siRNA-1	CGACACAACUUGAUAAUGA	Exon 2 (close to sgRNA region)
siRNA-2	CAGCAAAGAUACCCAUUUA	Exon 10 start
siRNA-3	GAACGAGAUUUACUGACUA	Exon 10 mid
siRNA-4	GAGACUUCAUCGAUUAAUA	Exon 10 end

A day prior siRNA KD, cells were plated for western blotting (300k/6-well plates) and proliferation assays (10k/96-well plates) respectively. DharmaFECT-1 (#T-2001-03) or RNAiMAX (Invitrogen, #13778075) transfection reagents were diluted 1:100 in OptiMEM Reduced Serum Medium (RSM) (Gibco, #31985070). siRNAs were diluted in OptiMEM-RSM and incubated (5 minutes, RTP) before 1:1 mixing with transfection reagent and incubating again (20 minutes, RTP). Cells were then incubated (3 days, 37°C) in transfection-siRNA mixes diluted in antibiotic-free cell line media. Untreated controls were cultured with the same volumes of OptiMEM-RSM:antibiotic-free media but without any siRNA or transfection reagent.

2.8 Western blotting

The below methodology depicts the final western blotting (WB) protocol used in this thesis and any deviations from this will be referred to in the relevant chapter.

2.8.1 Protein extraction, quantification and preparation

Cells were lysed with RIPA lysis buffer (ThermoFisher, #89901) and 1X cOmplete Protease Inhibitor Cocktail (Roche, #11697498001) then inverted (30 minutes, 4°C) before centrifuging (4°C, 8000g, 10 minutes). Protein concentration was determined using the Pierce BCA Protein Kit (ThermoFisher, #23225), POLARstar Omega Plate Reader and BCA standard curve-interpolation. Samples were diluted in nuclease-free water and reduced in 1X Laemmli sample buffer (Bio-Rad, #161-0747) supplemented with fresh 2-mercaptoethanol (ThermoFisher, #125472500) at 2.5% as by manufacturer's instructions. Samples were heated (10 minutes, 70°C) and frozen at -80°C. Before loading, samples were thawed and heated again before cooling (10 minutes, RTP).

2.8.2 SDS-PAGE and protein transfer

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was performed differently depending on molecular weight of proteins of interest (Table 2.9). Electrophoresis and transfer were run using a Bio-Rad PowerPacHC (Bio-Rad, #1645052) and respective manufacturer's instructions. High molecular weight protocol was optimised from a published protocol for ZFHX4 (Perez Baca et al., 2025).

Table 2.9 | SDS-PAGE and protein transfer parameters

	Parameter	Standard Molecular Weight (12-225kDa)	High Molecular Weight (>350kDa)
SDS-PAGE	NuPAGE Mini Gel	4-12% Bis-Tris 1.5mm (Invitrogen, #NP0335/6)	3-8% Tris-Acetate 1.5mm (Invitrogen, #EA0378/5)
	Running Buffer	1X MES SDS in dH_2O (Invitrogen, #NP0002)	1X Tris-Acetate SDS in dH_2O (Invitrogen, #LA0041)
	Running Settings	150V 1.5 hours, RTP	100V 1 hour then 150V 1 hour, RTP
	Protein Ladder	Amersham Rainbow Full Range (Cytiva, #RPN800E) or PageRuler Plus (Thermo Fisher, #26619)	HiMark Pre-stained (Invitrogen, #LC5699)
	Running System	XCell SureLock Mini-Cells (Invitrogen, #EI0001)	
Transfer	Transfer System	XCell II Blot Module (Invitrogen, #EI9051)	Mini Trans-Blot (Bio-Rad, #1703930/5)
	Transfer Buffer	1X NuPAGE, 10% Methanol in dH_2O (Invitrogen, #NP00061)	
	Transfer Settings	30V 1.5 hours, RTP	30V ONT – 24 hours, 4°C and on ice with ice pack inside
	Membrane Type	0.2μm Nitrocellulose (NCM; Cytiva, #10600006)	

ONT=overnight, RTP=room temperature

2.8.3 Blotting

Ponceau (Thermo, #A40000279) was used to check protein transfer. Membranes were then washed with Tris-Buffered Saline with Tween 20 (TBST; Sigma, #T9039) 3 times for 5 minutes. Membranes were blocked (1 hour, RTP) with blocking buffer (BB; 5% Milk in TBST) and primary antibodies (Table 2.10) were prepared in BB at 1:2000 and incubated (overnight; ONT, 4°C). Antibodies probing housekeeping proteins were incubated for 1 hour at RTP at 1:10000. Membranes were washed and secondary antibodies conjugated to Horseradish Peroxidase (HRP) were prepared in BB (Table 2.10) at 1:10000 and incubated (1 hour, RTP).

Table 2.10 | Primary and secondary antibodies used for western blotting

Ab ref.	Target* (size kDa)	Host Species	Clonality	Supplier and Cat. No.
Primary Antibodies (non-housekeeper)				
PA5	ZFHX4 (393)	Rabbit	Poly	Thermo Fisher, #PA5-99757
abc	ZFHX4 (393)	Rabbit	Poly	Abcam, #ab254654
HPA	ZFHX4 (393)	Rabbit	Poly	Sigma, #HPA023837
M11	ZFHX4 (393)	Mouse	Mono (2B12)	Abnova, #H00079776-M11
M02	ZFHX4 (393)	Mouse	Mono (5E6)	Abnova, #H00079776-M02
Primary Antibodies (housekeeper)				
GAPDH (36)		Rabbit	Poly	Abcam, #9485
Lamin B1 (66)		Mouse	Mono (119D5-F1)	Santa Cruz, #sc-56143
β-actin (42)		Rabbit	Mono (13E5)	CST, #4970S
HSP60 (60)		Rabbit	Mono (D6F1)	CST, #12165S
Vinculin (117)		Mouse	Mono (VIN-11-5)	SLS, #V4505
Secondary Antibodies**				
Anti-rabbit IgG(H+L)		Goat	Poly	Proteintech, #SA00001-2
Anti-mouse IgG(H+L)		Goat	Poly	Proteintech, #SA00001-1

**all target species are anti-human, **all secondary antibodies are HRP conjugated*

Electrochemiluminescence (ECL) was performed using Amersham ECL Prime (Cytiva, #RPN2232) for housekeeper proteins and SuperSignal West Atto/Femto (ThermoFisher, #A38554/#34094) for ZFHX4. Membranes were imaged with the ChemiDocXRS+ System (Bio-Rad) using the “Chemi High Sensitivity” automated exposure program. Where applicable, bands were quantified using Fiji (Schindelin et al., 2012) as by Supplementary Methods 2.3. For some ZFHX4 experiments, membranes were stripped to validate bands using another primary antibody. Membranes were incubated (5 minutes, 37°C) with Restore PLUS Stripping Buffer (ThermoFisher, #46430), washed and ECL confirmed secondary antibody removal.

2.8.4 Nuclear fractionation

Buffer A (2X) was made with final concentrations of 20mM HEPES, 20mM KCl (Sigma, #P3911), 680mM Sucrose (Sigma, #S0389) and 20% Glycerol (Sigma, #G5516) in dH_20. NaOH (Sigma, #72068) was used to obtain an accurate pH of 7.45. Cells were washed in ice-cold DPBS taking half for a whole cell extract control which followed standard RIPA

lysis, and half for fractionation. Here, cells were centrifuged (500g, 3 minutes, 4°C) and resuspended in 2x pellet-volumes of 1X Buffer A (2X buffer A and 25X PIC in dH_20). Cells were permeabilised (0.1% Triton X-100) with gentle tapping and incubated for 5 minutes on ice. Samples were centrifuged (300g, 3 minutes, 4°C) and the supernatant was used as the cytoplasmic fraction. Nuclear pellet fractions were washed three times with 1X Buffer A, centrifuged (500g, 3 minutes, 4°C) and then resuspended with RIPA. Both nuclear and cytoplasmic fractions were then centrifuged as normal (8000g, 10 minutes, 4°C).

2.9 Flow Cytometry

At least 20k T-cells were resuspended in MACS (Section 2.1.4) and plated in U-bottomed 96-well plates. Plates were centrifuged (Eppendorf 5804R; 500g, 5 minutes) and aspirated by motion before washing with MACS. Except for fluorescence minus one (FMO), unstained and live/dead controls, a master mix of flow antibodies and dyes (Table 2.11) were added to each well and incubated in the dark (30 minutes, 4°C). FMOs were prepared in respective wells, omitting the antibody/dye of interest in the panel. Live/dead controls were prepared by heat-killing 50% sample (10 minutes, 70°C) before staining. Where appropriate, compensation controls constituted UltraComp-eBeads (Invitrogen, #01-2222-42) with antibodies diluted 1:100 as by manufacturer's instructions. For markers utilising dyes, compensation controls were created by mixing 0% and 100% stained populations 1:1.

Table 2.11 | Flow cytometry antibodies and dyes used in various flow panels

Laser (EM Filter)		Marker	Colour/ Format	Host Species, Isotype and Clone	Dil.	Supplier and Cat. No.
PANEL ONE						
405	(440/50 417LP)	PD-1	BV421	Mouse IgG1κ EH12.2H7	1:100	BioLegend, #329920
	(710/50)	CD25	BV711	Mouse IgG1κ BC96	1:100	BioLegend, #302636
488	(530/30 503LP)	CFSE	CFSE	-	1:100	Invitrogen, #C34554
637	(670/14 654LP)	CD69	AF647	Mouse IgG1κ FN50	1:100	BioLegend, #310918
	(780/60)	Viability	Live/Dead NearIR	-	1:100	Invitrogen, #L34976

PANEL TWO						
561	(585/16 577LP)	TCR/CD3	PE	Mouse IgG1κ IP26	1:100	BDBiosciences, #564728
637	(780/60)	Viability	Live/Dead NearIR	-	1:100	Invitrogen, #L34976
PANEL THREE						
405	(440/50 417LP)	PD-1	BV421	Mouse IgG1κ EH12.2H7	1:100	BioLegend, #329920
	512/25	CD8a	BV510	Mouse IgG1κ SK1	1:200	BDBiosciences, #563919
	(710/50)	CD25	BV711	Mouse IgG1κ BC96	1:100	BioLegend, #302636
488	(530/30 503LP)	CD4	FITC	Mouse IgG1κ RPA-T4	1:50	BDBiosciences, #555346
561	(585/16 577LP)	IFNγ	PE	Mouse IgG1κ B27	1:200	BDBiosciences, #554701
637	(670/14 654LP)	CD69	AF647	Mouse IgG1κ FN50	1:100	BioLegend, #310918
	(780/60)	Viability	Live/Dead NearIR	-	1:500	Invitrogen, #L34976

Samples were resuspended in fixation buffer (BioLegend, #420801) and incubated in the dark (20 minutes, RTP) before washing and transferring into flow tubes (Corning, #352052) or 96-well plates. Flow cytometry was performed on the AttuneNxT Flow Cytometer (Invitrogen, #A24858) and data was analysed using FlowJo-v10.10.0 software (BDBiosciences, Oregon, USA). For panel three, intracellular staining for IFNγ was conducted using the BD Cytofix/Cytoperm Kit (BDBiosciences, #554714). Cells were incubated with eBioscience protein transport inhibitor cocktail (Invitrogen, #00-4980-93) 6 hours prior to staining which followed protocol as above until fixation. Here, BD Fix/Perm solution was added (20 minutes, 4°C) before washing twice with BD Perm/Wash solution. IFNγ antibody (Table 2.11) was added into BD Perm/Wash and incubated in the dark (ONT, 4°C) before washing twice in BD Perm/Wash and resuspending in MACS. For antibody titration experiments, separation indices (SI) were calculated using the following formula on median fluorescence intensity (MFI) values (UWCCC Flow Cytometry Laboratory, 2016):

$$SI = \frac{MFIPositive - MFINegative}{(84\%Negative - MFINegative)/0.995}$$

2.10 Bioinformatics

2.10.1 Programming language and software

All bioinformatic analyses were conducted in RStudio (v2023.09.1+494; (Posit Team, 2023)) using R-programming (v4.2.2;(R Core Team, 2022)). Relevant scripts were uploaded to my GitHub repository (Flower, 2025) and R packages and versions are referenced in Table 2.12. For some packages, an online RStudioIDE was used via positCloud (Posit Team, 2025) due to version incompatibility with Rv4.2.2.

Table 2.12 | R packages used throughout the project (mostly installed with Bioconductor v3.16)

Package Name	Version	Package Name	Version
ASCAT	3.1.1	GSVA*	2.2.0
Biobase	2.58.0	gtsummary	1.7.2
BiocManager	1.30.26	limma	3.54.2
BiocOncoTK	1.18.0	maftools	2.14.0
Biostrings	2.66.0	msigdbr	7.5.1
BSgenome.Hsapiens.UCSC.hg38	1.4.5	NMF	0.26
CMScaller	2.0.1	org.Hs.eg.db	3.16.0
ComplexHeatmap	2.21.2	pROC	1.18.5
ConsensusClusterPlus	1.62.0	purrr	1.0.2
ConsensusTME*	0.0.1.9000	RColorBrewer	1.1-3
corrplot	0.92	readxl	1.4.3
data.table	1.17.8	rstatix	0.7.2
data.table	1.17.8	scatterplot3d	0.3-44
dplyr	1.1.4	sigminer	2.2.2
edgeR	3.40.2	stringr	1.5.1
EnhancedVolcano	1.16.0	survival	3.5-7
EnsDb.Hsapiens.v86	2.99.0	survminer	0.4.9
fgsea	1.24.0	Tibble	3.2.1
fmsb	0.7.6	tidyestimate*	1.1.1.9000
ggfortify	0.4.16	tidyr	1.3.1
ggplot2	3.5.2	tidyverse	2.0.0
ggplot2	3.5.2	VennDiagram	1.7.3
ggpubr	0.6.0	viridis	0.6.4
ggrepel	0.9.5	writexl	1.4.2
ggsurvfit	0.3.1		

**package employed with R v4.5.1 and Bioconductor v3.21 in positCloud*

2.10.2 Dataset acquisition and processing

2.10.2.1 TCGA-COAD

The Cancer Genome Atlas (TCGA) Colon Adenocarcinoma (COAD) dataset was used as the project for study (Hoadley et al., 2018). For RNAseq data, raw STAR count files were downloaded from Genomic Data Commons (GDC; (NIH National Cancer Institute, 2025)) and bash-compiled into a single file by Dr Amit Mandal. Filtering retained 1 sample per patient, selecting primary colon adenocarcinoma as by the ICD-10 and downloaded cBioportal annotations (World Health Organisation, 2019, Cerami et al., 2012). Counts per million (CPM) thresholding of ≥10 counts at the median library size (35.6M) in 10% all samples (22885 genes) removed lowly expressed genes. Limma-voom allowed count normalisation and subsequent batch correction using TSS_ShipDate annotated from MBatchOmicBrowser (Akbani et al., 2025). Whole Exome Sequencing (WES) data were downloaded from GDC combining muse, mutect2, varscan2 and pindel variant callers. Other sample annotations were obtained from other studies (Thorsson et al., 2018, Charoentong et al., 2017, Saltz et al., 2018) or other TCGA sources (Cancer Genome Atlas, 2012, Cerami et al., 2012). Consensus molecular subtype (CMS) was characterised using the original CMS caller (Eide et al., 2017) as by two other studies (Greene et al., 2022, Khaliq et al., 2022) with 93% accuracy to original CMS labels (Guinney et al., 2015). Lymph Node (LN) status was annotated from cBioportal as by two previous studies (Lal et al., 2022, Bundred et al., 2025). Tumour immune dysfunction and exclusion (TIDE) scores were produced following gene expression matrix upload to its respective online portal (Jiang et al., 2018).

2.10.2.2 Other CRC Big-Data

Supporting TCGA-COAD for further validation throughout this thesis, three other datasets accompanying matched transcriptomic and genomic data were downloaded and processed similarly to retain only treatment-naïve primary colon adenocarcinomas (Table 2.13).

Table 2.13 | Summary of supporting colon cancer datasets used throughout the thesis

Study and Reference	Data Type	Download and Format	Data Processing	N* (att.)	Final N**
TCGA-COAD[1]	RNAseq	Described in Section 2.10.2.1		376 (22885)	342
	WES			343 (88661)	
	Clinical variables			376 (210)	
Sidra-LUMC AC-ICAM (SIDRA)[2]	RNAseq	S3.1 Supplementary Source Data (GRCh38, quantile normalised log2 transformed, HUGO)	Conversion to csv., manual renaming of date genes and removal of duplicates	348 (19456)	224
	WES	Figshare online download (Roelands et al., 2023b) (GRCh37, Mutect and Strelka)	Conversion to MAF file with maftools	281 (140127)	
	Clinical variables	S2.1 Supplementary Source Data	None	349 (36)	
Nunes[3]	RNAseq	EMBL-EBI database, #E-MTAB-12862 (GRCh38, raw counts, ENSEMBLID)	Same as TCGA processing (read count >10 at median library size 29.6M in at least 10% samples) but no batch correction + removal of pretreated samples	597 (16472)	597
	WES	ZFHX4 mutations only (with thanks to Dr James Bundred)	None	328 (1)	
	Clinical variables	Supplementary Table S1	None	1063 (63)	
CPTAC-2 (CPTAC)[4]	RNAseq	cBioportal download – 2025 GDC version (GRCh38, raw counts, ENTREZID)	Same as TCGA processing (read count >10 at median library size 24.2M in at least 10% samples) but no batch correction	82 (19895)	84
	WES	cBioportal download	None	106 (53891)	
	Clinical variables	cBioportal download	None	105 (53)	

**representing unique samples:patient 1:1, **final number of samples following RNAseq-WES-clinical data matching and subsetting treatment-naïve primary colon adenocarcinoma only. Processed files were used when raw files were not available/had to be officially accessed. [1](Cancer Genome Atlas, 2012), [2](Roelands et al., 2023a), [3](Nunes et al., 2024), [4](Vasaikar et al., 2019) [att.=attributes, WES=whole exome sequencing, RNAseq=RNA sequencing]*

One pan-cancer immunotherapy WES dataset was downloaded directly from cBioportal, representing 294 samples from 6 cancer types with immunotherapy response (Miao et al., 2018).

2.10.2.3 Organoid and cell line RNAseq

For cell line model selection, mRNA and protein expression of all cell lines were accessed and downloaded from DepMap (25Q2 release; (DepMap Broad Institute, 2025)). Organoid model selection was based upon a matched WES and RNAseq dataset (8 lines) obtained from Dr Scott Collins and pre-processed and normalised by Dr Mandal. RNAseq ZFHX4$^{-/-}$ count matrices obtained from Dr Mandal were processed similar to Section 2.10.2 (Table 2.14). For some analyses, transcript-level estimation matrices were provided by Dr Mandal (StringTie processed). All gene set enrichment analyses (GSEAs) and pathway enrichments were performed using the fgsea and msigdbr package in RStudio and employed 10930 gene sets (15-500 size limits) comprising Hallmark, C2 (BIOCARTA, KEGG) and Gene Ontology (Biological Process, Cellular Component, Molecular Function).

Table 2.14 | RNAseq processing of cell line and organoid *ZFHX4*$^{-/-}$ experiments

Experiment	Limma-CPM thresholds allowing removal of lowly expressed genes	Specialised design matrix?
Cell line KO	≥5 counts at the median library size (51.8M) in at least 3 samples (25199 genes)	No
Organoid KO	≥5 counts at the median library size (39.1M) in at least 3 samples (21826 genes)	Electroporation-related effects were regressed out/included in limma design

2.10.3 Multiomic clustering on TCGA-COAD data

A large matrix representing multiple determinants of immunogenicity (mDOI; Figure 1.3) was built following extensive literature searches (Table 2.15). 'Single' gene expression parameters were further filtered to reduce multicollinearity. Here, genes spanning multiple groups were only represented once (339 total genes reduced to 281 unique genes), for example IFNG was included as by the interferon group but also the stimulatory immunomodulatory gene (IMG) group. To note, some overlap may exist from genes used in multiple gene expression scores. Samples columns containing NA values were removed, reducing sample number from 376 to 275. This allowed z-score scaling for parameter normalisation, and clustering was performed using the ConsensusClusterPlus (CCP) package (Wilkerson and Hayes, 2010). A k-means clustering algorithm of 5000 iterations using Euclidean distance was performed, testing k values from 1-15. Optimal cluster number was chosen based upon the elbow method, selecting the k for which a large increase was followed by smaller increases in the area under the corresponding cumulative distribution function (CDF) curve as by its original methodology (Monti, 2003, Wilkerson and Hayes, 2010). The CCP package also produced internal validation indices, namely mean consensus scores which infer how often items cluster together over multiple iterations.

For other bioinformatic methodology, see Supplementary Methods 2.4.

2.11 Data visualisation and statistical analyses

Data and plots produced in RStudio were exported in pdf format to allow vector-based graphic editing in Adobe Illustrator (version 28.7.10, 2024), of which all figures in this thesis were produced in. All statistical analyses were performed using RStudio. Datasets were always checked for normality by Shapiro-Wilkes testing, and then respective non-parametric or parametric testing would be performed using base R packages.

Table 2.15 | **Parameters employed in the ICS clustering pipeline, each representing a determinant of tumour immunogenicity (DOI)**

DOI	Clustering Parameter	Derivation/Rationale	N (r)	N (c)
 **1. Ability to Express Antigen**	MSI Sensor Score	Annotated as described in Supplementary Methods 2.4	1	334
	Mismatch Repair Deficiency (MMRd) Score		1	343
	Polymerase Proofreading Deficiency (PPd) Score		1	343
	Tumour Mutation Burden (TMB)		1	343
	Missense_SNV Count	MAF counts (muse, mutect2, varscan2 and pindel called) of MODERATE and HIGH mutations downloaded from GDC that have a gVAF < 0.001	1	343
	Nonsense_SNV Count		1	343
	Nonstop_SNV Count		1	343
	Splice_Site_SNV Count		1	343
	TSS_SNV Count		1	343
	FS_Del Count		1	343
	InFrame_Del Count		1	343
	FS_Ins Count		1	343
	InFrame_Ins Count		1	343
	AIM SNV Count	Counts of SNV/indel mutations predicted to be antigenic (can be presented to cell surface) and immunogenic (can be recognised on surface) (termed AIMs). Annotated from TSNAdbv2.0[1] with rationale by Turajlic et al.[2]. Counts represent individual mutations per sample ID irrespective of HLA allele or peptide. Filtering includes deepbind>0.5 and deepimm>0.5	1	320
	AIM Indel Count		1	320
	AIM SNV_Indel Count		1	320
	AIM SNV Rate	AIM rate - above counts divided by TMB as a ratio, with rationale by three studies[3-5] and similar to the neoantigen index[6,7]	1	320
	AIM Indel Rate		1	320
	AIM SNV_Indel Rate		1	320
	Predicted Tumour Neoantigen Burden (pTNB)	Number of potential peptides strongly binded as neoantigen to various HLA alleles derived from various SNV and indel AIMs per sample (i.e accounts for HLA alleles and peptides now). Rationale for this suggested in a review[8] and three studies[2,9,10]	1	320
	Neoantigen Load Score (NLS)	Neoantigen scores represent binding affinity scores (i.e. antigenicity) and immunogenicity scores per antigen (generated by deepHLApan - data available in TSNAdbv2.0) multiplied together. Summed neoantigen scores per sample gives NLS. Method as by Luo et al.[11] but just SNVs and indels (no fusions).	1	320

	Cancer Germline Antigen Expression Scores			Summed expression values of available CGA genes taken from the CTA database[12] [79]	1	376
				Summed expression values of available CGA genes taken from Wang et al.[13] [437]	1	376
	Intratumoural Heterogeneity			MATH scores[14] calculated using the maftools R package on the MAF file	1	343
	Genomic Instability Index			Values annotated from ASCAT[15] via their GitHub site	1	345
	Ploidy				1	345
	Fraction Genome Altered			Values annotated from Thorsson et al.[16]	1	356
2. Ability to Process and Present Antigen	Antigen processing/presentation machinery (APM) gene expression			Expression values of genes curated from various studies[5,16-19] and the HNGC HLA gene set[20] [44]	Within 281	376
	APM Score			Calculated as by Thompson et al.[19] and represents the sum of z-scores of 8 APM genes	1	376
	APM Signature			ES from GSVA of gene set from Gong and Karchin[18] [23]	1	376
	Tumour Immunogenicity Score (TIGS)			Similar to Wang et al.[21], TIGS represents 0-1 scaled APM scores multiplied by log(TMB+1)	1	343
3. Ability to Attract Immune-Destructive Environment	Immune Cell Abundance	Adaptive Infiltrate	9 types	ES from immune deconvolution performed by ConsensusTME[22]	17	376
		Innate Infiltrate	7 types	ES from immune deconvolution performed by ConsensusTME[22]		
			MDSC	ES from GSVA of gene set from Charoentong et al.[23] [20]		
		Infiltration Scores	Leukocyte Score	Estimated Leukocyte content (calculated using DNA methylation patterns and a mixture model) annotated from Thorsson et al.[16]	1	362
			Immune Score	Values calculated using tidyestimate R package which uses ESTIMATE[24]	1	374
	Immune Cell Function	Gene Signatures/ Scores	Cytotoxicity Score	Mean average of GZMA and PRF1 expression, calculated as by Rooney et al.[25]	1	376
			T-Cell Inflamed Signature	ES from GSVA of gene set from Ayers et al.[26] [18]	1	376

 **4. Ability to Exclude Immune-Suppressive Environment**			6-gene IFNg Signature	ES from GSVA of gene set from Ayers et al.[26] [6]	1	376
			Effective T-Cell Sig.	ES from GSVA of gene set from McDermott et al.[27] [5]	1	376
			Immunologic Constant of Rejection	ES from GSVA of gene set from Bertucci et al.[28] [19], rationale by Bedognetti et al.[29]	1	376
			Immune hot score	ES from GSVA of gene set from Wang et al.[30] [12]	1	376
			Immune cold score	ES from GSVA of gene set from Wang et al.[30] [3]	1	376
			12-Chemokine Signature	ES from GSVA of gene set from Coppola et al.[31] [12]	1	376
			Complement System Score	Summed expression values of genes curated from Olcina et al.[32] and HNGC complement gene set[33], rationale by Krieg et al.[34] and Reis et al.[35] [44]	1	376
 **5. Ability to Succumb to Immune-Mediated Killing**		Stimulatory IMG expression		Expression values of genes curated from three studies[5,16,17] [52]	Within 281	376
		Inhibitory IMG expression		Expression values of genes curated from three studies[5,16,17] [28]		
		Cytokine signalling expression	Chemokines	Expression values of genes curated from four studies[17,31,36,37] [58]		
			Interleukins	Expression values of genes curated from two studies[17,38] [77]		
			Interferons	Expression values of genes curated from Parker et al.[39], rationale by von Locquenghien et al.[40] (all Type I/II/III) [7]		
			TGFs	Expression values of genes curated from Derynck et al.[41] [5]		
			TNFs	Expression values of genes curated from two studies[42,43], rational by Ward-Kavanagh et al.[44] [42]		
			CSFs	Expression values of genes curated from Metcalf[45] [8]		

		Immune-mediated cytotoxic activity gene expression		Expression values of genes curated from four studies[46-49] [18]		
	Other TME components	Stroma	Fibroblasts & Endothelial Infiltrate	ES from immune deconvolution performed by ConsensusTME[22]	2	376
			Stromal Score	Values calculated using tidyestimate R package which uses ESTIMATE[24]	1	376
			Tumour Purity		1	376
		ECM		ES from GSVA of gene set from Naba et al.[50] 'primary colon tumour', rationale by Gordon-Weeks et al.[51] [35]	1	376
				ES from GSVA of gene set from Naba et al.[50] 'core matrisome', rationale by Gordon-Weeks et al.[51] [135]	1	376
#parameter rows: 345 #sample columns: 376 **#non-NA sample columns: 275**						

[1](Wu et al., 2023), [2](Turajlic et al., 2017), [3](Roelands et al., 2023a), [4](Rooney et al., 2015), [5](Angelova et al., 2015), [6](Bortolomeazzi et al., 2021), [7](Sugawara et al., 2022), [8](Wang et al., 2021b), [9](Feng et al., 2022), [10](Jiang et al., 2020), [11](Luo et al., 2023), [12](Almeida et al., 2009), [13](Wang et al., 2016), [14](Mroz and Rocco, 2013), [15](Steele et al., 2022), [16](Thorsson et al., 2018), [17](Mao et al., 2022), [18](Gong and Karchin, 2023), [19](Thompson et al., 2020), [20](HGNC, 2025b), [21](Wang et al., 2019), [22](Jimenez-Sanchez et al., 2019), [23](Charoentong et al., 2017), [24](Yoshihara et al., 2013), [25](Rooney et al., 2015), [26](Ayers et al., 2017), [27](McDermott et al., 2018), [28](Bertucci et al., 2018), [29](Bedognetti et al., 2016), [30](Wang et al., 2021a), [31](Coppola et al., 2011), [32](Olcina et al., 2018), [33](HGNC, 2025a), [34](Krieg et al., 2022), [35](Reis et al., 2018), [36](Kohli et al., 2022), [37](Ozga et al., 2021), [38](Briukhovetska et al., 2021), [39](Parker et al., 2016), [40](von Locquenghien et al., 2021), [41](Derynck et al., 2021), [42](Dostert et al., 2019), [43](So and Ishii, 2019), [44](Ward-Kavanagh et al., 2016), [45](Metcalf, 2010), [46](Chang et al., 2022), [47](Balint et al., 2020), [48](Liu et al., 2020), [49](Prager and Watzl, 2019), [50](Naba et al., 2014), [51](Gordon-Weeks and Yuzhalin, 2020)

Numbers of square brackets represent gene numbers. [MDSC=myeloid-derived suppressor cell, TGF=transforming growth factor, TNF=tumour necrosis factor, CSF=colony stimulating factor, r=rows, c=columns, ES=enrichment score, GSVA=gene set variation analysis, IMG=immunomodulatory gene]

Chapter Three: Interrogating tumour immunogenicity in publicly available datasets

3.1 Introduction

The growing availability of large-scale transcriptomic, genomic and other multi-omic cancer datasets provides unprecedented opportunities to identify biomarkers and therapeutic targets, particularly in relation to tumourigenicity, immunogenicity and therapeutic response. By employing such data, the association of gene expression and mutation with specific biology such as immunogenicity can be unpicked bioinformatically and then later tested empirically with *in vitro* and *in vivo* models and methods. For colon cancer, four large matched RNA sequencing (RNAseq) and whole exome sequencing (WES) datasets exist, with three comprising >300 patients (Nunes et al., 2024, Roelands et al., 2023, Cancer Genome Atlas, 2012, Vasaikar et al., 2019). The most characterised and well-studied is The Cancer Genome Atlas (TCGA), with TCGA-CRC encompassing nearly 600 patients. Over the past decade TCGA has been extensively leveraged in both oncology and immuno-oncology, resulting in a wealth of patient-level annotations and metadata. TCGA-Colon Adenocarcinoma (COAD) was the only dataset available at the start of the DPhil and was thus used for the foundation of this chapter work given its broad acceptance across the field, rich data measures and integration of matched omics information.

Increasing evidence suggests mutation quality (i.e. the genes in which they occur) is just as critical as quantity (i.e. TMB) in shaping immunogenicity (Wang et al., 2023, Ishino et al., 2023, Porta-Pardo and Godzik, 2016). Regarding tumourigenicity, mutations in well-established colon cancer driver genes such as APC, BRAF and TP53 are known to impact tumourigenic pathways in specific and well-characterised ways. It's therefore plausible that mutations may also modulate multiple determinants of immunogenicity (mDOI; Figure 1.3) through disruption of the gene's original biological functions, independent of their potential to generate neoantigens. This may help explain the heterogenous immunotherapeutic responses observed in mCRC clinical trials irrespective of TMB status, where 30-50%

MMRd/MSI patients lack response (Andre et al., 2022, Andre et al., 2020, Diaz et al., 2022) and some MSS patients demonstrate sensitivity (Fakih et al., 2023, Chalabi et al., 2020, Saunders et al., 2022, Morris et al., 2022).

Identifying mutational drivers of immunogenicity could facilitate the development of therapies that restore anti-tumour immunity and enhance immunotherapy efficacy, parallelling personalised cancer therapies that emerged from the discovery of oncogenic driver mutations (Martinez-Jimenez et al., 2020). Such efforts will need to be employed in a TMBL setting to minimise the confounding effects of mutational load and neoantigen burden, which although insufficient to fully explain immunogenicity or immunotherapy response remain strongly associated with both as explained by the neoantigen hypothesis (Samstein et al., 2019, Monach et al., 1995). Beyond uncovering new mutational drivers of immunogenicity, this approach could address a major unmet need in TMBL and MSS tumours which constitute most CRCs. Although these tumours are typically classified as a uniformly non-immunogenic group, increasing evidence suggests they comprise distinct subtypes of immunogenicity potentially defined by differences in mutation quality (Acha-Sagredo et al., 2025, Wang et al., 2022, Taraborrelli et al., 2023).

3.2 Hypothesis

I hypothesise that independent of TMB and overall mutational load, specific gene mutations are associated with tumour immunogenicity in TCGA-COAD, detectable both through minimalist descriptors (such as cytotoxicity score and CD8A expression) and through broader, integrative measures that more comprehensively capture mDOI.

Aims

1. Compare mutational enrichment in TMBL immune-segregated groups using a few key parameters of immunogenicity
2. Compare mutational enrichment in TMBL immune-segregated groups defined through a holistic clustering approach using mDOI

3.3 Experimental Design

3.3.1 Overview

TCGA-COAD RNAseq and WES datasets were downloaded in a raw format to allow standardised processing. In addition, clinical variables and other patient annotations from the literature were downloaded and used as the basis for all work in this chapter. Only samples with 'COAD or not otherwise specified (NOS)' histology were selected. All analyses were conducted with R-programming on RStudio. Two complementary strategies were employed iteratively to investigate mutational drivers of tumour immunogenicity, defining immunogenicity using a minimalistic-targeted approach and a more robust systems-level holistic approach.

3.3.2 Defining mutational load (ML) phenotype

As outlined in the introduction, MMRd drives MSI which in turn results in TMBH. However, not all TMBH tumours are a result of MMRd/MSI. POLE and POLD1 mutant tumours are usually MMRp and MSS but display an ultramutator phenotype and are a distinct colon cancer subgroup (Campbell et al., 2017). Therefore, characterising ML by incorporating MMR-status, polymerase-proofreading (PP)-status, MS-status and TMB-status (and consequently hypermutation status; HM) would be a robust approach to select low-ML tumours for identifying mutational drivers of immunogenicity. TCGA-COAD patients were categorised into ML phenotypes using specified definitions and cut-offs which are commonly inconsistent amongst literature. Mutation comparisons were performed with respect to all-ML tumours and low-ML tumours.

3.3.3 Comparing mutations between Imm^{high} and Imm^{low} patients categorised by five immunogenicity parameters

The first strategy comprises a focused approach, segregating TCGA-COAD patients into poorly immunogenic (Imm^{low}) and highly immunogenic (Imm^{high}) groups based upon five well-characterised parameters comparable with variables associated with immunotherapy response and thus immunogenicity (Table 3.1) (Gonzalez-Montero et al., 2024). TMB was not used as a parameter because cohorts were analysed based upon ML phenotype to reduce the confounding for mutational driver investigation. Reducing immunogenicity to its simplest level, an immunogenic tumour could be described by the presence of a pre-existing anti-tumour immune response and the ultimate effector of anti-tumour immunity, cytotoxic $CD8^+$ T-cells (Rooney et al., 2015).

Table 3.1 | Key parameters of immunogenicity that can be mined from TCGA-COAD

Parameter	Description	Justification and Reference
CD8	Gene expression (log2)	Indicates presence of intratumoural CD8+ T-cells (Galon et al., 2012)
PD-L1	Gene expression (log2)	Associated with the presence of intratumoural CD8+ T-cells (Tumeh et al., 2014)
Cytotoxicity Score	Average GZMA and PRF1 expression (non-log2)	Indicates T-cell activation and cytotoxicity within tumours, predominantly through $CD8^+$ T-cells (Rooney et al., 2015)
Tumour Infiltrating Lymphocyte (TIL) Score	% TIL estimated by deep learning on H&E slides	Indicates presence of intratumoural T-cells (Saltz et al., 2018)
Immune Score	Normalised enrichment score following ssGSEA of a TCGA-COAD specific gene set to represent all immune infiltrate	Indicates presence of intratumoural immune cells (Jimenez-Sanchez et al., 2019)

Designation of Imm^{high} and Imm^{low} was based upon a match of at least 3 parameters in the top and bottom 15^{th} percentile of scores respectively. Mutations were compared between Imm^{high} and Imm^{low} groups and were also investigated for their associations with survival in an immunotherapy-treated cohort (Miao et al., 2018).

3.3.4 Comparing mutations between mDOI-defined subgroups characterised as Immunogenicity classification of Colon cancer Subtypes (ICS)

The second strategy aimed to provide a more robust and unbiased characterisation of complex, high-dimensional immunogenicity profiles within the TCGA-COAD cohort. Patients were clustered using k-means-based clustering on a substantially 69-fold broader set of features describing multiple determinants of immunogenicity (mDOI) to better capture the biological heterogeneity of tumour-immune interactions. This approach was intended to uncover biologically distinct immunophenotypic subgroups that may more accurately reflect the diversity of real tumour-immune states than the previously defined Imm^{high} and Imm^{low} categories. While the Imm^{high}/Imm^{low} stratification offers interpretability and clarity using well-established immune-related parameters, it likely captures only a limited slice of the immunogenicity landscape. For example, Imm^{high} could primarily reflect an inflamed tumour microenvironment (TME) without differentiating between functionally competent, suppressed or exhausted immune states, each of which carries different clinical implications (Chen and Mellman, 2017, Hegde and Chen, 2020, Galon et al., 2013). Moreover, composite scores such as the immune score aggregate signals from both pro- and anti-tumour immune populations, potentially obscuring meaningful distinctions. In contrast, clustering based on a wide spectrum of immunogenicity-related features accommodates the complexity of tumour-immune dynamics and allows the identification of distinct immune states beyond a simplified binary classification. This approach not only enhances immune-biological insight but may also yield clinically relevant subtypes, thereby offering greater translational value in addition to identifying mutational drivers of immunogenicity. For example, other studies have classified TCGA-COAD into immune subtypes using a similar approach and were able to test certain transcriptomic biomarkers *in vitro* following their identification *in silico* with respect to subtypes (Wang et al., 2022).

The rationale of parameter choice was manually curated from the literature and based on the mDOI spanning a tumour's ability to express antigen, process and present it, attract or

supress specific microenvironments and succumb to immune-mediated cytotoxicity. Parameter values were calculated from transcriptomic and genomic data or annotated from previous publications which analysed TCGA-COAD data. Clustering based on these parameters revealed distinct groups which were characterised and compared with particular attention to TMBL tumours exhibiting varying degrees of immunogenicity.

3.4 Results

3.4.1 Extrapolating immune-related phenotype in transcriptomic data reveals heterogeneity within COAD tumours irrespective of mutational load phenotype

Before investigating mutational drivers and immune-related phenotypes, ML phenotype was defined in the TCGA-COAD cohort (n=341) by MMR-status, PP-status, MS-status, TMB-status and HM-status (Figure 3.1). HM-status represents the cardinal descriptor and 'end-result' of ML since it's the accumulation of mutations that underpins the neoantigen hypothesis. It was defined to distinguish subsets of TMBH patients such as those with PPd who exhibit an ultramutator phenotype distinct from other TMBH cases. Defining and classifying these statuses and their relationship in forming ML is particularly important to understand so that we can obtain a 'true' non-hypermutated (nHM) population for mutational investigation.

An MMRd score was generated for each sample as a composite of mutational processes associated with MMRd (as defined by COSMIC; Supplementary Methods 2.4) given that MMRd-IHC data were not available for TCGA. Thresholding was performed using receiver operator curve (ROC) analysis and Youden's Index using MS-status as the response and MMRd score as the predictor, as MMRd consistently leads to MSI. Area under the curve (AUC) demonstrated great sensitivity (0.95) and specificity (0.99) at assigning MS-status based upon an MMRd score threshold of 0.4774415 (Figure 3.1A). PPd score was devised

similarly but with mutational processes associated with PPd as by COSMIC. The PPd phenotype is particularly rare (Campbell et al., 2017) so thresholding was performed manually using a density plot to segregate TMBL and TBMH (as MSI is not associated with PPd), identifying three potential samples with this phenotype (Figure 3.1B). TMBL and TMBH cut-offs were <10muts/Mb and ≥10Muts/Mb respectively. To note, both scores are predictors of likelihood rather than a definitive designation. ‘Likely MMRd’ and ‘likely PPd’ are consequently referred to as MMRd and PPd respectively.

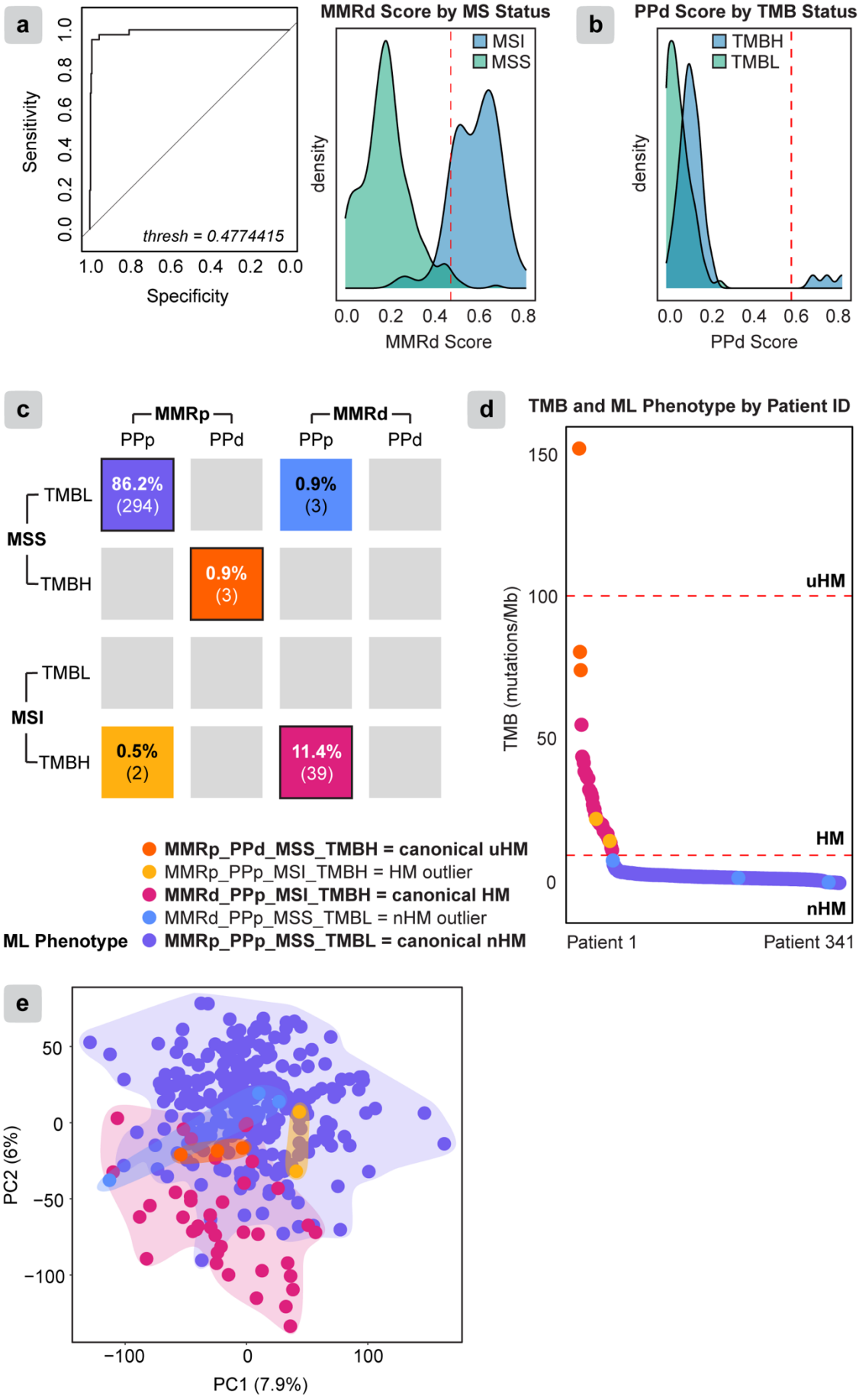

a
Sensitivity
Specificity
thresh = 0.4774415
MMRd Score by MS Status
MSI
MSS
density
MMRd Score
b
PPd Score by TMB Status
TMBH
TMBL
density
PPd Score
c
MMRp
MMRd
PPp
PPd
PPp
PPd
TMBL
MSS
TMBH
86.2% (294)
0.9% (3)
0.9% (3)
TMBL
MSI
TMBH
0.5% (2)
11.4% (39)
MMRp_PPd_MSS_TMBH = canonical uHM
MMRp_PPp_MSI_TMBH = HM outlier
MMRd_PPp_MSI_TMBH = canonical HM
MMRd_PPp_MSS_TMBL = nHM outlier
ML Phenotype
MMRp_PPp_MSS_TMBL = canonical nHM
d
TMB and ML Phenotype by Patient ID
TMB (mutations/Mb)
uHM
HM
nHM
Patient 1
Patient 341
e
PC2 (6%)
PC1 (7.9%)

Figure 3.1 | Defining mutational load (ML) phenotype in TCGA-COAD (n=341). (A) Receiver operator curve (ROC) analysis and Youden's index to identify a threshold (red dashed line) of MMRd score that designates samples as MMRp or MMRd based on known MS-status (coloured on the density plot of MMRd scores). **(B)** Density plot of PPd scores coloured by TMBL (<10muts/Mb) and TMBH (≥10Muts/Mb) groups with a red dashed line identifying selected threshold based on the plot. **(C)** Matrix of ML phenotype combinations identified indicating three predominant groups and two outlier groups. **(D)** Scatterplot of ordered TMB coloured by ML phenotype with dashed lines indicating HM-status cut-offs (nHM <10Muts/Mb, HM 10-100Muts/Mb, uHM >100Muts/Mb). **(E)** PCA plot of gene expression annotated by ML phenotype and a manually drawn shape surrounding the respective group. *[TMBL/H = tumour mutation burden high/low, MMRp/d = mismatch repair proficient/deficient, nHM = non-hypermutated, HM = hypermutated, uHM = ultrahypermutated]*

ML phenotype is described in this work as a combination of multiple ML descriptors and designate three predominant ML phenotypes of colon cancer; canonical ultrahypermutated (uHM), hypermutated (HM) and non-hypermutated (nHM) (Figure 3.1C-E). Two outlier groups were also described (Table 3.2). 2 patients had MSI and TMBH phenotypes but were classified as MMRp and PPp and clustered together with canonical nHM following principal component analysis (PCA; Figure 3.1E) likely as an artefact of being just below the MMRd threshold and slightly above the TMB threshold. 3 patients had evidence of MMRd mutational processes (MMRd scores > threshold) but with nHM features, one of which demonstrating values just under the thresholds needing to satisfy MSI and TMBH and clustering together with canonical HM patients (Figure 3.1E). Despite using HM-status as the cardinal ML output, these outlier groups weren't focused on in downstream analyses.

Table 3.2 | Outlier ML groups and their ML-descriptor values

ML Phenotype	Patient ID	MMRd score[a]	PPd score[b]	MSI Sensor[c]	MSI Mantis[d]	MS Status	TMB[e]
MMRp PPp MSI TMBH **(HM outlier)**	TCGA-G4-6309	0.401722	0.041972	26.62	0.9341	MSI	15.03
	TCGA-AZ-4313	0.267458	0.01235	NA	0.7225	MSI	22.81
MMRd PPp MSS TMBL **(nHM outlier)**	TCGA-A6-6648	0.505947	0.000	0.52	0.3926	MSS	2.03
	TCGA-AA-3519	0.525787	0.227052	0.69	0.3647	MSS	0.47
	TCGA-F4-6703	0.681707	0.165722	1.00	0.4396	MSS	8.17

Thresholds: [a]0.477, [b]0.6, [c]3.5, [d]0.4, [e]10

Furthermore, to obtain more complete data TMB was calculated by dividing SNV counts by the average coding exon size of 36Mb and whilst correlating with other sources (cBioportal $r=0.967$, $p<2.2x10^{-16}$, (Thorsson et al., 2018) $r=0.990$, $p<2.2x10^{-16}$) they were found to be conservative estimates (1.65-fold and 1.58-fold lower on average respectively). This explains the discordance on the TMB plot (Figure 3.1D) which designate 2/3 patients as HM despite PPd-status (TMB >100mut/Mb in cBioportal annotations indicative of uHM) and were thereafter referred to as uHM. This also explains the TCGA-F4-6703 outlier as this patient had 17.83 TMB in cBioportal annotations suggesting the conservative estimate was the reason for the nHM designation.

To assess whether immune-related heterogeneity is associated with each ML phenotype (specifically the canonical nHM phenotype) an immune score was calculated using ConsensusTME (Jimenez-Sanchez et al., 2019). This tool integrates multiple TME cell estimation methods tailored to TCGA datasets and employs single-sample gene set enrichment analysis (ssGSEA) to generate normalised enrichment scores (NESs) for immune deconvolution on RNAseq data. The resulting immune score reflects a composite gene set derived from various immune cell types, representing the overall immune content within the TME. When TCGA-COAD samples were plotted by TMB alone and annotated with ML phenotype, there is clear distinction between canonical nHM, HM and uHM groups (Figure 3.2A). However, when immune score is computed as an additional variable the canonical ML populations diversify and demonstrate heterogeneity in estimated immune cell prevalence. Immune score for canonical HM samples (median 0.0650) was statistically significantly higher than canonical nHM samples (median -0.00179) (two-sample t-test, $p=3.11x10^{-9}$), however some nHM samples exhibited similar immune scores to HM. Onto the same plot, additional parameters were overlayed to provide a visual representation of clinical variables with reference to spatial orientation of ML phenotype (Figure 3.2B). Canonical nHM samples appeared late stage, genomically unstable and CMS2-4 in comparison to canonical HM samples which appeared early stage and CMS1.

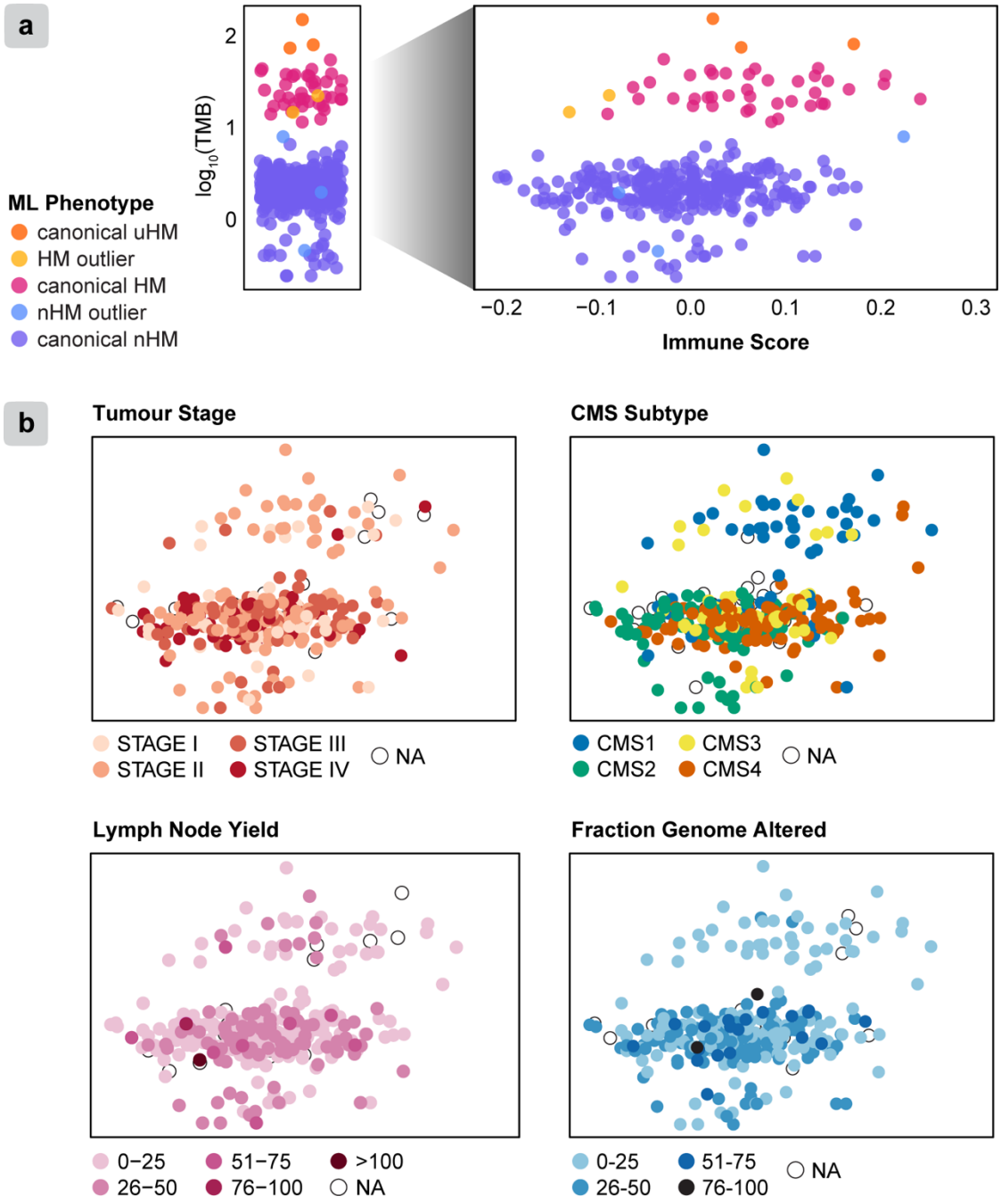
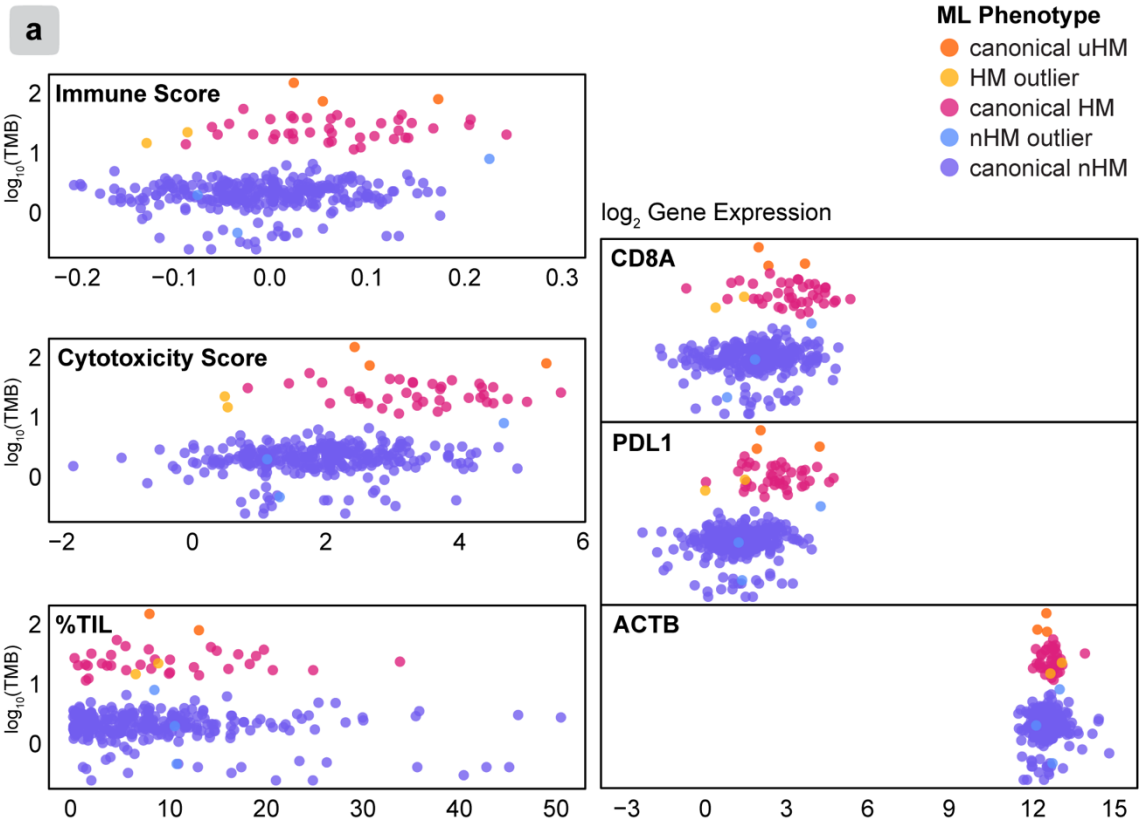
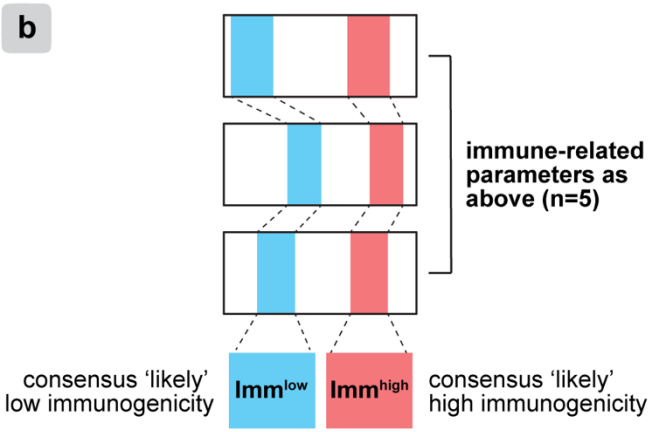


Figure 3.2 | Stratifying TCGA-COAD samples by Immune Score irrespective of ML phenotype uncovers a spectrum of immunogenicity that remains hidden when stratifying by TMB and ML phenotype alone (n=341). (A) TMB status of TCGA-COAD samples represented as dots and coloured by ML phenotype with and without the x-axis variable of immune score computed from ConsensusTME. A high immune score implicates an abundance of immune cell transcripts from gene expression data. **(B)** The same plot was further annotated by additional parameters downloaded as accompanying clinical data from TCGA, with samples plotted on the same axis enabling dots to retain the same spatial arrangement for ease of reference to plot A. *[TMBL/H = tumour mutation burden high/low, MMRp/d = mismatch repair proficient/deficient, nHM = non-hypermutated, HM = hypermutated, uHM = ultrahypermutated, CMS = consensus molecular subtype]*

In addition to immune score, we sought to characterise immunogenicity with a combination of parameters described in Supplementary Methods 2.4. Cytotoxicity score, TIL infiltration (%), CD8A expression and PD-L1 expression provided a similar degree of heterogeneity amongst TCGA-COAD samples, particularly in comparison to control parameters such as ACTB housekeeper expression (Figure 3.3A). Interestingly, the only parameter which demonstrated no significant differences between nHM and HM was %TIL infiltration (medians 6.45% vs 8.07% respectively, wilcoxon rank sum p=0.3064).

Taking all five parameters into account, TCGA-COAD samples were classified into groups of inferred immunogenicity dependent on specific criteria: if a sample's score was identified in the top or bottom 15th percentile in at least three parameters it was classified as Imm^{high} or Imm^{low} respectively (samples not matching this criteria were not classified) (Figure 3.3B). The 15th percentile was chosen as an arbitrary cut-off to obtain sufficient n numbers for analysis but also investigate extreme ends of parameters. This analysis was conducted in both the full cohort (n=341) and nHM-only subset (n=294), with Imm^{low} and Imm^{high} groups displaying the expected distributional differences across all parameter scores (Figure 3.3C).





*samples identified in the top or bottom 15th percentile in **at least 3** immune-related parameters*

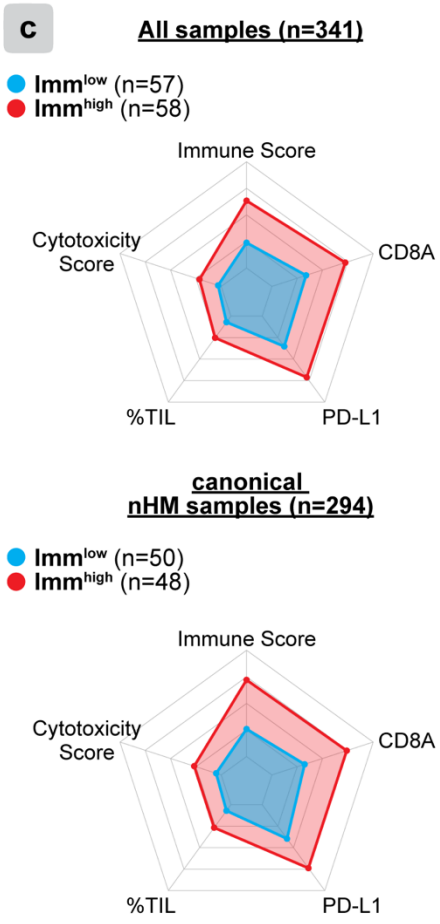


Figure 3.3 | Deriving a five-parameter based immunogenicity classification for assigning TCGA-COAD samples into Imm^{low} and Imm^{high} groups. (A) Scatter plots demonstrating scores of five separate immunogenicity parameters with each dot representing a TCGA-COAD sample (n=341) and coloured by ML phenotype. For visual representation, cytotoxicity scores were log2 transformed. **(B)** Schematic depicting assignment of TCGA-COAD samples into Imm^{low} and Imm^{high} groups. **(C)** Radar plots for the all-sample and canonical nHM cohorts demonstrating distribution of scores pertaining to each parameter. *[ML = mutational load, nHM = non-hypermutated]*

3.4.2 Mutation comparison between Imm^{low} and Imm^{high} groups reveals ZFHX4 as a potential mutational driver of immunogenicity

Following group assignment, mutational profiles of samples were compared between Imm^{low} and Imm^{high} groups derived from both the all-sample and nHM cohorts (Table 3.3). The percentage of samples in Imm^{low} harbouring a mutation in a given gene was compared against the percentage of samples in Imm^{high} harbouring a mutation in the same gene (denoted as % mutation frequency). If >1 mutation was found in the same gene in the same sample, the sample was still counted as a single 'mutated case' in binary fashion. Fisher's exact tests were performed to determine whether there was a non-random association between mutation status and Imm^{group} and thus test whether these proportions were significantly different, agnostic to mutation type. However, as a limitation, only nominal (NOM) p-values were reported and weren't adjusted (adj.) for multiple testing using false discovery rate (FDR) correction. Selecting the top 10 mutated genes in Imm^{high} and ordering by fold difference of %Imm^{high}/%Imm^{low} identified genes of interest that could be mutational drivers of immunogenicity (i.e. more mutations in Imm^{high} than Imm^{low}).

Mutated in 20% of all TCGA-COAD samples, *ZFHX4* displayed the largest fold difference in mutation frequency when comparing Imm^{high} and Imm^{low} groups from the all-sample cohort. Imm^{high} demonstrated a 3.11-fold higher ZFHX4 mutation frequency and this difference was significant (NOM p=0.006). Furthermore, following differential gene expression (DEG) analysis of *ZFHX4^{mt}* vs *ZFHX4^{wt}* samples, there was significant enrichment of the 'hot tumour signature' in *ZFHX4^{mt}* samples (adj.p-=5.86x10^{-6}) (Wang et al., 2021). Additionally, *ZFHX4* mutations were also significantly associated with clinical

benefit (CB) following immunotherapy in an independent pan-cancer dataset (Miao et al., 2018), reflected by a 2.34-fold higher *ZFHX4* mutation frequency in the CB group comparative to the no-CB (NCB) group (NOM p=0.002).

To accurately predict a mutational driver of immunogenicity, a TMBL setting must be employed which is why these comparisons were also performed in the nHM cohort. *ZFHX4* was the only gene to maintain these significant associations in the nHM cohort. Despite being in the nHM cohort, *ZFHX4* remained mutated at a relatively high frequency (12.24%) and exhibited the second-largest fold difference when comparing Imm^{high} and Imm^{low} groups. *ZFHX4* mutations were 4.17-fold more frequent in Imm^{high} comparative to Imm^{low} and this difference was statistically significant (NOM p=0.0488). Similarly, significant enrichment of the 'hot tumour signature' was found in $ZFHX4^{mt}$ samples (adj.p=0.0003) upon DEG analysis and $ZFHX4^{mt}$ samples were again significantly associated with clinical benefit from immunotherapy (NOM p=0.0261) in TMBL samples.

Linear regression was performed for each mutated gene to see if it associated with the immune score parameter irrespective of TMB, a parameter selected to represent all five for this model (Supplementary Methods 2.4, Supplementary Table 3.1). In both all-sample and nHM-only TCGA-COAD cohorts, *ZFHX4* regression effect sizes demonstrated a positive association of having a *ZFHX4* mutation with immune score (0.0173 and 0.0161-unit immune score increases respectively), though these did not reach significance. Logistic regression was performed for each mutated gene to see if it associated with the cohort-respective Imm^{high} or CB group. Again, *ZFHX4* demonstrated positive associations (odd ratios >1) in every analysis. In the canonical nHM TCGA-COAD cohort, patients with a *ZFHX4* mutation had 4.28-fold higher odds of being in the Imm^{high} group compared to those without the mutation, though this was just under significance (p=0.0811). Furthermore, the same trend was observed in the TMBL-filtered immunotherapy-treated cohort where patients with a *ZFHX4* mutation had 2.56-fold higher odds of having CB compared to those without the mutation (p=0.2).

Table 3.3 | Mutation comparison between Imm^{high} and Imm^{low} groups of all and nHM-only TCGA-COAD samples, with reference to GSEA results of mt vs wt hot tumour signatures as well as an independent immunotherapy-treated cohort of pan-cancer samples

All samples in TCGA-COAD cohort (n=341)						**All samples in immunotherapy cohort[a]** (n=249)			**Canonical nHM samples in TCGA-COAD cohort** (n=294)						**TMBL samples in immunotherapy cohort[a]** (n=149)		
Gene	Mt freq (%)	% mt in Imm^{high} (n=58) vs Imm^{low} (n=57)		Fold diff. [p-value]*	GSEA NES** of mt vs wt hot tumour sig.[b] [adj.p value]	% mt in CB (n=70) vs NCB (n=123)***		Fold diff. [p-value]*	Gene	Mt freq (%)	% mt in Imm^{high} (n=48) vs Imm^{low} (n=50)		Fold diff. [p-value]*	GSEA NES** of mt vs wt hot tumour sig.[b] [adj.p value]	% mt in CB (n=31) vs NCB (n=81)***		Fold diff. [p-value]*
ZFHX4	19.06	33	11	3.11 **[0.0060]**	2.19 **[5.86x10^{-6}]**	34	15	2.34 **[0.0020]**	**DNAH5**	12.59	19	4	4.69 **[0.0262]**	1.78 **[0.005]**	16	14	1.19 [0.7662]
FAT4	21.11	38	12	3.09 **[0.0023]**	2.18 **[1.88x10^{-5}]**	39	24	1.58 **[0.0487]**	**ZFHX4**	12.24	17	4	4.17 **[0.0488]**	1.97 **[0.0003]**	19	5	3.92 **[0.0261]**
PIK3CA	28.74	40	14	2.83 **[0.0029]**	1.83 **[0.0036]**	10	4	2.46 [0.1245]	**FBXW7**	11.56	23	8	2.86 [0.0513]	1.67 [0.0736]	0	1	0.00 [1.0000]
OBSCN	17.01	29	11	2.78 **[0.0185]**	2.36 **[1.97x10^{-5}]**	23	20	1.17 [0.5847]	**MUC16**	18.37	23	8	2.86 [0.0513]	1.84 **[0.0019]**	42	31	1.36 [0.2751]
SYNE1	25.22	40	16	2.51 **[0.0064]**	2.10 **[2.17x10^{-6}]**	30	15	2.05 **[0.0149]**	**TTN**	35.03	42	28	1.49 [0.2034]	1.90 **[0.0005]**	48	42	1.15 [0.6707]
MUC16	25.81	43	18	2.46 **[0.0043]**	2.08 **[2.78x10^{-5}]**	64	51	1.26 [0.0973]	**PIK3CA**	27.21	29	20	1.46 [0.3508]	1.71 **[0.0141]**	10	4	2.61 [0.3446]
TTN	42.52	59	32	1.86 **[0.0049]**	2.31 **[2.00x10^{-8}]**	70	59	1.18 [0.1635]	**TP53**	64.97	71	62	1.14 [0.3976]	1.65 [0.0698]	29	25	1.18 [0.6372]
KRAS	42.52	43	40	1.07 [0.8506]	0.94 [0.8004]	16	2	6.44 **[0.0011]**	**KRAS**	43.20	46	42	1.09 [0.8389]	1.24 [0.4500]	16	2	6.53 **[0.0169]**
TP53	59.82	53	63	0.85 [0.3460]	0.78 [0.7892]	39	26	1.48 [0.0757]	**SYNE1**	18.37	19	18	1.04 [1.0000]	1.90 **[0.0007]**	13	5	2.61 [0.2137]
APC	72.43	60	77	0.78 [0.0701]	-2.09 **[8.36x10^{-6}]**	11	9	1.28 [0.6196]	**APC**	76.87	60	78	0.77 [0.0797]	-1.49 [0.0603]	10	6	1.57 [0.6826]

*Ordered by fold difference of % mt Imm^{high} vs Imm^{low} samples of respective TCGA-COAD cohort. Bold p-values indicate significance ($p < 0.05$). *Fisher's exact test with nominal p-values (not FDR adjusted), **NES with respect to mt samples, ***CB = CR/PR, NCB = PD (SD ignored), [a](Miao et al., 2018), [b](Wang et al., 2021). [Mt = mutation, Wt = wildtype, GSEA = gene set enrichment analysis, NES = normalised enrichment score, CB = clinical benefit, NCB = no clinical benefit, CR = complete response, PR = partial response, PD = partial disease, SD = stable disease]*

3.4.3 Clustering patients into Immunogenicity Classification of Colon Cancer Subtypes (ICS1-5) offers an alternative method to identify somatic mutations in distinct immune-biological subgroups

275 TCGA-COAD patients were utilised for clustering based upon data availability of 345 parameters relating to mDOI (Table 2.15, Methods 2.10.3). k-means clustering by ConsensusClusterPlus (CCP) (Wilkerson and Hayes, 2010) was employed to identify distinct immune-biologically relevant subgroups of COAD and achieved 5 clusters – an evaluation based on stability and biological interpretability. Optimal k was selected based on delta area, demonstrating a flattening curve for consensus cumulative distribution function (CDF) and a diagonal block consensus matrix at k=5 (Figure 3.4A-C). To measure cluster stability, the CCP package generated consensus scores which are average consensus values for all pairs of items within the same cluster. Scores close to 1 indicate items in the cluster are always grouped together whilst scores close to 0 indicate items rarely grouped together across multiple iterations. At k=5, 3/5 clusters had good stability (mean consensus score >0.8) and 2/5 had moderate stability (>0.6) (Figure 3.4E). PCA of the scaled clustering matrix coloured by cluster depicted clustering success as their respective assignment using the first three principal components (PCs) in 3D space is distinct. Plotting a PCA of gene expression and colouring by cluster demonstrated spatial closeness of same-cluster samples indicating these subtypes were relevant biologically (Figure 3.4D). Furthermore, this is echoed by significant survival differences between all clusters (Figure 3.6). Whilst cluster stability validation is important, biological interpretability matters too and the 5 clusters stratify into ML phenotype-specific groups of varying immunogenicity. This is consistent with current literature and permits the interpretation of differences in immunogenicity independent of ML phenotype. For example, given some HM patients fail to respond and some nHM patients do respond to immunotherapy, these data may help infer potential underlying explanations.

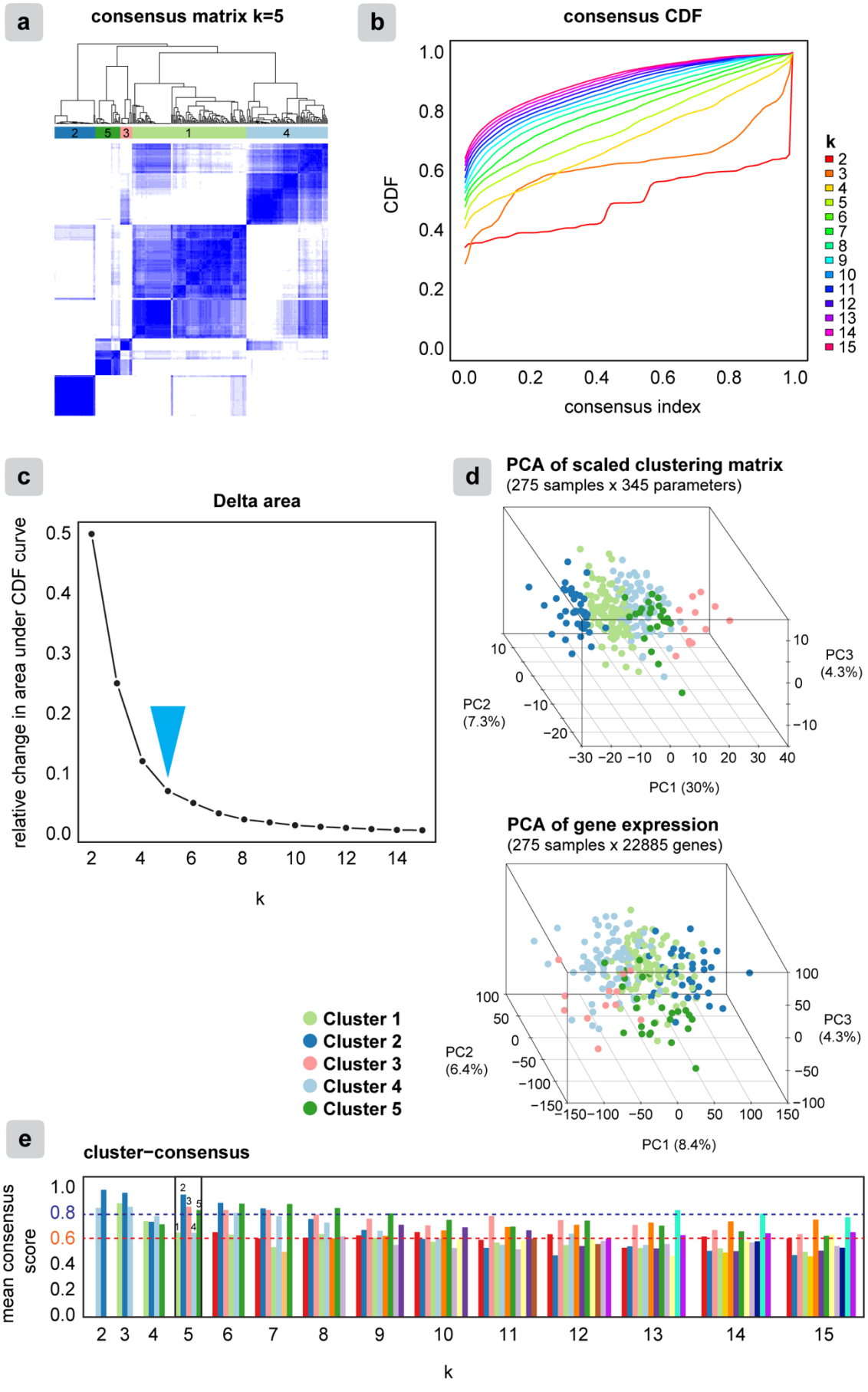

a
consensus matrix k=5
2
5
3
1
4
b
consensus CDF
CDF
consensus index
k
c
Delta area
relative change in area under CDF curve
k
d
PCA of scaled clustering matrix
(275 samples x 345 parameters)
PC1 (30%)
PC2
(7.3%)
PC3
(4.3%)
PCA of gene expression
(275 samples x 22885 genes)
PC1 (8.4%)
PC2
(6.4%)
PC3
(4.3%)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
e
cluster−consensus
mean consensus score
k

Figure 3.4 | Internal validation of cluster number selection and consequent identification of TCGA-COAD subgroups following k-means clustering on multiple determinants of immunogenicity (mDOI). (A) Consensus matrix heatmap produced as a result of ConsensusClusterPlus, depicting the proportion of times each pair of samples were clustered together across all subsampling iterations. Clear blocks along a diagonal demonstrate strong, stable clusters. The dendrogram describes sample-sample relationships, colouring samples by cluster. **(B)** Consensus Cumulative Distribution Function (CDF) is a plot representing consensus values across all sample pairs for each k. A flatter CDF curve/plateau indicates more stable and distinct clustering. **(C)** Line plot representing the area under the CDF curve as a readout of clustering stability as k increases. The selection of k=5 is highlighted by the triangle, chosen based upon the elbow method (the point after which increases in stability drop off) as the most parsimonious and stable k. **(D)** PCA plots representing the scaled clustering matrix or gene expression, coloured by cluster identity. **(E)** The average consensus score within each cluster, representing how often members of the same cluster were grouped together during resampling (values >0.6 indicate moderate stability, >0.8 indicate good stability), grouped by k iterations.

After validating cluster number selection and identification, the resulting clusters were then characterised using both the original clustering parameters and additional external measures (Figure 3.5). Cluster naming of Immunogenicity classification of Colon cancer Subtypes (ICS1-5) followed the order of cytotoxicity score, the ultimate effector for tumour immunogenicity as discussed in the introduction (these were coded from cluster numbers 3, 5, 4, 1, and 2 respectively in Figure 3.4).

ICS1 and ICS2 were predominantly canonical HM and demonstrated two extremes of immunogenicity, compared to ICS3-5 which were nearly entirely canonical nHM and also displayed extremes of immunogenicity within the same ML phenotype. Antigen presentation machinery (APM) signature scoring and hot tumour signature scoring also followed the same pattern as cytotoxicity. Furthermore, tumour immune dysfunction and exclusion (TIDE) scores > 0 inferring samples of likely non-responders to immunotherapy and < 0 inferring samples of likely responders to immunotherapy (Jiang et al., 2018) exhibited the same correlation. ICS1 and ICS2 both appeared CMS1-rich and low tumour stage, consistent with HM-ML phenotype observed in Figure 3.2. ICS5 had the highest proportion of Stage IV tumours which were mostly CMS2-rich. ICS3 mostly comprised CMS4-rich tumours which were more early stage (I-II) comparative to ICS5. Taking together, these data indicate that ICS1 represents a HM subgroup with heightened immunogenicity relative to

ICS2. ICS3 is a nHM subgroup with heightened immunogenicity that slowly decreases to the extreme of ICS5, a subgroup with apparent poorest immunogenicity of all five. Furthermore, all groups exhibited biologically interesting features warranting deeper investigation.

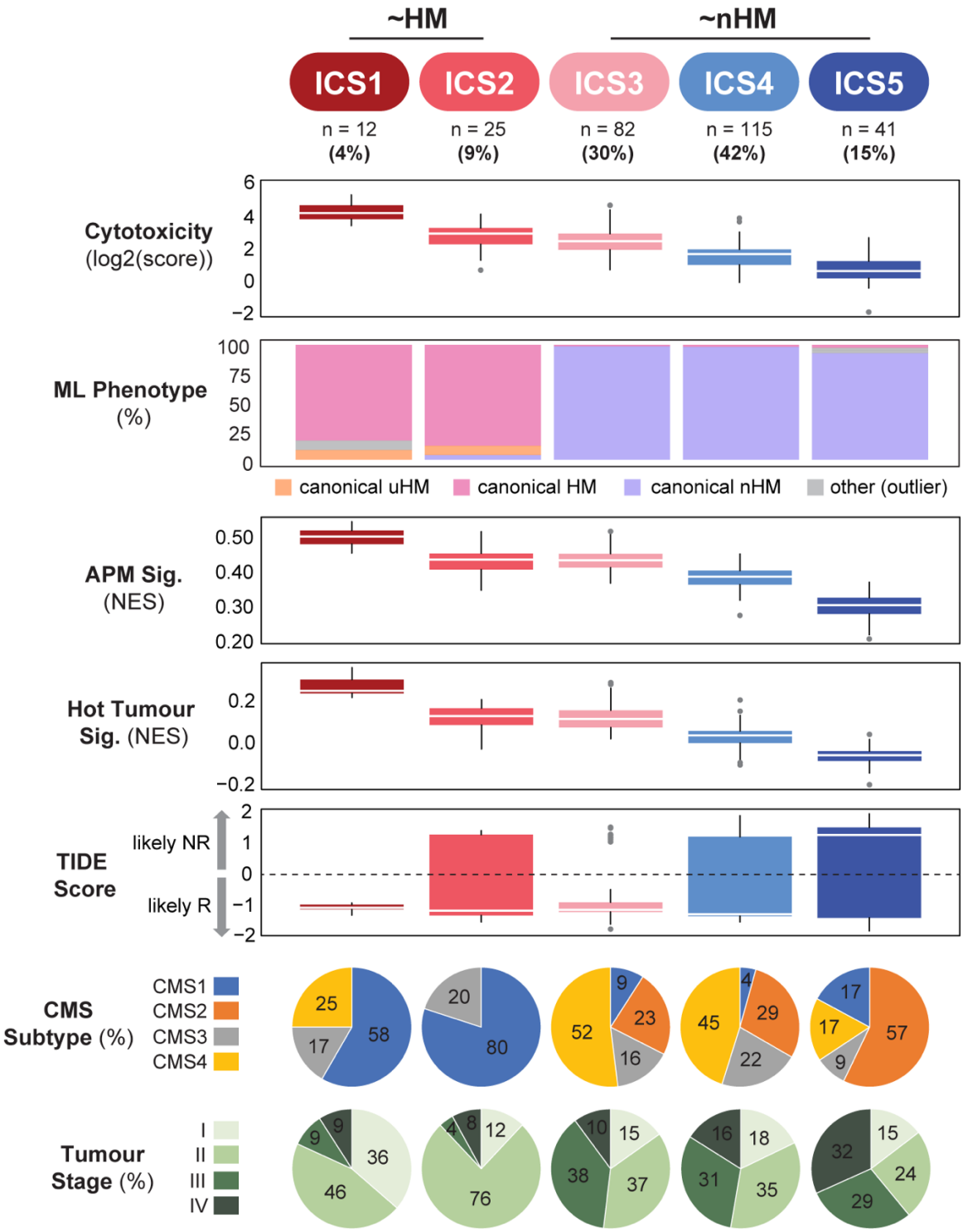


Figure 3.5 | Characterisation of TCGA-COAD clusters as Immunogenicity classification of Colon cancer Subtypes (ICS1-5). Multiple parameters used for clustering were used to initially characterise the biology of each cluster. Cytotoxicity scores (Rooney et al., 2015), mutational load (ML) phenotypes, antigen presentation machinery (APM) signature enrichment scores (Gong and Karchin, 2023) and hot tumour signature enrichment scores (Wang et al., 2021) were involved in clustering parameters and are presented as examples of key mDOI categories. Tumour immune dysfunction and exclusion (TIDE) scores, CMS subtypes (Guinney et al., 2015) and tumour stage were not employed in clustering but are useful to ascertain further biology. TIDE scores > 0 likely infer non-responders to immunotherapy and vice versa (Jiang et al., 2018). *[NR=non-responder, R=responder, CMS=consensus molecular subtype, HM=hypermutated, nHM=non-hypermutated, mDOI=multiple determinants of immunogenicity, NES=normalised enrichment score]*

To explore potential cell types inferred from tumour microenvironment deconvolution by ConsensusTME, normalised enrichment scores were plotted per cell type per cluster (Supplementary Figure S3.1). Generally, ICS1 displayed the highest enrichment of any immune cell type or cluster, with reduced enrichment for endothelial cell and fibroblast signatures that were both highest in ICS3. ICS5 demonstrated the lowest enrichment of any cell type, following a decreased trend from ICS4 and ICS3 with every cell type.

All ICS groups demonstrated survival differences which were significant for Progression-Free Survival (PFS) and Disease-Specific Survival (DSS) (log-rank p=0.0452 and p=0.0242 respectively) (Figure 3.6). Furthermore, ICS3 demonstrated significantly better PFS than ICS5 by log-rank measures (p=0.0485), and a 28% risk reduction in progression by cox proportional hazards which just missed significance (hazard ratio;HR=0.72, 95% confidence interval;CI[0.51-1.00], p=0.0527) (Figure 3.6B). ICS1 consistently had the worst Overall Survival (OS), PFS and DSS of any group, exhibiting 42%, 36% and 47% survival probability respectively at 5 years (Supplementary Table 3.2). This was echoed by a 2.6-fold increased risk of death (p=0.0277), 2-fold increased risk of progression (p=0.100) and significant 3.1-fold increased risk of disease-specific death (p=0.018) when compared to the rest of the cohort by cox proportional hazards (Supplementary Table 3.3). Conversely, ICS2 displayed the best survival of any group, with 72% patients not progressing for 5 years and a 2.7-fold lower risk of progression than the rest of the cohort, though this wasn't significant (HR=0.37, 95%CI[0.12-1.16], p=0.0878) (Supplementary Table 3.2/3).

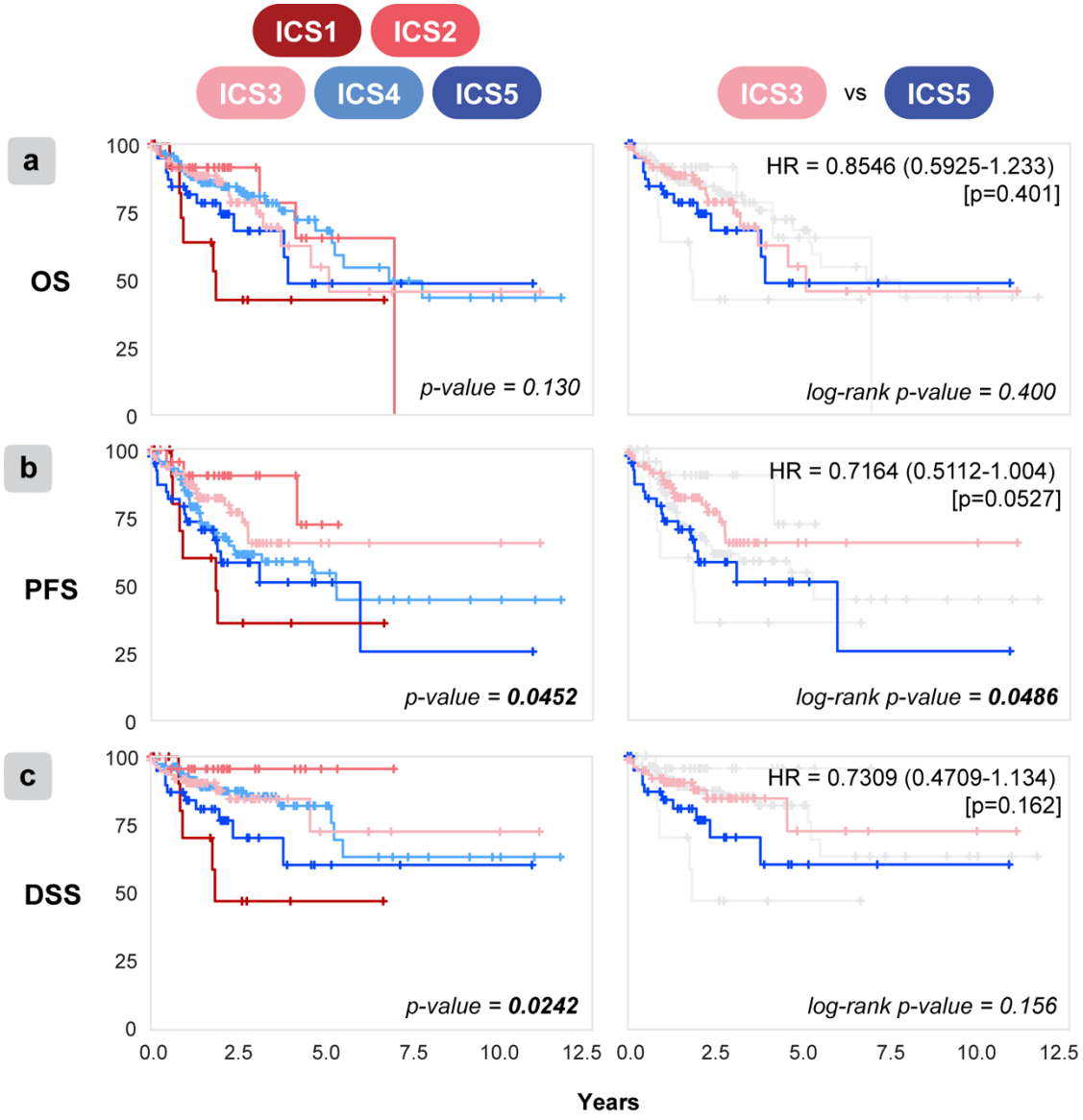


Figure 3.6 | Kaplan-Meier (KM) curves depicting Overall (OS; A), Progression-Free (PFS; B) and Disease-Specific Survival (DSS; C) of TCGA-COAD patients stratified by ICS group. KM curves on the left depict all five ICS clusters and p-values were produced following log-rank analyses. KM curves on the right are specific to the comparison between ICS3 and ICS5, depicting results from log-rank measures as well as hazard ratios (HR) from cox proportional hazards. Bold p-values are indicative of significance ($p < 0.05$).

It is important to note here that tumour-node-metastasis (TNM) stage was not adjusted in these survival analyses. Further work would include stratifying each ICS group into early (I-III) and late (IV) stage tumours before comparing survival differences, considering the strong association between stage and baseline immunogenicity as discussed in the introduction (Section 1.5).

Next, I sought to further characterise ICS groups to extrapolate relevant meaning. An ideal option for this would be DEG analysis, however this can only be performed between 2 groups so instead a PCA-centric strategy was devised. To interpret the biological meaning of each cluster, the predominant drivers of clustering were examined given all 345 mDOI-parameters were equally weighted despite being mostly (93%) transcriptomic derived. Figure 3.4D demonstrates a PCA plot produced on the scaled clustering matrix, indicating PC1, PC2 and PC3 explaining 30%, 7.3% and 4.3% variance respectively. These PC axes will reflect the predominant drivers of clustering thus interrogating these may provide insight into the biology underlying each ICS. As each PC is a weighted combination of the original variables, weight (loadings) can be extracted which indicate how much each original variable contributes to a PC. High positive or negative loadings suggest that variable has a strong influence on the PC. The top 20 positive and negative PC loadings were extracted and presented in Supplementary Table 3.4. Manual internet searches were performed to look for common biological associations with identified genes alongside different parameters such as scores and signatures used in clustering. This was further summarised in Table 3.4. It appears PC1/2/3 represent distinct immune-genomic-stromal axes.

Table 3.4 | Summary of biological themes predominantly contributing to PC1, PC2 and PC3 in the PCA of scaled clustering matrix

PC	% Var.	Biological Theme	Positive PC Drivers	Negative PC Drivers
PC1	30	Immune infiltration and cytotoxicity/ immunogenicity vs genomic instability	"Hot" tumours with high T-cell infiltration and cytotoxicity (CD8, GZMA, PRF1, IFNG, APM, immune signatures)	"Cold" tumours with low immune activity demonstrating genomic instability and aneuploidy
PC2	7.3	Stromal enrichment and immune exclusion vs antigenicity	Stromal-rich, fibroblast/CAF-inflamed, TGF-β-driven, immune-excluded tumours (high ESTIMATE stromal, fibroblast signatures)	Hypermutated, MSI-high, neoantigen-rich tumours (high TMB, INDELs, APM expression, MMRd)

PC3	4.3	Stromal/myeloid TGF-β inflammation vs functional adaptive immunity	TGF-β-inflamed, fibroblast-rich, myeloid-infiltrated tumours (TGFB1/3, CCL2/3/7/8, ECM signatures)	Adaptive immune activity – T-cell co-stimulation, Th17/Th1 cytokines, antigen presentation (CD40LG, IL17A, HLA-A, CCR7, CD27)

As predicted, the primary positive driver explaining PC1 (30% variance) and thus clustering was immune infiltration and cytotoxicity in comparison to the primary negative driver which appeared to be absence of immune activity and enrichment of tumours with genomic instability and aneuploidy. PC2 described a stromal enrichment vs antigenicity axis, with high PC2 values describing stromal-rich likely-immune-excluded tumours and low PC2 values pertaining to tumours with high antigenicity but not necessarily immune-infiltrated. Contributing to 4.3% variance, the PC3 axis described differences between stromal- and myeloid-enriched tumours vs those with suggestions of functional adaptive immunity. The distinct differences in stromal and TGFβ signals between high PC2 and high PC3 scores warrant further investigation as the biological themes captured by each PC axis are broad and intended to suggest rather than definitively define underlying biology.

Taken together, the general immune-biology of ICS1-5 can be summarised (Table 3.5). Mean PC loading values were calculated for each ICS group which enabled an approximate positioning and biological anchoring on each PC1/2/3 axis, roughly equivalent to their spatial position on the PCA plot in Figure 3.4D. ICS1 and ICS2 appeared to display similar biology with respect to antigenicity (i.e. TMB) and CMS enrichment (mostly CMS1) as expected for HM tumours but intriguingly diverged dramatically in survival and prognosis. Key differences between the two included a substantial enrichment of immune infiltration and cytotoxicity in ICS1, a slightly higher degree of antigenicity in ICS2 and a much higher enrichment of stromal and myeloid inflammation in ICS2 via the PC3 axis. Comparing the two extremes of the nHM spectrum, ICS3 represented mostly early-stage tumours that were CMS4 rich and positioned with an enriched immune infiltration and cytotoxicity phenotype on the PC1 axis and stromal enrichment on the PC2 axis. Interestingly, ICS3 associated

with a functional adaptive immunity phenotype on the PC3 axis (the most of any group) and demonstrated better prognosis than ICS5. ICS5 appeared devoid of immune infiltration (reflected in Supplementary Figure S3.1) and activity, accompanying neutral positioning on both PC2 and PC3 axes due to absence of stromal infiltration and a nHM phenotype. ICS5 also demonstrated the most likely genomically unstable group with the lowest PC loading mean of -16.3 alongside enrichment of mostly late-stage tumours.

Table 3.5 | Proposed characterisation of ICS groups in TCGA-COAD, reflecting differences in multiple determinants of immunogenicity following clustering

	ICS1	ICS2	ICS3	ICS4	ICS5
ML pheno.	HM		nHM		
CMS	CMS1 rich		CMS4 rich	Intermediate	CMS2 rich
Stage	Mostly early (I/II)		Mostly early (I/II)	Intermediate	Mostly late (III/IV)
Prognosis	Very poor	Very good	Good	Intermediate	Poor
Immune infiltration and cytotoxicity ↑ **PC1 axis*** ↓ Genomic instability	Very high (21.2)	Moderate (4.88)	High (8.26)		
				Low (-3.37)	Very low (-16.3)
Stromal enrichment and immune exclusion ↑ **PC2 axis*** ↓ Antigenicity			High (3.09)	Neutral (0.789)	Neutral (-0.741)
	Very low (-6.11)	Very low (-9.60)			
Stromal/myeloid TGF-β inflammation ↑ **PC3 axis*** ↓ functional adaptive immunity	High (1.52)	Very high (4.4)		Neutral (-0.675)	Neutral (0.874)
			Low (-1.05)		

**values in brackets are indicative of mean PC loadings for that ICS group*

3.4.4 Mutation comparison between ICS3 and ICS5 groups revealed ZFHX4 as a potential mutational driver of immunogenicity

Mutational profiles were identified for each ICS group (Supplementary Table 3.5) in the same way as Section 3.4.2. With a specific focus on ICS3 vs ICS5 to investigate potential mutational drivers of immunogenicity in a nHM setting, mutations were compared (Table 3.6). Hypothesising that these mutations will occur more frequently in ICS3 than ICS5, the top 10 frequently mutated genes in ICS3 were selected and ranked by fold difference of mutation frequency compared to ICS5. *ZFHX4* was again identified as a potential mutational driver, demonstrating a 2.17-fold higher *ZFHX4* mutation frequency in ICS3 compared to ICS5, however this enrichment didn't meet significance (NOM p=0.2586). This likely reflected unequal group sizes as ICS3 included nearly twice as many patients as ICS5, reducing statistical power and increasing noise in frequency comparisons. For example, when significance was reached in the previous analysis of Imm^{high} vs Imm^{low} (albeit with a greater fold difference), these groups had roughly identical size. To note, all *ZFHX4* mutations identified in ICS3 and ICS5 were missense.

Table 3.6 | Comparison of ICS3 (n=82) vs ICS5 (n=41) mutational profiles. The top 10 mutations in ICS3 were selected and then ordered by fold difference in mutation frequency to ICS5 with respect to ICS3

Gene	Mut freq. in ICS3 (%)	Mut freq. in ICS5 (%)	Fold difference in mut freq.	Significance*
ZFHX4	15.85	7.32	2.17	0.2586
PIK3CA	35.37	19.51	1.81	0.0951
SOX9	20.73	12.20	1.70	0.3213
MUC16	21.95	14.63	1.50	0.4697
SYNE1	29.27	21.95	1.33	0.5178
TTN	42.68	34.15	1.25	0.4361
FAT4	17.07	14.63	1.17	0.8011
TP53	70.73	68.29	1.04	0.8360
KRAS	41.46	46.34	0.89	0.7000
APC	71.95	90.24	0.80	**0.0219**

Bold indicates significance (p<0.05). **Fisher's exact test with nominal p-values (not FDR adjusted)*

Across analyses, whether using a minimalist, targeted approach or more broader systems-level framework, *ZFHX4* mutations consistently showed a suggestive association with heightened immunogenicity even within the nHM setting. In addition to this, no mutation in *ZFHX4* in TCGA-COAD was predicted to form neoantigens defined as antigenic (presented to and recognised on the cell surface) in the tumour-specific neoantigen (TSNA) database v.2.0 (Wu et al., 2023a) in comparison to most other mutations denoted in the tables presented in this chapter, supporting the mutation quality over quantity theory. Most importantly, these *ZFHX4* mutations were frequent and significantly enriched in TMBL patients, displaying clinical benefit following immunotherapy in an independent pan-cancer cohort.

3.5 Discussion

Both bioinformatic approaches using TCGA-COAD revealed *ZFHX4* mutations to be associated with immunogenicity, however both have strengths and limitations.

3.5.1 A specific focus on the Imm^{group} approach

Firstly, despite their simplicity Imm^{high} and Imm^{low} groups were based on multi-faceted immune parameters – immune score and CD8A (quantity), cytotoxicity score (activity), PD-L1 (checkpoint tone) and TIL% (histology) capture different axes. However, the 15th percentile threshold was convenient but arbitrary, dichotomising potential continuous immune biology leaving many samples “unclassified”, cutting power and potentially biasing towards outliers. Additionally, cytotoxicity score, CD8A and immune score are features which are not independent, thus classifying on ≥3 out of 5 may overweight a single latent axis (inflamed vs non-inflamed) and underweight orthogonal signals (such as myeloid/TGFβ or inflamed-functional vs inflamed-non-functional). This could have been improved through other metrics such as IFNγ-related signatures or genes which correlate with immunotherapy response for multiple cancers with apparent functional adaptive immunity (Ayers et al., 2017) or computing an effector index taking T-cell exhaustion markers into account relative to activation markers such as cytotoxicity and IFNγ. Furthermore, these groups may not be biologically distinct. For example, if one sample meets thresholds for three parameters but lacks PD-L1 expression and another sample meets thresholds for four parameters including PD-L1 expression, both would be assigned Imm^{high} but could be different biologically. An improvement on this approach would be to build a composite immunogenicity score encompassing all metrics by z-scoring each and averaging. Different cut-offs could be tested iteratively to show robustness by reporting classification stability. Additionally, this continuous immunogenicity score could also be used in per-gene linear models adjusted with covariates such as TMB to enable greater inclusion of samples for mutational driver investigation.

On the other hand, previous studies have used single scores or parameters to successfully identify mutational drivers of immunogenicity. For example, using the cytotoxicity score in TCGA pan-cancer analyses, Rooney et al. found *CASP8* and *B2M* mutations positively associated with cytolytic activity and *CD8A* expression, linking them to immunogenicity (Rooney et al., 2015). Rather than defining 'immune-high' within a TMBL subset, the authors used cytotoxicity score as a single metric. They applied a regression-based model that controlled for non-silent point mutation rate (TMB) across all samples. A similar approach could have been tested in this thesis, alongside subgroup analyses using the previously proposed composite immunogenicity score. This could show superior results to the immune score-based (single metric example) linear regression performed in this work demonstrating little significance (Supplementary Table 3.1). It would also be valuable to compare mutational drivers between TMBH Imm^{high} and Imm^{low} groups. Similar analyses have already been reported in MSI tumours, contrasting inflamed and non-inflamed colon cancers (Ishino et al., 2023, Kim et al., 2021), and parallel differences were observed between my ICS1 and ICS2 subgroups (Figure 3.5). Another study also employing TCGA and the TIL score metric as by here (Saltz et al., 2018) found *NF1* and *RB1* mutations enriched in TIL infiltrated glioblastoma samples (Rutledge et al., 2013). A similar TIL score was also used in a CRC dataset which found significant associations of HLA-related mutations in TIL-high MSS subsets by chi-square testing (Giannakis et al., 2016). The *CD8A* metric has also been used bioinformatically within TCGA to investigate mutations associated with immunogenicity (Brown et al., 2014). Taken together, this approach has limitations but has been used previously to identify mutational drivers of immunogenicity. These examples support its validity despite its simplicity.

3.5.2 A specific focus on the ICS clustering approach

The second bioinformatic approach offered a more holistic method to study immunogenicity in colon cancer. It not only identified potential mutational drivers but also revealed biologically meaningful subtypes with translational relevance and opportunities to explore

additional biomarkers. Key strengths included its breadth, as 345 mDOI features offered a large multi-parameter immune map to reflect composite immune biology compared to a single axis/few axes as with the initial approach. This approach showed strong internal validation, with PCA plots confirming separation in expression space and survival differences. These findings align better with biological reality and clinical behaviour, unlike the Imm-grouping method which also excluded 70% samples. Additionally, clustering with ConsensusClusterPlus permitted integration of both transcriptomic and genomic data. Such combined approaches are rarely reported in the current literature. For example, colon cancer and even TCGA-COAD has been stratified into immune subtypes before using ConsensusClusterPlus and other similar clustering approaches, providing ideas for parameter choice in my work (Table 3.7). However, most employ a solely transcriptomic-based set of ‘immune-related’ genes and/or immune deconvolution signatures as their parameters to describe immunogenicity.

Table 3.7 | **An exhaustive chronological selection of previously published immune subtypes of colon cancer employing a clustering-style approach**

Study	Discovery cohort [n]	Validation cohort [n]	Clustering parameters	Immune deconv. method	Clustering method	Subtypes		IO valid.?	*In vitro* follow-up?
						#	ML pheno*		
Lal[1]	TCGA-COAD [n/a]	n/a	Discrete clinical variables e.g. mut status	n/a	Unsupervised hierarchical	4	1 MSI, 3 MSS	✗	✗
Tang[2]	TCGA-CRC [618]	1 GSE set [316]	4 cell-type ssGSEA signatures and PDL1 expression	ssGSEA of immune cell gene sets	Unsupervised not specified	4	2 MSS, 2 mixed	✗	✗
Zheng[3]	TCGA-CRC [488]	3 GSE sets [1015]	150 immune-related genes derived from NMF	ssGSEA of immune cell gene sets	NMF, clusterProfiler and NTP	3	mixed	1 set¶	✗
Lin[4]	TCGA-COAD [447]	1 GSE set [176]	29 immune signatures by ssGSEA	TIMER	Unsupervised hierarchical	3	n/a	✗	Yes (H19)
Liu[5]	7 GSE sets [1198]	TCGA-CRC [638]	2798 immune-related genes	ssGSEA of immune cell gene sets	Unsupervised k-means (CCP)	2[ab]	mixed	3 sets¶	✗
Mao[6]	TCGA-CRC [619] and 7 GSE sets [1802]	1 indep. cohort [447]	31 TME cell signatures	CIBERSORT and MCPcounter	NMF	3[ab]	1 MSI, 2 MSS	✗	✗
Sun[7]	TCGA-COAD [433]	1 GSE set [177]	8 immune cell signatures	MCPcounter	Unsupervised k-means	3[a]	1 mixed, 2 MSS	✗	✗
Xu[8]	TCGA-COAD [437]	2 GSE sets [706]	17 immune cell signatures (prognostic in >2 datasets)	IOBR integration e.g. CIBERSORT, MCPcounter…	Unsupervised PAM (CCP)	3[ab]	mixed	✗	✗
Sugawara[9]	1 indep. colon cancer set [89]	n/a	1 IFNy related gene signature	MCPcounter	Unsupervised hierarchical (wards)	3	1 mixed, 2 MSS	✗	✗
Qi[10]	TCGA-COAD [424]	1 GSE set [512]	1921 immune-related genes	ssGSEA of immune cell gene sets	Unsupervised PAM (CCP)	3[ab]	2 mixed, 1 MSS	✗	✗

Wang[11]	TCGA-COAD and 10 GSE sets [1881]	n/a	6238 genes (MAD >0.5 in one cohort)	ssGSEA of immune cell gene sets	Unsupervised k-means (CCP)	3[a]	1 MSI, 2 MSS	2 sets¶	Yes (GBP2)
Wu[12]	TCGA-COAD, CPTAC and 8 GSE sets [2196]	n/a	61 TME-related signatures	MCPcounter and CIBERSORT	Unsupervised PAM (CCP)	4[abc]	1 MSI, 3 MSS	4 sets	✗
Borràs[13]	1 indep. colon cancer set** [63]	TCGA-COAD [376]	Signalling modules based off IFNγ$^+$CD8$^+$ T-cell transcriptomics, separately for MSS & MSI	quanTIseq	Unsupervised hierarchical (MSI & MSS separately)	7	2 MSI, 5 MSS	2 sets	✗
Pei[14]	TCGA-CRC [647]	1 GSE set [232]	200 'hallmark inflammatory' genes	ssGSEA of immune cell gene sets	NMF	2[ab]	1 mixed, 1 MSS	✗	Yes (RIPK2)
Li[15]	TCGA-CRC [563] and 5 GSE sets	n/a	28 immune cell signatures + 17 immune-related signatures	xCell	Unsupervised hierarchical	3[a]	1 mixed, 2 MSS	2 sets	Yes (IGF1Ri)
Acha-Sagredo[16]	1 indep. CRC set*** [59]	TCGA-CRC [389] and CPTAC [104]	7 manually curated gene signatures consistently differing between MMRp and MMRd tumours	Derived from orthogonal scRNAseq	Unsupervised hierarchical	4 (5 in val cohorts)	1 mixed, 2 MSI, 1 MSS (1 mixed, 1 MSI, 3 MSS in val cohorts)	This was an IO cohort	Yes (CD74)

*rough approximations for simplicity, **the only scRNAseq-based study, ***the only spatial transcriptomics-based study, a=DEG comparisons, b=DMG comparisons, c=DmeG comparisons, ¶=SubMap analysis, [ ]=n numbers

[1](Lal et al., 2015), [2](Tang et al., 2020), [3](Zheng et al., 2021), [4](Lin et al., 2021), [5](Liu et al., 2021), [6](Mao et al., 2022), [7](Sun et al., 2022), [8](Xu et al., 2022), [9](Sugawara et al., 2022), [10](Qi and Zhang, 2022), [11](Wang et al., 2022), [12](Wu et al., 2023b), [13](Borras et al., 2023), [14](Pei et al., 2024), [15](Li et al., 2025), [16](Acha-Sagredo et al., 2025)

[DEG/DMG/DmeG=differential expressed/mutated/methylated gene analysis, CCP=consensus cluster plus, NMF=non-matrix factorisation, GSE=gene set expression from gene expression omnibus, ML=mutational load, IO=immunotherapy dataset, IO=immunotherapy-treated cohort]

The largest and most comprehensive effort to date to define immune subtypes across cancers is the ‘Immune Landscape of Cancer’ (Thorsson et al., 2018) which used TCGA data to identify six pan-cancer subtypes. Whilst this work was seminal and provided the principles of many mDOI parameters in my clustering analysis (Table 2.15), TCGA-COAD samples were predominantly classified into only two of these six subtypes highlighting the need for a more refined, disease-specific approach focused solely on colon cancer. Since this work, several colon cancer-centric studies have classified immune subtypes (Table 3.7). My work incorporated ideas and principles from these studies but diverges on the integration of genomic-based data to supplement transcriptomic-based data to describe immunogenicity on a more holistic level based on mDOI.

The 16 studies in Table 3.7 classified immune subtypes in colon cancer using a range of approaches spanning bulk-RNAseq, scRNAseq and spatial transcriptomics. These studies differ in their methodological validity, reproducibility and clinical relevance but across nearly all of them, TCGA-COAD was consistently used as either a discovery or validation cohort. Albeit this popularity may be a result of its well-annotated nature, this reaffirms its utility in dissecting immunogenicity and supports its use in my work to identify mutational drivers of immunogenicity. Despite employing different clustering methods and parameters representing immunogenicity or ‘immunity’, all converge on subtypes that stratify nHM tumours across multiple immunogenicity states. For example, nearly all bulk-RNAseq-based studies revealed MSS/TMBL samples segregate into both “immune-hot” and “immune-cold” subtypes. This is supported by a spatial transcriptomics study in Table 3.7 (Acha-Sagredo et al., 2025) which showed that IFNγ-high immunophenotypes contained both MSI and MSS tumours and were associated with better immunotherapy response. Importantly, their clustering validation in TCGA-COAD and CPTAC also yielded 5 immune subtypes with similar ML phenotype distributions to ICS subtypes identified by my work and exhibited a balanced mix of MSI and MSS samples in IFNγ-high groups. Furthermore, another study presented significant differences in estimated immune infiltration within both

MSS and MSI tumours, indicating that immune variation exists even within the same ML phenotype (Li et al., 2025). Moreover, an AI-based study using IHC-derived TIL patterns in a large pan-cancer immunotherapy cohort found that ~25% nHM tumours demonstrated an inflamed immune phenotype similar to HM tumours (Shen et al., 2024). This inflamed immune phenotype was positively prognostic of response to immunotherapy, and interestingly was prognostic of favourable response in MSS/TMBL but not in MSI/TMBH subgroups. scRNAseq work further confirms these patterns. One study found that MSS and MSI tumours differed in their $IFN\gamma^{+}CD8^{+}$ compartments and used this to stratify TCGA-COAD into multiple immune subtypes within identical ML subsets (Borras et al., 2023). Using random forest classification in an immunotherapy treated cohort, they further showed these subtypes had distinct responses to immunotherapy (Borras et al., 2023).

Collectively, these findings challenge traditional dogma that MSS is “immune-cold” and instead reveal clear immunological heterogeneity within the nHM setting. Of the 16 studies reviewed, 14 reported fewer than 5 immune subtypes, and most often 3. This likely reflects their reliance on transcriptomic data alone and omission of key features such as TMB, MMRd score and neoantigen load, which are critical for clinical stratification of colon cancer. In contrast, my work identifies five subtypes of immunogenicity following an integration of transcriptomic and genomic parameters which thus more clearly map onto clinically meaningful ML phenotypes. For example, supplying genomic parameters have led to 2-HM and 3-nHM subtypes rather than mixed groups, which would enable investigations within ML phenotypes of differing immunogenicity to explain differential response to immunotherapy.

The observation that TMBL and TMBH tumours are not immunologically distinct as single entities is consistently supported across multiple platforms comprising bulk-RNAseq, scRNAseq and spatial transcriptomics as well as across many novel, TCGA-based or GEO-based datasets. Furthermore, where immunotherapy response data were available, response varied significantly between subtypes within the same ML category reinforcing

that TMB alone is not sufficient to explain immunogenicity. Altogether, these data strongly support the validity and clinical relevance of my study's ICS subtypes and reflect a broader shift in the field toward recognising immune diversity within MSS/TMBL and nHM tumours.

All studies including my own show that immune-related subtypes carry distinct prognoses. This highlights the potential clinical actionability of *in silico* subtype classification, not only for understanding underlying tumour biology but also for predicting outcomes. For instance, it's now well-established that not all HM patients respond to immunotherapy, and defining biologically distinct immune subtypes can help explain why. With respect to chemotherapy treatment that applies to most TCGA-COAD samples, my analysis identifies a subset of HM patients with poor prognosis, consistent with findings in other studies (Tang et al., 2020). Specifically, ICS1 exhibited the poorest prognosis of any ICS subtype across all survival types (between 2-3-fold increased event risks following ICS1 vs rest comparisons with cox proportional hazards) yet presented the group of highest immunogenicity. Strikingly, both ICS1 and ICS2 constitute early-stage (I/II) HM tumours (82% and 88% respectively) but their 5-year PFS survival probabilities of 36% vs 72% differ dramatically (Supplementary Table 3.2). This strongly suggests that differences in biological phenotype or treatment interactions rather than stage are driving these divergent prognoses. However, this finding is counterintuitive because HM tumours are generally associated with favourable prognosis and high T-cell infiltration has even been shown to outperform MSI status as a prognostic factor (Galon et al., 2006, Mlecnik et al., 2016, Pages et al., 2018). Given ICS1 displayed the highest estimated T-cell infiltration and immunogenicity phenotype, as well as the considerably highest mean PC loading for PC1 towards immune infiltration and cytotoxicity (Table 3.5), its poor prognosis appears at first glance paradoxical. One possible explanation is "immune overdrive" where excessive immune activation leads to dysfunction or exhaustion (Fakih et al., 2019). Reports of exhausted T-cell populations in MSI immunotherapy non-responders (Borras et al., 2023) further highlight the need to compare activation and exhaustion markers between ICS1 and ICS2 within the mDOI framework.

Furthermore, HM tumours show differential sensitivity to certain adjuvant chemotherapeutics, such as resistance towards 5-fluorouracil (5-FU) (Ribic et al., 2003). This is classically attributed to deficient mismatch repair which fails to recognise 5-FU-modified DNA and thus attenuates 5-FU-induced cytotoxicity (Tajima et al., 2004, Carethers et al., 1999). Combinations of 5-FU with other chemotherapeutics such as oxaliplatin have prompted significant survival benefits in MMRd colon cancer patients compared to 5-FU alone (Tougeron et al., 2016). Given post-resection adjuvant regimens in TCGA-COAD vary, prognostic differences between ICS1 and ICS2 may partly reflect treatment heterogeneity rather than immunogenicity per se. Moreover, this paradox doesn't appear in the nHM setting as ICS3 had both highest immunogenicity and best prognosis compared to ICS5 which had the lowest immunogenicity and worst prognosis, suggesting that this phenomenon is ML-context specific. These findings underscore the need for deeper biological investigation, for example through differential gene expression and mutation analysis between ICS1 and ICS2 and also ICS1 vs ICS3 whilst controlling for TMB. These types of analyses have already been done before with success, for example one study demonstrated 61% MSI colon cancer patients had *TGFBR2* mutations (79% DFS at 5-years) and these had significantly better prognoses following fluorouracil-based chemotherapy than *TGFBR2*^{wt} MSI patients (40% DFS at 5-years, log-rank p=0.007) (Watanabe et al., 2001). Stratifying survival analyses by stage, as performed in some studies (Tang et al., 2020), would permit stage-matched comparisons between subtypes. This is important because ICS3 contains 13% more early-stage tumours than ICS5, which may account for the prognostic differences rather than underlying biology, and stage is closely associated with baseline immunogenicity as discussed in the introduction. Collectively, this argues for immunotherapy-treated cohorts as more appropriate for prognostic analyses, since immunogenicity is defined by its ability to predict response to such therapies. TCGA patients were not treated with immunotherapy thus this type of survival data can't be used, however the utilisation of external immunotherapy-treated datasets as described below is imperative.

Another interesting finding included enrichment of CMS4 tumours in ICS3 comparative to ICS5 and thus a weak association with *ZFHX4* mutations considering the 2.17-fold difference in mutation frequency which wasn't significant (Table 3.6). CMS4 tumours are characterised by stromal TGFβ activation, angiogenesis, immune exclusion and subsequent poor relapse-free and overall survival (Guinney et al., 2015). This pattern is consistent with immune-subtyping studies that reported CMS4 enrichment in poorly prognostic immune-related subtypes (Mao et al., 2022, Xu et al., 2022). However, other studies have demonstrated CMS4 enrichment in high immunogenicity subtypes (Acha-Sagredo et al., 2025, Li et al., 2025, Roelands et al., 2023) which reflects my analyses and provides more evidence of a CMS4 subset of nHM colon cancers that appear amenable to immunotherapy, particularly in combination with other agents (Acha-Sagredo et al., 2025, Li et al., 2025, Leonard et al., 2024, Takei et al., 2024). For example, one study demonstrated not all stromal signalling was completely immunosuppressive, showing comparable immunomodulatory TNFα signalling in CMS4 to the immunogenic CMS1 and preferential effects on the macrophage compartment following PD-1 inhibition (Leonard et al., 2024). Another study showed CSM4 was significantly associated with response (p=0.035) and PFS (p=0.006) in a colon cancer cohort treated with immunotherapy plus anti-VEGFR2 targeting (Takei et al., 2024). The effect was attributed to inhibition of M2 macrophage activation, reduced Treg recruitment and indirect release of immunosuppression on anti-tumour $CD8^+$ T-cells. Furthermore, cornerstone to CMS4, many TGFβ targeting strategies are now emerging which can remove the T-cell exclusion barrier rendering tumours susceptible to anti-PD-1/PD-L1 immunotherapy through immunosensitisation (Tauriello et al., 2018). Additionally, an immune-enriched fibrotic pan-cancer subtype has been described with similar responses following immunotherapy comparative to the classical HM-immune enriched non-fibrotic subtype (Bagaev et al., 2021). In the PCA (Tables 3.4/5), ICS3 separated along both PC2 (stromal enrichment and immune exclusion) and PC3 (functional adaptive immunity), whereas other nHM subtypes

showed little separation. The differing stromal and TGFβ signals between high PC2 and high PC3 scores remain unclear, warranting further investigation.

Taken together, CSM4 and thus potentially ICS3 enriched in *ZFHX4* mutations represent a complex stromal microenvironment that could be manipulated through immunosensitisation to permit and subsequently improve immunotherapy response. To note, the original CMS authors did find CMS4 associated with samples of low tumour purity and thus high stromal and immune infiltration scores as per ESTIMATE which were used in the clustering pipeline, highlighting the need for tumour purity investigation and potential adjustment in ICS-related analyses. However, it appeared that most studies in Table 3.7 did not control for this. ICS5 showed CMS2 enrichment, the lowest immune enrichment scores (Supplementary Figure S3.1) and a 90% predominance of APC mutations, consistent with a true immune-desert phenotype. This group requires further genomic and transcriptomic investigation, for example by ranking the top 10 ICS5 mutations against ICS3 to identify variants associated with poor immunogenicity.

Beyond the classifications in Table 3.7, other published subtypes such as CRIS (Isella et al., 2017), iCMS/IMF (Joanito et al., 2022) and CRPS (Nunes et al., 2024) could provide additional clinical context for my ICS subtypes. Linking ICS to these frameworks, as others have done with CMS (Sun et al., 2022), may strengthen their translational relevance. For example, Joanito et al. describe some colon MSS tumours as iCMS3 which display similarities to MSI, enrichment for immune cell signatures and anti-tumour cytotoxicity and it would be interesting to see if my ICS3 group is transcriptionally similar to this subtype (Joanito et al., 2022). Furthermore, Nunes et al. reports CRPS2 as containing CMS4 samples with longest OS and RFS and thus it would be interesting to see if ICS3 is related to this group utilising their ResNet50 CRPS-based classifier (Nunes et al., 2024).

Another strength to my work is the employment of TCGA-COAD-specific immune deconvolution by ConsensusTME (Jimenez-Sanchez et al., 2019) marking what is to my

knowledge the first application of this method for immune subtype classification in colon cancer. TME deconvolution tools such as CIBERSORT, MCPcounter and xCell are commonly employed to estimate the relative abundance of cell infiltrate from bulk-RNAseq data and have been used for clustering or characterisation post-clustering in all studies in Table 3.7. These approaches are central to quantifying immunogenicity given cellular composition of the TME strongly describes and subsequently influences immune response. ConsensusTME developed by Jimenez-Sanchez et al. is a consensus-based framework that integrates gene sets from seven leading deconvolution methods including CIBERSORT, xCell, TIMER and MCPcounter to generate robust consensus gene supersets for 18 immune and stromal cell types. With explicit validation using TCGA data, ConsensusTME was well suited for application in a clustering approach using TCGA-COAD data in my work.

In addition to ConsensusTME, another strength in my approach was the use of internal clustering validation using CDF and delta area for selecting optimal K and observing cluster stability, metrics of which are provided by the ConsensusClusterPlus package (Wilkerson and Hayes, 2010, Monti, 2003). Studies in Table 3.7 using ConsensusClusterPlus employed k-means clustering with similar metrics. However, cluster number selection often appeared subjective, with some studies providing no justification for their chosen value. Internal validation could have been strengthened by applying additional computational methods to assess consensus clustering outputs. For example, proportion ambiguous clustering (PAC) provides a quantitative measure of CDF curves for optimal K selection and has outperformed other indices such as GAP, silhouette width and CLEST (Senbabaoglu et al., 2014); it was also used in an immune-subtyping study (Liu et al., 2021). R packages are available which can compute many of these metrics and would have further supported my work (Charrad et al., 2014). Further validation of cluster numbers, sizes and distribution could also have been achieved using other datasets for external validation, and this is described below.

One limitation to my work lies in cluster parameter choice and selection (CPCS) which represents a double-edged sword as the rationale behind CPCS reflects both a strength through biological relevance and a limitation due to subjectivity in selection. The concept of mDOI introduced in Chapter 1 guided CPCS to represent key immunological features using available TCGA-COAD annotations to cover the basic principles of immunogenicity in a holistic manner. While clustering was unsupervised and tailored to capture various axes of immunogenicity through an improved method comparative to the initial bioinformatic approach, CPCS remained inherently subjective especially compared to the other studies in Table 3.7. These used a limited number of pre-defined immune genes or signatures offering more objectivity, albeit at the cost of biological complexity. An improvement to this CPCS process would involve a more systematic and formal literature review of mDOI-related features. In my work, selection was largely based on general literature exploration which may have missed some potentially informative or non-redundant parameters. Another important limitation is the risk of multicollinearity between input variables. Multicollinearity arises when multiple parameters are highly correlated and provide overlapping information (Usefi, 2022). This can be problematic in clustering which fundamentally depends on variation between features to define structure. If many variables are colinear, this can result in overemphasis of certain biological signals, reducing the diversity and interpretability of clusters. Despite efforts to use only unique genes in the expression-based parameters (Methods 2.10.3), some will also inevitably contribute to multiple gene signatures also used as parameters. For example, IFNG expression was included both as an individual gene and indirectly within several immune signatures related to cytotoxicity and T-cell activation. This introduces redundancy, which could distort clustering results by overrepresenting certain axes of immunogenicity such as cytotoxicity. Addressing this, further work could include (1) removing individual genes already represented in composite scores and signatures, (2) filtering out highly correlated variables using correlation matrices and thresholds (i.e. $r>0.8$) or parameters that show little variability across samples (i.e. median absolute deviation cut-offs), (3) reducing dimensionality

through PCA to collapse colinear variables into uncorrelated components and (4) construction of module eigengenes where a group of correlated features is summarised by its first PC thereby capturing domain signal without redundancy. While some correlation between variables is expected given all features relate to immunogenicity, it's the excessive multicollinearity that poses a risk and this would need to be investigated. Overrepresentation of a single latent axis can skew the clustering process and obscure other relevant biological dimensions. Another consideration is data origin imbalance, with 93% of input parameters derived from transcriptomic data. This raises questions about whether genomic features such as TMB were sufficiently weighted in the clustering. However, early analyses (data not shown) demonstrated even the small subset of genomic parameters meaningfully contributed to clustering structure, consistent with their prominence among top loadings for PC1–3 (Supplementary Table 3.4). This is likely due to their high variability across samples, particularly for metrics like TMB. Refinement of CPCS has been performed in other studies (Angelova et al., 2015, Thorsson et al., 2018) and further suggests its necessitation to provide enhanced validity to ICS groups.

Some studies in Table 3.7 extend their immune subtype classification work into *in vitro* validation based on biomarkers identified from their respective analyses, highlighting the translational potential of biomarker investigation *in silico*. It’s important to underscore that bioinformatic analysis primarily yields associative and predictive insights which require experimental validation to establish causality. While genomic biomarkers are less frequently pursued in this context, transcriptomic biomarkers have gained prominence in immunogenicity research, with recent studies successfully exhibiting functional links between epithelial gene expression and immune phenotype (Acha-Sagredo et al., 2025, Li et al., 2025). In support of this, the clustering approach offers great translational value in addition to identifying mutational drivers of immunogenicity. For example, in some preliminary analyses *IHH* expression was found significantly higher in ICS5 vs ICS3 and significantly inversely correlated with cytotoxicity score (cor = -0.4967, $p<2.2x10^{-16}$). IHH is

a known immunosuppressive factor in intestinal tissue (Westendorp et al., 2018) and led to the formation of an ongoing MSc project, highlighting the multiple functionalities of the clustering approach in generating testable hypotheses for downstream validation.

3.5.3 Evaluating both bioinformatic approaches

Arguably the largest limitation of my work is the lack of external validation in other independent colon cancer datasets across both bioinformatic approaches. As outlined in Section 3.1, several large cohorts such as Nunes (Nunes et al., 2024), SIDRA (Roelands et al., 2023) and CPTAC (Vasaikar et al., 2019) represent promising candidates for such validation and are further explored in Chapter 4 and 5. With respect to the Imm-grouping approach, external validation of Imm^{high} and Imm^{low} groups would be relatively straightforward. Subsequent mutational comparisons could test whether *ZFHX4* mutations are again enriched in Imm^{high} patients, reinforcing their potential role as a consistent mutational driver of immunogenicity. This approach would require minor adjustments, for example excluding TIL score as this parameter is only available for TCGA-based data.

With respect to the ICS clustering approach, external validation is more complex. As shown in Table 3.7, most published studies performed discovery-validation using multiple cohorts, yet the novelty and specificity of the mDOI parameters used in my work particularly make direct translation difficult. For example, neoantigen-based parameters from TNSAdbv2.0 (Wu et al., 2023a) contributed to the theme of ‘ability to express antigen’. A total of 8 such parameters were annotated from original studies because of their access to HLA data and specialised prediction tools, whereas most other parameters were calculated *de novo* based on my own transcriptomic or genomic data. Therefore, to ensure reproducibility and assess the biological generalisability of the ICS groups, two complementary strategies could be pursued: (1) parameters could be reduced for cross-dataset portability by selecting

only those that can be calculated *de novo* in other cohorts and (2) transcriptomic classifiers could be derived from ICS subtypes and mapped onto external datasets.

Arguably more crucial than validation in standard datasets, the same classification strategies described above could also facilitate external validation in pan-cancer immunotherapy-treated cohorts, a strategy employed by nearly half the studies referenced in Table 3.7. This validation could test whether certain immune-related subtypes predict response to immunotherapy. In the absence of immunotherapy-response data, many studies in Table 3.7 including my own have inferred immunotherapy sensitivity using proxy gene signatures such as TIDE (Jiang et al., 2018), IMPRES (Auslander et al., 2018) and IPRES (Hugo et al., 2016). These are derived from transcriptional profiles of known immunotherapy responders and non-responders of several cancers. While this approach is widely adopted for estimating *likely* responders or *likely* non-responders, it's important to emphasise they are surrogates and not actual measures of clinical response. To address this limitation, several studies take the inverse approach by mapping immune-related subtypes onto real-world immunotherapy-treated cohorts and then assessing clinical outcomes such as response and survival. Although such datasets are often pan-cancer rather than colon-specific due to the relatively recent approval of immunotherapy for CRC, they offer a powerful route to validate immune subtyping frameworks particularly if they include RNAseq and/or WES/WGS data. For example, Li et al. derived a transcriptomic-based immunogenicity score from their immune subtypes and successfully applied it to an immunotherapy cohort to predict response (Li et al., 2025). Furthermore, an especially valuable extension of my work would involve mapping the ICS subtypes derived from TCGA-COAD onto immunotherapy-treated colon cancer datasets. This would allow direct testing of whether ICS groups (particularly MSS immune-high such as ICS3) are associated with differential response rates, validating their potential as clinically actionable subtypes. The recent proliferation of immunotherapy-combination clinical trials in nHM colon cancer (Table 1.2) such as NICHE (Chalabi et al., 2020), IMblaze370 (Eng et al., 2019) and

AtezoTRIBE (Antoniotti et al., 2022), which report both clinical response and RNAseq/WES data, provides a unique opportunity. For example, IMblaze370 (Taraborrelli et al., 2023) and AtezoTRIBE (Acha-Sagredo et al., 2025, Bortolomeazzi et al., 2021) datasets have already enabled immunogenicity biomarker-driven publications respectively. However, access to IMblaze370 for my own independent analysis was denied due to an ongoing Genentech-CRUK dispute. This would have permitted validation of TCGA-COAD findings in an independent, treatment-relevant cohort. Other available datasets are more limited in scope, for example some are restricted to HM patients with no available RNAseq data and thus don't allow exploration of immunogenicity in the MSS setting (Le et al., 2017, Le et al., 2015). A final consideration is that many immunotherapy trials, especially those in early development, select patients based on pre-existing markers of immunogenicity such as MSI status, PD-L1 positivity and TMB. Consequently, these cohorts may be enriched for more immune-active tumours, limiting their generalisability to "immune-cold" or MSS tumours which are precisely the groups where subtype-based insights like ICS3 vs ICS5 may have the greatest clinical utility.

In summary, while external validation is currently a limitation, several practical and established bioinformatics strategies can address this. These are essential to confirm the biological and clinical relevance of the identified subtypes, support the translational potential of ICS classification in colon cancer, support the enrichment of *ZFHX4* mutations in nHM high immunogenicity groups (whether Imm^{high} or ICS3) and correlate with immunotherapy response.

3.5.4 The association of *ZFHX4* mutations with enhanced immunogenicity

ZFHX4 consistently emerged as a gene of interest in multiple analyses focused on nHM immune-high subgroups. *ZFHX4* (missense) mutations were significantly enriched in TMBL Imm^{high} groups (NOM p=0.0488). However, in the ICS3 vs ICS5 comparison significance was lost despite a 2.17-fold higher mutation frequency (NOM p=0.2586), likely due to

unequal group sizes and reduced power in the second clustering approach. Some immune subtype studies in Table 3.7 reported *ZFHX4* mutations to be up to 4.31-fold enriched in high-immunogenicity groups (Wu et al., 2023b, Xu et al., 2022) but their analyses didn't adjust for TMB and these high-immunogenicity groups contained up to 50% MSI samples. A more rigorous interrogation into all these studies whilst controlling for TMB could help validate ZFHX4's importance.

Other comparisons between Imm^{high}/Imm^{low} and ICS subgroups such as mutations at the pathway level could give further insight, for example ICS5 comprises 90% patients with APC mutations and thus considering the CMS2 enrichment it would be interesting to see if the other 10% harbour alternate mutations in the Wnt pathway. Likewise comparing HM tumours in ICS1 vs ICS2 would be interesting, considering even within HM backgrounds different mutations have been associated with differential anti-tumour immune responses and response to immunotherapy (Ishino et al., 2023, Wang et al., 2023).

Most crucially, *ZFHX4* mutations were significantly enriched in TMBL pan-cancer immunotherapy responders (p=0.0261) (Miao et al., 2018) and was the only gene identified by Imm^{high} enrichment other than KRAS which displayed significance. This suggests validation of *ZFHX4* mutations in more immunotherapy datasets alongside TMB adjustment could be informative. Large pan-cancer cohorts with WES and matched response data, such as the CPI1000 and upcoming CPI3000 would provide ideal validation datasets (Litchfield et al., 2021, Thakkar et al., 2024). This is especially relevant since smaller immunotherapy studies already suggest *ZFHX4* mutations correlate with treatment response (Fu et al., 2024, Mao et al., 2023, Shen et al., 2022). Furthermore, a recent study analysing a pan-cancer metastatic immunotherapy-treated cohort investigated all possible molecular and clinical factors associating with response to immunotherapy and identified several hundred significant associations following univariate regressions and controlling for multiple testing – thus mining their results for ZFHX4 associations would be a logical next step (Usset et al., 2024).

Mechanistic predictions in how missense *ZFHX4* mutations could enhance immunogenicity are also necessitated and described in Chapter 5. Additionally, despite generating neoantigens some mutations can still affect immunogenicity by their original gene function as already demonstrated with RNF43 in TMBH patients where the authors describe the notion of paradoxical neoantigenic mutations (Ishino et al., 2023). This is supported by others who've shown mutations in specific protein domains were associated with immunogenicity by mechanisms other than neoantigen generation (Porta-Pardo and Godzik, 2016) and also that not all mutations in TMBH lead can enhance anti-tumour immune response (Wang et al., 2023). Considering most *ZFHX4* mutations are missense, additional work could compare their positional distribution and predicted functional impact between groups. Such approaches have previously been shown to be informative as distinct mutations within the same protein can confer differential immunogenicity and be associated with varying levels of immune infiltration (Porta-Pardo and Godzik, 2016). Additionally, this functional impact could be assessed, for example using cancer cell fraction (CCF) analysis. CCF adjusts variant allele frequency for copy number and purity to infer clonality, which has been linked to immunotherapy response, particularly through neoantigen generation (McGranahan et al., 2016). If *ZFHX4* mutations enriched in Imm^{high} or ICS3 prove clonal by CCF, this would strongly support a role for ZFHX4 in modulating the immune landscape. Furthermore, no *ZFHX4* mutations were predicted to be antigenic (presented to cell surface) or immunogenic (recognised on cell surface) in TSNAdbv2.0 for TCGA-COAD (Wu et al., 2023a) compared to many other genes described in this chapter, suggesting that their potential immunogenic relevance is independent of direct antigenicity. This supports a more nuanced view of immunogenicity beyond TMB or antigenicity alone.

It's important to emphasise that *ZFHX4* is not considered a traditional cancer driver gene (CDG) – a gene that when perturbed affects essential cellular functions prompting tumourigenesis (Hanahan and Weinberg, 2000). This is because its somatic mutational patterns across cancer types isn't indicative of positive selection (Martinez-Jimenez et al.,

2020, Gonzalez-Perez et al., 2013) and there is lack of sufficient functional and mechanistic evidence linking ZFHX4 mutations with tumourigenesis (Sondka et al., 2018). However, this does not preclude ZFHX4's potential relevance to tumour immunogenicity. Unlike classical CDG mutations, mutations associated with immune modulation may arise under alternative selective pressures or even as passengers that still influence tumour-immune interactions. For example, *ZFHX4* mutations may alter gene function in ways impacting immune infiltration or stromal signalling, even if they are not under strong positive selection for driving tumour growth. Few studies have sought to identify somatic mutational drivers of immunogenicity, and those that exist focus mainly on CDGs (Misetic et al., 2023) or neoantigens (Hoyos et al., 2022, Zapata et al., 2023). Further work is required, since a gene does not need to be a canonical driver to influence immunogenicity, and immunogenicity extends well beyond antigenicity alone. These concepts have been elaborated further in the main discussion (Chapter 6).

3.5.5 Conclusion

In summary, *ZFHX4* displays consistent mutational enrichment in immunogenic TMBL subgroups and together with published associations with immunotherapy response across cancer types (Fu et al., 2024, Mao et al., 2023, Shen et al., 2022) strongly implicate it as a potential regulator of immunogenicity. Interestingly, brief analyses using cBioportal showed *ZFHX4* expression to be significantly prognostic in TCGA-COAD (Cerami et al., 2012), consistent with reports in ovarian cancer (Ha et al., 2020, Millstein et al., 2020, Zong et al., 2022). Together with one *in vitro* study exhibiting oncogenic epigenetic regulation in glioblastoma (Chudnovsky et al., 2014), these findings and the gaps in colon cancer research highlight ZFHX4 as a candidate for further investigation in both tumourigenicity and immunogenicity contexts with regard to its expression and mutations respectively. Future analyses incorporating functional modelling as explored in Chapters 4 and 5 will be essential to elucidate a mechanistic role to explain these apparent associations.

Chapter Four: ZFHX4 as a potential regulator of colon tumourigenicity

4.1 Introduction

Zinc-Finger Homeobox 4 (ZFHX4) is a poorly characterised transcription factor (TF). Therefore, this introduction provides a small literature review on what is currently known about its classification, structure and function in physiological and disease contexts.

4.1.1 ZFHX4 classification and structure

ZFHX4 is classified within the homeobox gene superclass, characterised by ~180bp homeobox DNA motifs that encode structurally conserved homeodomains (Holland, 2013, Bürglin, 2011, Holland et al., 2007). Homeodomains are ~60 amino acids (aa)-long protein structures that bind DNA through a triple alpha-helix core, with one helix responsible for making sequence-specific contacts in the DNA backbone (Bürglin, 2011). All homeodomain-containing proteins are thus transcription factors (TFs), defined as proteins able to bind DNA in a sequence specific manner and regulate transcription of genes such as those involved in development and differentiation of all multicellular eukaryotes (Bürglin, 2011, Lambert et al., 2018). 11 classes of homeobox genes exist and are mostly defined by other domains, for example the Zinc-Finger (ZF) class which contain 28aa-long cysteine$(C)_2$histidine$(H)_2$-type ZF structures, the most common domain in metazoan TFs which interact with nucleic acid, proteins and lipids (Stubbs et al., 2011). In humans, this class encodes many polydactyl ZFs and homeodomains and is further subclassified into 5 families of 14 genes including *ZFHX4* (Bürglin, 2011, Holland et al., 2007) (Figure 4.1A).

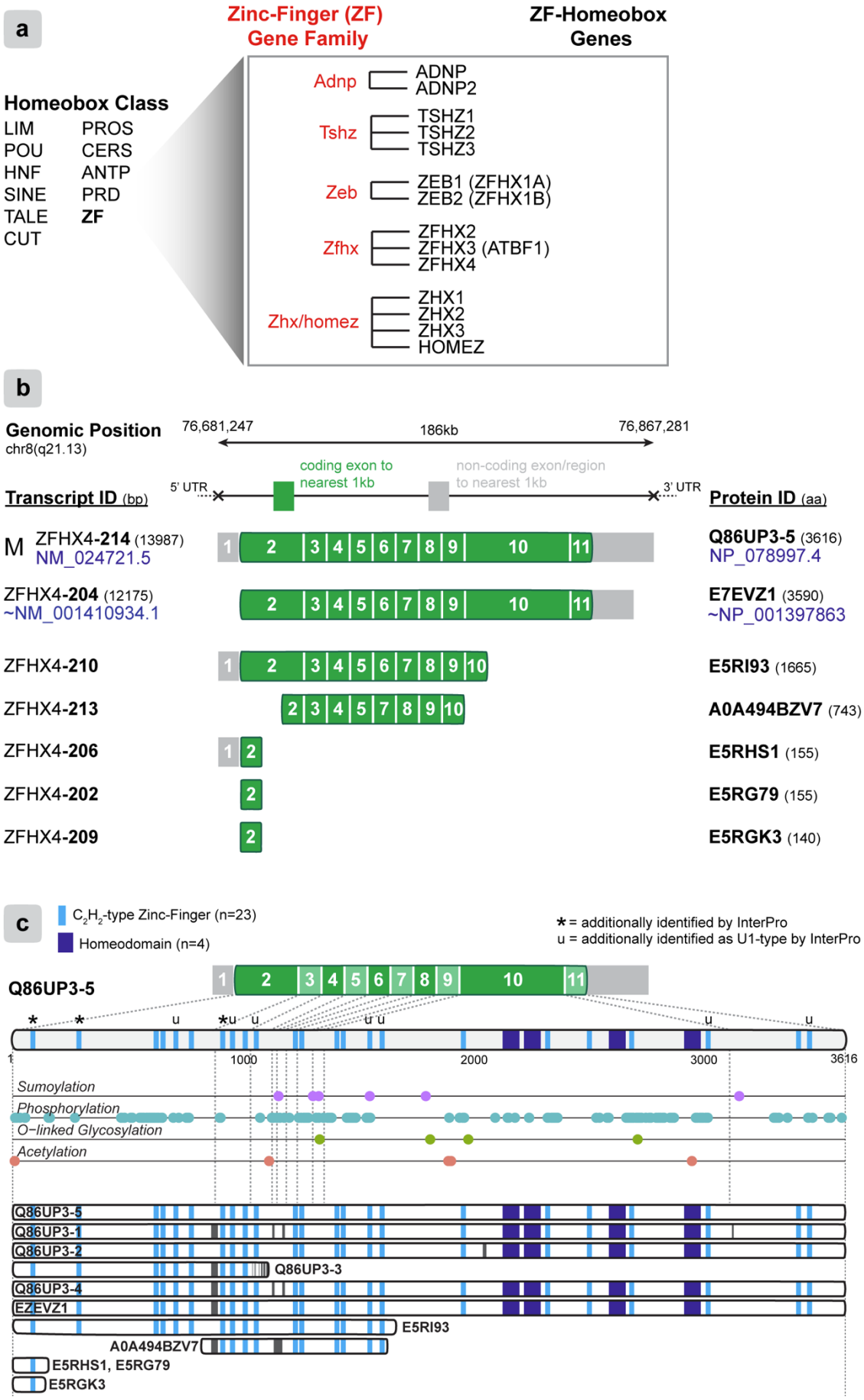

a
Zinc-Finger (ZF) Gene Family
ZF-Homeobox Genes
Homeobox Class
LIM
POU
HNF
SINE
TALE
CUT
PROS
CERS
ANTP
PRD
ZF
Adnp
ADNP
ADNP2
Tshz
TSHZ1
TSHZ2
TSHZ3
Zeb
ZEB1 (ZFHX1A)
ZEB2 (ZFHX1B)
Zfhx
ZFHX2
ZFHX3 (ATBF1)
ZFHX4
Zhx/homez
ZHX1
ZHX2
ZHX3
HOMEZ
b
Genomic Position
chr8(q21.13)
76,681,247
186kb
76,867,281
coding exon to nearest 1kb
non-coding exon/region to nearest 1kb
5' UTR
3' UTR
Transcript ID (bp)
Protein ID (aa)
M ZFHX4-214 (13987)
NM_024721.5
Q86UP3-5 (3616)
NP_078997.4
ZFHX4-204 (12175)
~NM_001410934.1
E7EVZ1 (3590)
~NP_001397863
ZFHX4-210
E5RI93 (1665)
ZFHX4-213
A0A494BZV7 (743)
ZFHX4-206
E5RHS1 (155)
ZFHX4-202
E5RG79 (155)
ZFHX4-209
E5RGK3 (140)
c
C2H2-type Zinc-Finger (n=23)
Homeodomain (n=4)
* = additionally identified by InterPro
u = additionally identified as U1-type by InterPro
Q86UP3-5
1000
2000
3000
3616
Sumoylation
Phosphorylation
O-linked Glycosylation
Acetylation
Q86UP3-5
Q86UP3-1
Q86UP3-2
Q86UP3-3
Q86UP3-4
E7EVZ1
E5RI93
A0A494BZV7
E5RHS1, E5RG79
E5RGK3

Figure 4.1 | Domain-based classification of ZFHX4 and its variation in structure at the genomic, transcriptomic and proteomic level. (A) Classification of ZFHX4 within the homeobox superclass, zinc-finger class and Zfhx gene family. Adapted from (Holland et al., 2007, Bürglin, 2011). **(B)** Seven *ZFHX4* mRNA transcript variants reported by ENSEMBL. Introns are not represented and exon schematic size is proportional to size to the nearest kb. Matching UniProt protein IDs are also reported per variant (only protein coding variants are shown). 'M' denotes MANE transcript. Dark blue IDs under the top two variants refer to matching RefSeq entries. Numbers refer to exon numbers in ENSEMBL. **(C)** Domain annotation using Q86UP3-5 as the reference protein sequence. Exon positioning to amino acids is accurate, with coordinates obtained from ENSEMBL. Light and dark green indicates alternating exons. White rectangles outlined in black depict protein sequence for all available isoforms, aligned using BLAST on UniProt. Alignment was in reference to Q86UP3-5, with dark grey gaps indicating absence of amino acid and light grey indicating alternate amino acids (only in the Q86UP3-3 isoform). PTMs were mined from dbPTMv2025 with reference to Q86UP3-5. All information obtained from (Frankish et al., 2023, Morales et al., 2022, UniProt, 2025, Blum et al., 2025, Dyer et al., 2025, Chung et al., 2025) *[aa=amino acid, bp base pairs, PTMs=post-translational modifications]*

Spanning 186kb, *ZFHX4* is located on chromosome 8 (q21.13) and comprises 11 exons, 9 classed as coding exons and 1 (exon 11) a mix of non-coding and coding regions (Figure 4.1B). EMBL-EBI-ENSEMBL (release 114, GRCh38.p14) presents 14 mRNA variants by alternative splicing, 7 of which predicted as protein coding and match with a protein isoform accession in the UniProtKB database (Figure 4.1B) (UniProt, 2025, Dyer et al., 2025). The *ZFHX4-214* mRNA variant is classed as the MANE transcript (a match to NCBI's RefSeq transcript variant 1) and will thus be utilised as reference throughout the project, demonstrating 13987 base-pair (bp) exon size and 10851bp coding sequence (Morales et al., 2022).

The matching protein isoform Q86UP3 is the only Swiss-Prot reviewed isoform of ZFHX4 and is presented in 5 variations in UniProtKB. Despite Q86UP3-1 annotated here as the 'canonical' isoform, Q86UP3-5 of 3616aa is the matching MANE isoform and will also be used as reference throughout the project (Figure 4.1C). All other unreviewed protein isoforms are predicted based on computational mapping from ENSEMBL mRNA transcript variants and thus may not exist biologically (UniProt, 2025). With additional domain annotation by InterPro that collates SMART, PFAM, PROFILE, PROSITE and CDD database annotation, ZFHX4 protein has 23 ZFs (7 matrin/U1-type and 13 C_2H_2-type) and 4 homeodomains with the latter entirely coded by exon 10, the largest *ZFHX4* exon of 5415bp (Blum et al., 2025). The second largest exon (2) codes for amino acids present in

nearly all predicted ZFHX4 protein isoforms. Around half (5/11) of protein isoforms contain all four homeodomains at the C-terminus end whilst the majority (8/11) contain most N-terminus ZFs. Whilst not depicted in Figure 4.1, 92% of *ZFHX4's* gene length is represented by introns, ranging up to 70.2kb between exon 3 and 4 (Dyer et al., 2025, Frankish et al., 2023). Furthermore, it's long aa structure results in a large 399kDa protein displaying many predicted post-translational modifications (PTMs) including SUMOylation, phosphorylation, O-linked glycosylation and acetylation (dbPTMv2025 (Chung et al., 2025)). 92% PTMs are phosphorylation, which mostly occur on aa's coded from exon 2 and 10 due to their large sizes, with the latter exon mostly coding aa's hosting all other PTM types.

4.1.2 ZFHX4 in normal physiology

ZFHX4 was first identified in mouse, showing significant homology to the previously characterised *ZFHX3* and exhibiting low but detectable mRNA in most embryonic tissues (Kostich and Sanes, 1995, Sakata et al., 2000). Protein expression was first demonstrated in rat central nervous tissue (Nogami et al., 2005). Characterisation in humans demonstrated genomic mapping to chr8q, with protein displaying binding sites for SRY, CdxA, MZF1, S8 and deltaEF1 TFs (Hemmi et al., 2006). Alongside an interesting similarity found between mouse *Zfhx4* and human *ZFHX3*, potential roles of ZFHX4 in neuronal and muscle differentiation were suggested (Hemmi et al., 2006).

The only known study to examine human ZFHX4 in physiological conditions used induced pluripotent stem cells (iPSCs) differentiated into neuronal stem cells, neuronal progenitor cells (NPCs) and neurons, and demonstrated *ZFHX4* mRNA expression increased during human neuronal differentiation (Perez Baca et al., 2025). Immunoprecipitation-Mass Spectrometry (IP-MS) of NPCs and neurons to identify potential ZFHX4-protein interactions exhibited 47 significantly enriched proteins relative to an IgG isotype control (including its family member ZEB1) involved in histone modifications, embryonic development, transcriptional regulation, protein trafficking and neurodevelopmental disorders. To identify

potential ZFHX4-DNA interactions, CUT&RUN was performed on NPCs, retrieving 1640 peaks of 55.62% promoter regions concerning genes involved in neuronal development, mesenchymal differentiation and Wnt signalling. Furthermore, genes were significantly enriched for homeobox, SOX-related and ZF motifs which occur in transcription factors crucial for transcriptional regulation and proliferation. Using CRISPR to disrupt *Zfhx4* in zebrafish (72% aa similarity) investigators demonstrated its requirement for craniofacial and skeletal development. This aligns with two other studies showing *Zfhx4* expression across most cell types during embryogenesis, with temporospatial patterns similar to mice (Ishorst et al., 2025, Liu et al., 2025).

Additional work in mice has uncovered roles in neural development and differentiation through *Phox2b*, *Math1* and *Lbx1* regulation (Zhang et al., 2021a) and endochondral ossification and ECM-related processes alongside *Osterix*, *Runx2* and *Sox9,* regulating genes such as *Mmp13* (Nakamura et al., 2021). One study identified *Zfhx4* as a target of the TF Zfp260 which is crucial for maturation of skeletal stem cells into progenitors and directing osteogenic differentiation through maintaining chromatin accessibility within a super-enhancer complex with *Med1*, *Brd4* and *P300* (Weng et al., 2024). Whilst identification of a super-enhancer region of *Zfhx4* was described but not further elaborated on, its ATAC-signal accessibility was reduced following Zfp260 knockout along with signals pertaining to genes associated with ECM organisation, development and differentiation as well as the *Runx2* super-enhancer region. This complex was found to interact with ß-catenin, Smad5 and Gli1 protein and regulate genes involved in cell adhesion, development, ECM and Wnt signalling as identified by ChIP-seq. Loss of Zfp260 impaired osteogenesis and reduced skeletal stem cell potency, underscoring the importance of maintaining *Zfhx4* accessibility during lineage progression. Notably, *Zfhx4* expression increased during osteogenic fate commitment, suggesting a role in late-stage differentiation. These findings position *Zfhx4* as a chromatin-regulated gene within a key osteogenic super-enhancer network, with Zfp260 as a primary upstream regulator in mice.

4.1.3 ZFHX4 in disease and disorders

Most ZFHX4 research pertains to human disease and developmental disorders associated with genomic changes to *ZFHX4* and surrounding regions. Since an initial case report of congenital bilateral isolated ptosis associated with 8q21 breakpoints and consequent *ZFHX4* disruption (McMullan et al., 2002), an 8q21.1 microdeletion syndrome (8q21.1-MDS) has now been identified. These individuals display intellectual disability, distinct facial features and eye abnormalities, with alterations ranging from 0.66-14.5Mb (Fontana et al., 2021, Palomares et al., 2011). *ZFHX4* resides in this critical region and all individuals with germline inherited or *de novo ZFHX4* alterations present with 8q21.1-MDS features, whilst some 8q21.1-MDS individuals can have many genes affected such as neighbouring *PEX2* (Fontana et al., 2021). Postulation surrounding this phenotype suggests it could result from haploinsufficiency for a single gene and that *ZFHX4* could be the candidate disease-causing gene. For example, heterozygous loss by intragenic deletion of *ZFHX4* was found in an individual presenting with 8q21.1-MDS suggesting haploinsufficiency and the underlying cause (Vulto-van Silfhout et al., 2013). Furthermore, heterozygous germline *de novo* point mutations have also been reported, for example in splice sites (c.3093+1G>T and c.3646-1G>A) predicted to be pathogenic and loss of function (LoF) (Eising et al., 2019, Fontana et al., 2021). Moreover, two large-scale exome sequencing studies identified *ZFHX4* as a key neurodevelopmental disorder gene, with LoF and missense *de novo* variants significantly enriched in individuals affected by orofacial clefts (Kaplanis et al., 2020, Bishop et al., 2020), and more studies continue to report further neurodevelopmental-related *ZFHX4* variants (Creton et al., 2023, Sorrentino et al., 2024, Goel and O'Donnell, 2025, Ishorst et al., 2025).

In agreement with this, the ClinVar database of disorder-associated variants presents 8 *ZFHX4* single nucleotide variants (SNVs; <50bp) and 32 structural variants (SVs; ≥50bp) predicted to be likely pathogenic or pathogenic by LoF (Landrum et al., 2014). There was an additional 422 SNVs with uncertain significance, with 90% classified as missense

mutations. Collating 76156 human genomes gnomAD (v4.1.0) displays 8218 germline variants affecting *ZFHX4*, with 60 SNVs and 112 SVs predicted LoF by frameshift, stop-gained or splice-site-affecting (Chen et al., 2024). Based on the MANE *ZFHX4* transcript gnomAD also calculates constraint metrics predicting a gene's tolerability to genetic variation based upon observed SNVs in the population vs expected SNVs by mutational modelling (o/e), considering sequence context and mutation rates. For synonymous *ZFHX4* SNVs, an o/e score of 1.02 is consistent with expectation for neutral variation and an o/e score of 0.94 for non-synonymous missense *ZFHX4* SNVs suggest slightly less tolerability than synonymous SNVs. Interestingly, the o/e score for predicted LoF-SNVs in *ZFHX4* (0.19) and probability of LoF intolerance value (1.0) reveals that *ZFHX4* is likely highly LoF-intolerant, dosage-sensitive, strongly selected against and consistent with haploinsufficiency at the germline level.

Further work on germline *ZFHX4* variants has been undertaken by the Vergult group and ZFHX4 Consortia who studied a novel patient group harbouring various *ZFHX4* aberrations and displaying neurodevelopmental disorder phenotypes (Perez Baca et al., 2025). Collating 57 individuals with germline aberrations affecting *ZFHX4* of mostly *de novo* origin, 21 displayed LoF SVs ranging from 1.8kb to 12.92Mb and mostly consisting of multigenic and intragenic deletions (Perez Baca et al., 2025). The other 36 displayed LoF SNVs consisting of heterozygous frameshift, nonsense and splice-site variants. Furthermore, 20 missense SNVs in gnomAD affect codons also mutated in affected individuals in this cohort, indicating some *ZFHX4* missense mutations could contribute to disease despite missense constraint metrics suggesting only mild selection at the population level. This real-world phenotypic validation further supports haploinsufficiency, demonstrates heterozygous LoF (SNVs and SVs) in *ZFHX4* could cause disease and hints that some missense SNVs could have deleterious functional impact.

4.1.4 ZFHX4 in cancer

Pertaining to its role in cancer biology, a mix of *in vitro* and *in silico* approaches are summarised in Table 4.1. Most publications suggest increased *ZFHX4* mRNA expression promotes tumourigenesis and is associated with poor survival irrespective of cancer type. To my knowledge, there is no publication on ZFHX4 in the context of healthy colon or colon cancer.

Table 4.1 | Publications reporting on ZFHX4 in cancer and tumourigenicity through *in vitro* (IV) or *in silico* (IS) means

IV/IS	Study	Cancer Type	Findings
IV 2014	Chudnovsky[1]	GBM	Described below
IV 2017	Qing[2]	OAC	• *ZFHX4* KD by siRNA in KYSE150 and TE-1 cell lines demonstrated decreased *ZFHX4* mRNA expression • Transwell and scratch in vitro migration assays showed reduced migration in *ZFHX4* KD (no significant effects on proliferation or apoptosis)
IV 2019	Xia[3]	LUAD	• A competing endogenous RNA (ceRNA) network was contructed from mRNA-miRNA-mRNA co-expression relationships in TCGA of *TWIST1-miR-514a-3p-ZFHX4* as well as *TWIST1-miR-194-3p-SLC12A5* • High expression of *TWIST1/ZFHX4/SLC12A5* correlated with late-stage and poor overall survival • KD of *TWIST1/ZFHX4/SLC12A5* reduced proliferation, migration and invasion in H1299 cells (high *TWIST1*), and overexpression of 3'UTRs increased proliferation, migration and invasion in A549 cells (low *TWIST1*). These effects were eliminated upon respective miRNA inhibitors • BALB/c mice injected with H1299-shRNA displayed similar tumour volumes to controls but growth was reduced by addition of respective miRNA inhibitors • Overexpression of *SLC12A5/ZFHX4* 3'UTRs both upregulated *TWIST1* mRNA and protein expression in A549s, and vice versa with shRNA KD in H1299s with effects eliminated upon respective miRNA inhibitors • AGO2 and miRNA pulldowns confirmed *SLC12A5* and *ZFHX4* compete with *TWIST1* for the binding of miRNAs by their 3' UTRs
IS 2006	Fevre-Montange[4]	Pineal	• High mRNA expression selectively in 'papillary tumour of the pineal region' comparative to normal brain and other pineal tumours

IS 2015	Stacey[5]	Basal cell	• GWAS identified an Indel variant SNP (rs28727938[G]) identified 100kb proximal to *ZFHX4* associated with BCC risk, suggesting potential susceptibility loci to BCC • Demonstrated limited evidence for a functional role in BCC pathogenesis
IS 2017	Lu[6]		• WES of BCC patients revealed ZFHX4 damaging mutations • STRING network analysis predicted associations with *CHD4*, *ZEB1*, *PIAS1*, *TMEM14A*, *HIVEP3* and *NOTCH* genes, though these aren't fully consistent with online
IS 2017	Qing[2]	OAC	• *ZFHX4* mutations significantly correlated with poor overall survival (n=281) • *ZFHX4* expression is higher in tumour than normal
IS 2020	Ha[7]	Ovarian	• *ZFHX4* mRNA expression was significantly higher in late-stage ovarian cancer than early-stage in 2 separate cohorts • High *ZFHX4* mRNA expression was significantly associated with poor prognosis in 5 independent cohorts irrespective of stage and grade
IS 2020	Millstein[8]		• *ZFHX4* mRNA expression was significantly associated with poor overall survival alongside *TAP1*, *CXCL9*, *FBN1* and *PTGER3*, conferring a 20% risk of death (n=3769) • A 101 gene signature including *ZFHX4* enriched for immune-related pathways identified patients least likely to achieve 5-year survival in a separate cohort
IS 2022	Zong[9]		• DEG analysis identified high *ZFHX4* expression in metastatic samples comparative to matched primary tumours in two cohorts, and its expression conferred an increased 86% risk in progression • *ZFHX4* was significantly positively correlated with EMT markers, ion trans-membrane channel genes, collagen formation, ECM-related genes and TGFß gene sets suggesting a potential role in metastasis • *ZFHX4* was significantly negatively correlated with DNA replication and proliferation, suggesting a reduction in proliferative phenotype whilst boosting metastatic potential

[1](Chudnovsky et al., 2014), [2](Qing et al., 2017), [3](Xia et al., 2019), [4](Fevre-Montange et al., 2006), [5](Stacey et al., 2015), [6](Lu et al., 2017), [7](Ha et al., 2020), [8](Millstein et al., 2020), [9](Zong et al., 2022). *[KD=knockdown, IV=in vitro, IS=in silico, OAC=oesophageal carcinoma, GBM=glioblastoma, LUAD=lung adenocarcinoma]*

The most notable study exploring ZFHX4 in tumourigenesis identified it as a key regulator of tumour-initiating cell (TIC) state in Glioblastoma (GBM) (Chudnovsky et al., 2014). TICs are classed as stem-like multipotent cells representing a therapy-resistant population responsible for recurrence and poor prognosis. *ZFHX4* suppression by RNA interference (RNAi) prompted profound differentiation, reduced proliferation and G0/G1 arrest but not cell death across five patient-derived TIC lines, downregulating stem-cell marker proteins

(SOX2, NESTIN, NG2, SSEA-1, ITGA6, EGFR) and upregulating neuronal proteins (DCX, p35). Injection of *ZHFX4* shRNA-suppressed TIC cells into NOD-SCID mice significantly prolonged median cancer-free survival respective to controls and this survival was proportional to shRNA-mediated expression suppressing efficiency, suggesting *ZFHX4* was indispensable for TIC-driven tumourigenesis. ZFHX4-protein interaction investigation by immunoprecipitation-mass spectrometry (IP-MS) of FLAG-tagged ZFHX4 in HEK239T cells demonstrated 15 hits including ZFHX4 and CHD4, a catalytic subunit of the regulatory epigenetic nucleosome remodelling and deacetylase (NuRD) complex (Lai and Wade, 2011). The NuRD complex can activate and repress transcription of target genes through multiple context-dependent mechanisms such as nucleosome remodelling and has been associated with stem cell fate, differentiation, tumourigenic pathways such as EMT and maintenance of genomic stability (Lai and Wade, 2011, Whyte et al., 2012, Ramirez and Hagman, 2009, Fu et al., 2011). ChIPseq identified 29% ZFHX4-bound DNA regions overlapped with those bound by CHD4, with target genes spanning known cancer drivers such as *MYC*, *PDGFRB* and *ETS1* as well as tumour suppressors such as *SPRY1*. 30.6% of these target genes had upregulated expression and 34.2% had downregulated expression upon ZFHX4/CHD4 shRNA-mediated suppression supporting the pleiotropic regulatory nature of NuRD in activating and repressing numerous specific genes. These findings were consistent with ZFHX4 acting as a TF and modulating NuRD-mediated gene expression potentially through genomic localisation of the complex to specific genes imperative for stem cell-like functions and tumourigenicity.

4.2 Hypothesis

ZFHX4 expression can attribute to aggressive tumorigenicity in colon cancer via transcriptional regulation, proliferation, migration and invasion and related oncogenic mechanisms.

Aims

1. Characterise ZFHX4 expression in normal and cancerous intestinal tissue
2. Generate and validate colon-specific *ZFHX4*$^{-/-}$ models using CRISPR/Cas9 technology
3. Perform multiomic profiling on these models to infer potential mechanisms of tumourigenicity
4. Perform functional profiling on these models to ascertain whether *ZFHX4* perturbations alter mechanisms of tumourigenicity
5. Validate results from transcriptomic experiments with the ZFHX4 phenotype observed clinically

4.3 Experimental design

Published datasets spanning single-cell RNA sequencing (scRNAseq), bulk-RNAseq and survival data were mined to provide additional inferences of ZFHX4 biology due to the current scarcity of work regarding known function. CRISPR/Cas9-mediated *ZFHX4*$^{-/-}$ in metastatic colon cancer organoids and cell lines was validated at the gDNA level with sanger sequencing and further profiled by bulk RNAseq. Cell lines were subsequently focused on due to organoid culture limitations, investigating potential ZFHX4 transcriptional target genes through CUT&RUN and functional profiling spanning proliferation, cell-cycle, secretome, angiogenesis and migration and invasion. Finally, an integrative analysis of four colon cancer multiomic datasets with my *in vitro* work was performed to observe any parallels between my models and the clinical ZFHX4 phenotype.

4.4 Results

4.4.1 ZFHX4 expression in normal and cancer intestinal tissues

4.4.1.1 ZFHX4 gene expression in normal tissue

Because little is reported in the literature regarding *ZFHX4* expression, bioinformatic mining was employed to draw temporary associations. According to the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2013), *ZFHX4* mRNA is expressed in nearly all healthy adult human tissues and this is largely with respect to its most characterised 'MANE' isoform ZFHX4-214 (Figure 4.1, (GTEx Consortium, 2025)). *ZFHX4* mRNA expression transcript per million (TPM) values (median 6.605) are around 1000-fold less than housekeeper genes (median 6778 *ACTB* in Colon-Sigmoid) suggesting generally low *ZFHX4* expression in bulk-RNA sequencing (RNAseq) (GTEx Consortium, 2025).

To investigate whether *ZFHX4* expression pertains to a particular cell type within the colon, relevant pre-processed single-cell RNAseq (scRNAseq) studies were mined from the cellxgene portal (C.Z.I.CellScienceProgram et al., 2025). Six studies were selected based upon utilisation of healthy intestinal tissue from a range of developmental ages (Table 4.2). In general, most cells (~99%) in any dataset didn't express *ZFHX4*. In line with previous evidence suggesting ZFHX4's role in foetal development, ~15-fold more cells of foetal origin expressed *ZFHX4* comparative to cells of adult origin.

Table 4.2 | Summary of scRNAseq datasets

Dataset	Total cells by developmental age (N)	Total cells expressing ZFHX4, N (%)
Space-Time Gut Atlas (Teichmann-A)[1]	Healthy foetal (247 344)	40 521 (16.38)
	Healthy paediatric (29 594)	26 (0.09)
	Healthy adult (124 367)	448 (0.36)
PanGI Cell Atlas (Teichmann-B)[2]	Healthy foetal (333 624)	49 954 (14.97)
	Healthy preterm (9512)	79 (0.83)
	Healthy adult (942 394)	4128 (0.44)
HTAN Cell Atlas[3]	Set 1 Healthy adult (30 340)	284 (0.94)

	Set 2 Healthy adult (34 008)	263 (0.77)
Human Cell Atlas - Intestine (Teichmann-C)[4]	Healthy adult (263 275)	419 (0.16)
Tabula Sapiens Human Cell Atlas - Intestine[5]	Healthy adult (80 103)	787 (0.98)
Human Cell Landscape - Intestine[6]	Healthy foetal (23 516)	82 (0.35)
	Healthy adult (35 898)	18 (0.05)

[1](Elmentaite et al., 2021), [2](Oliver et al., 2024), [3](Chen et al., 2021), [4](Xu et al., 2023), [5](Tabula Sapiens et al., 2022), [6](Han et al., 2020)

Focusing generally on *ZFHX4*-expressing cells (non-zero log-normalised counts per 10k reads), ZFHX4 expression in healthy intestine was predominantly restricted to mesenchymal and neuronal cells and some epithelium and endothelium subsets (Supplementary Figure S4.2). Specific cell types expressing *ZFHX4* in each dataset were then summarised (Figure 4.2). There was a noticeable change from stromal and neural predominance in foetal datasets to epithelial subsets in adult datasets. For example, small proportions of transit-amplifying, crypt-base stem cells, enterocytes, goblet cells and tuft cells were found to express *ZFHX4* in nearly all six datasets in the adult intestine. Out of all *ZFHX4*-expressing cells, the largest proportions represented fibroblast, glial and lymphatic endothelial cell populations which followed similar trends in every dataset. Furthermore, *ZFHX4* was expressed in normal colon-derived fibroblasts in an in-house scRNAseq dataset (Yi Zhou, data not shown). Interestingly, expression per cell was higher relative to housekeeping genes than in bulk-RNAseq. For example, in the Space-Time Gut Atlas, *ZFHX4* reached 17 counts/10k vs *ACTB* 830 counts/10k, highlighting a 50-fold difference compared with the 1000-fold difference observed in bulk-RNAseq. This contrast indicates dilution in bulk profiles by many non-expressing cells and underscores scRNAseq for studying ZFHX4 in specific cell types.

Collectively, these data suggest *ZFHX4* expression is generally rare in the colon but most frequently observed in stromal and neural populations. Within epithelium, *ZFHX4* expression is less common but detectable at moderate levels across several lineages, indicating a broad but low-frequency cell-type-specific expression pattern.

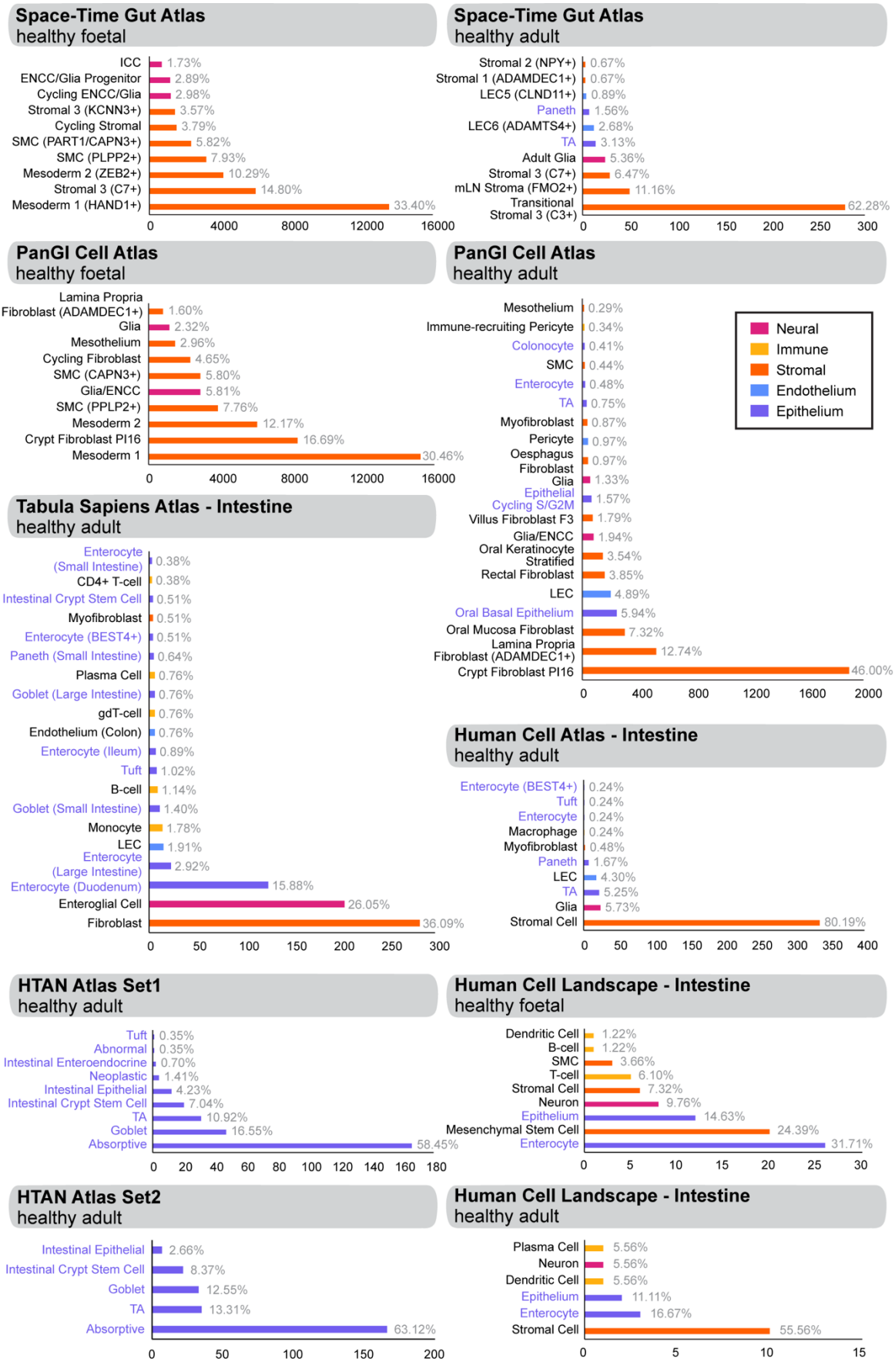

Space-Time Gut Atlas
healthy foetal
ICC 1.73%
ENCC/Glia Progenitor 2.89%
Cycling ENCC/Glia 2.98%
Stromal 3 (KCNN3+) 3.57%
Cycling Stromal 3.79%
SMC (PART1/CAPN3+) 5.82%
SMC (PLPP2+) 7.93%
Mesoderm 2 (ZEB2+) 10.29%
Stromal 3 (C7+) 14.80%
Mesoderm 1 (HAND1+) 33.40%
0 4000 8000 12000 16000
Space-Time Gut Atlas
healthy adult
Stromal 2 (NPY+) 0.67%
Stromal 1 (ADAMDEC1+) 0.67%
LEC5 (CLND11+) 0.89%
Paneth 1.56%
LEC6 (ADAMTS4+) 2.68%
TA 3.13%
Adult Glia 5.36%
Stromal 3 (C7+) 6.47%
mLN Stroma (FMO2+) 11.16%
Transitional Stromal 3 (C3+) 62.28%
0 50 100 150 200 250 300
PanGI Cell Atlas
healthy foetal
Lamina Propria Fibroblast (ADAMDEC1+) 1.60%
Glia 2.32%
Mesothelium 2.96%
Cycling Fibroblast 4.65%
SMC (CAPN3+) 5.80%
Glia/ENCC 5.81%
SMC (PPLP2+) 7.76%
Mesoderm 2 12.17%
Crypt Fibroblast PI16 16.69%
Mesoderm 1 30.46%
0 4000 8000 12000 16000
PanGI Cell Atlas
healthy adult
Mesothelium 0.29%
Immune-recruiting Pericyte 0.34%
Colonocyte 0.41%
SMC 0.44%
Enterocyte 0.48%
TA 0.75%
Myofibroblast 0.87%
Pericyte 0.97%
Oesphagus Fibroblast 0.97%
Glia 1.33%
Epithelial Cycling S/G2M 1.57%
Villus Fibroblast F3 1.79%
Glia/ENCC 1.94%
Oral Keratinocyte Stratified 3.54%
Rectal Fibroblast 3.85%
LEC 4.89%
Oral Basal Epithelium 5.94%
Oral Mucosa Fibroblast 7.32%
Lamina Propria Fibroblast (ADAMDEC1+) 12.74%
Crypt Fibroblast PI16 46.00%
0 400 800 1200 1600 2000
Neural
Immune
Stromal
Endothelium
Epithelium
Tabula Sapiens Atlas - Intestine
healthy adult
Enterocyte (Small Intestine) 0.38%
CD4+ T-cell 0.38%
Intestinal Crypt Stem Cell 0.51%
Myofibroblast 0.51%
Enterocyte (BEST4+) 0.51%
Paneth (Small Intestine) 0.64%
Plasma Cell 0.76%
Goblet (Large Intestine) 0.76%
gdT-cell 0.76%
Endothelium (Colon) 0.76%
Enterocyte (Ileum) 0.89%
Tuft 1.02%
B-cell 1.14%
Goblet (Small Intestine) 1.40%
Monocyte 1.78%
LEC 1.91%
Enterocyte (Large Intestine) 2.92%
Enterocyte (Duodenum) 15.88%
Enteroglial Cell 26.05%
Fibroblast 36.09%
0 50 100 150 200 250 300
Human Cell Atlas - Intestine
healthy adult
Enterocyte (BEST4+) 0.24%
Tuft 0.24%
Enterocyte 0.24%
Macrophage 0.24%
Myofibroblast 0.48%
Paneth 1.67%
LEC 4.30%
TA 5.25%
Glia 5.73%
Stromal Cell 80.19%
0 50 100 150 200 250 300 350 400
HTAN Atlas Set1
healthy adult
Tuft 0.35%
Abnormal 0.35%
Intestinal Enteroendocrine 0.70%
Neoplastic 1.41%
Intestinal Epithelial 4.23%
Intestinal Crypt Stem Cell 7.04%
TA 10.92%
Goblet 16.55%
Absorptive 58.45%
0 20 40 60 80 100 120 140 160 180
Human Cell Landscape - Intestine
healthy foetal
Dendritic Cell 1.22%
B-cell 1.22%
SMC 3.66%
T-cell 6.10%
Stromal Cell 7.32%
Neuron 9.76%
Epithelium 14.63%
Mesenchymal Stem Cell 24.39%
Enterocyte 31.71%
0 5 10 15 20 25 30
HTAN Atlas Set2
healthy adult
Intestinal Epithelial 2.66%
Intestinal Crypt Stem Cell 8.37%
Goblet 12.55%
TA 13.31%
Absorptive 63.12%
0 50 100 150 200
Human Cell Landscape - Intestine
healthy adult
Plasma Cell 5.56%
Neuron 5.56%
Dendritic Cell 5.56%
Epithelium 11.11%
Enterocyte 16.67%
Stromal Cell 55.56%
0 5 10 15

Figure 4.2 | Numbers and proportions of ZFHX4-expressing cell types in healthy intestine scRNAseq datasets. Each grey heading depicts a separate dataset or developmental stage within one dataset. X axes represent distinct cell numbers and percentages highlighted in grey refer to the proportion of that ZFHX4-expressing cell type out of all ZFHX4-expressing cell types in that dataset. Horizontal bar colours are coded by the key which roughly segregates all different cell types into five main groups. Specific focus is placed upon epithelial-based cell types, with text additionally highlighted in dark blue. Not all cell types are shown, just the top select few for visual representation. *[ICC = interstitial cells of cajal, LEC = lymphatic endothelial cell, SMC = smooth muscle cell, TA = transit-amplifying cell, ENCC = enteric neural crest cell]*

4.4.1.2 ZFHX4 gene expression in cancer and colon cancer tissue

ZFHX4 mRNA is expressed in multiple malignancies, with most high expressing cancers conferring neuronal-related origin (Figure 4.3A, (Hoadley et al., 2018)). Whilst colorectal cancer isn't the highest expressing by median, it presents tumours with a wide range of expression values (colon median 4.177 [0.235-8.842], rectal median 4.046 [0.235-8.648], RSEM log2(x+1)) and this diversity is apparent in multiple tumour types. Furthermore, annotating copy number status encompassing the *ZFHX4* region, whilst 54.2% cancers remain diploid, most changes appear to be amplifications and gains (39.8%) which are 6.7-fold more frequent comparative to shallow and deep deletions (5.97%). Most importantly, *ZFHX4* expression is highly prognostic irrespective of cancer type (Figure 4.3B). When patients are stratified by the 25th and 75th percentiles of *ZFHX4* expression, those with high expression consistently exhibit poorer outcomes compared to those with low expression across all survival endpoints (overall survival;OS, progression-free survival;PFS, and disease-specific survival;DSS). Specifically, low *ZFHX4*-expressing tumours are associated with a significant 32% reduction in overall mortality, significant 39% reduction in risk of disease progression and significant 50% reduction in disease-specific mortality, independent of cancer type.

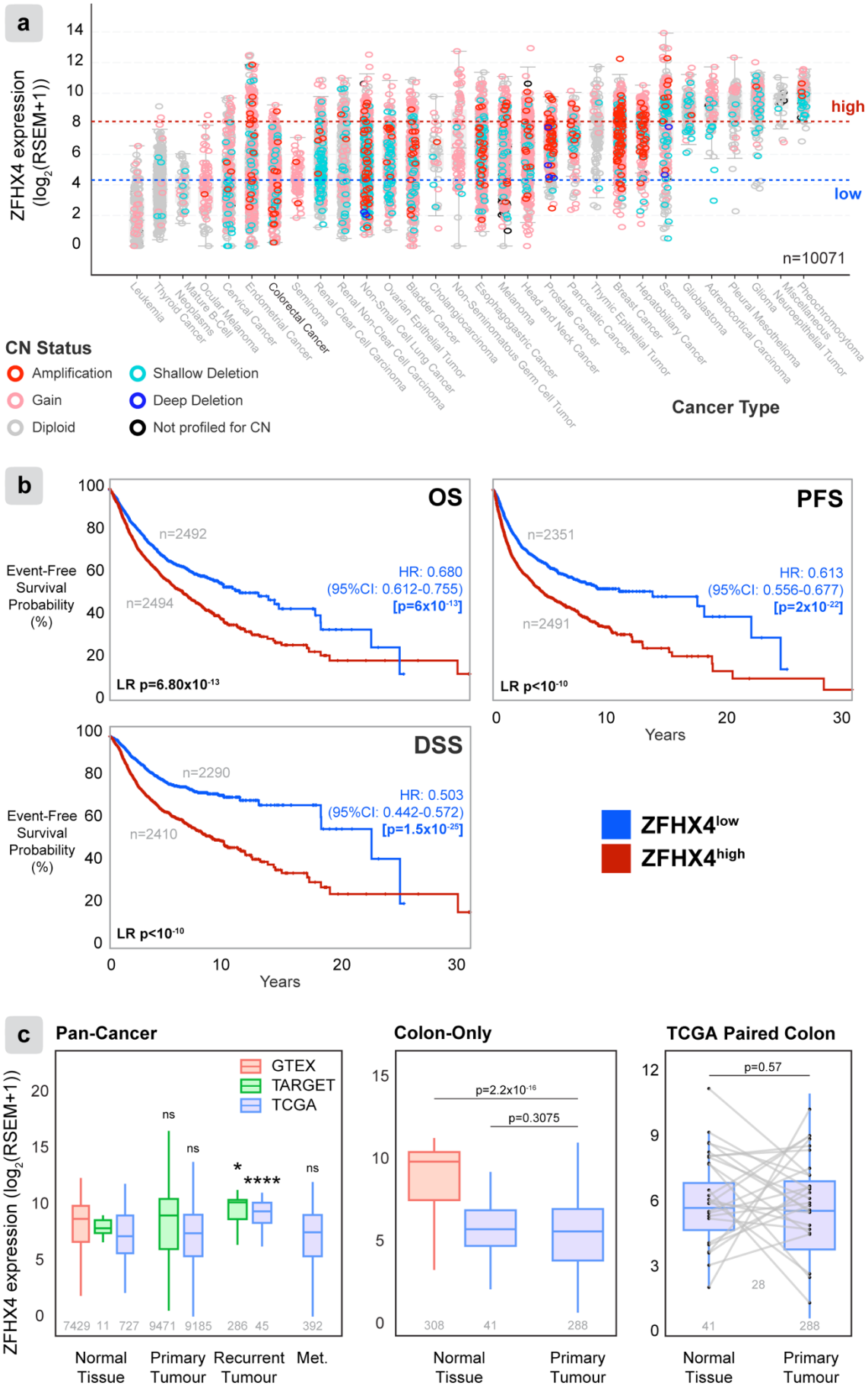

a
ZFHX4 expression (log2(RSEM+1))
high
low
n=10071
Leukemia
Thyroid Cancer
Neoplasms Mature B-Cell
Ocular Melanoma
Cervical Cancer
Endometrial Cancer
Colorectal Cancer
Seminoma
Renal Clear Cell Carcinoma
Renal Non-Clear Cell Carcinoma
Non-Small Cell Lung Cancer
Ovarian Epithelial Tumor
Bladder Cancer
Cholangiocarcinoma
Non-Seminomatous Germ Cell Tumor
Esophagogastric Cancer
Melanoma
Head and Neck Cancer
Prostate Cancer
Pancreatic Cancer
Thymic Epithelial Tumor
Breast Cancer
Hepatobiliary Cancer
Sarcoma
Glioblastoma
Adrenocortical Carcinoma
Pleural Mesothelioma
Glioma
Miscellaneous
Neuroepithelial Tumor
Pheochromocytoma
CN Status
Amplification
Gain
Diploid
Shallow Deletion
Deep Deletion
Not profiled for CN
Cancer Type
b
OS
n=2492
n=2494
HR: 0.680
(95%CI: 0.612-0.755)
[p=6x10-13]
LR p=6.80x10-13
PFS
n=2351
n=2491
HR: 0.613
(95%CI: 0.556-0.677)
[p=2x10-22]
LR p<10-10
DSS
n=2290
n=2410
HR: 0.503
(95%CI: 0.442-0.572)
[p=1.5x10-25]
LR p<10-10
Event-Free Survival Probability (%)
Years
ZFHX4low
ZFHX4high
c
Pan-Cancer
Colon-Only
TCGA Paired Colon
ZFHX4 expression (log2(RSEM+1))
GTEX
TARGET
TCGA
ns
*
****
p=2.2x10-16
p=0.3075
p=0.57
7429 11 727 9471 9185 286 45 392
308 41 288
41 28 288
Normal Tissue
Primary Tumour
Recurrent Tumour
Met.

Figure 4.3 | ZFHX4 mRNA expression across multiple cancers and their prognosis in the TCGA-PanCancerAtlas dataset. (A) ZFHX4 mRNA expression across multiple cancers, ordered by increasing median expression by grey box plots overlaid by data points. CN status has also been annotated with coloured circles represented in the key. Values depict RSEM batch normalisation from Illumina HiSeq RNASeqV2. TCGA-PanCancer Atlas data was obtained from cBioPortal (Cerami et al., 2012). **(B)** Overall (OS), progression free (PFS) and disease specific (DSS) survival of patients with either high or low ZFHX4-expressing tumours. Low and high expressing tumours were segregated by upper and lower quartiles of total ZFHX4 expression, depicted with y-axis lines in (A). Log-rank and cox proportional hazard testing was performed by cBioPortal and HRs are with respect to the $ZFHX4^{low}$ group. All data was accessed and downloaded from cBioPortal directly as the whole PanCancerAtlas dataset was never analysed in raw format. Without direct mentioning, the assumption of ZFHX4 mRNA expression is based on the MANE transcript. **(C)** Comparison of ZFHX4 expression (RSEM) between normal and cancer tissue. Data was obtained from UCSC Xena Browser and GTEx-TARGET-TCGA harmonisation dataset (n=18089 samples). All three plots are coloured by respective study. N numbers are displayed in grey underneath each box plot displaying median and quartiled expression (outliers are not shown). P-values for the pan-cancer and colon-only plots were generated by a non-parametric unpaired wilcoxon test and are respective to the normal tissue of the same dataset (no GTEx comparisons are depicted). For the TCGA paired-colon plot, extra dots are overlaid which represent patients that had both a normal and primary tumour sample, with lines connecting the match. A non-parametric paired wilcoxon test was performed between the two groups of these pairs (n=28 patients, n=52 samples). All x-axes show ZFHX4 expression in TPM. *[LR = log-rank, HR = hazard ratio, ns = p>0.05, * = p<0.05, **** = p<0.0001]*

Comparing both pan-cancer and colon cancer or colon adenocarcinoma (COAD) expression to normal tissue samples using the GTEx-TARGET-TCGA harmonisation dataset from UCSC Xena Browser (Goldman et al., 2020), *ZFHX4* expression was not significantly different when compared to normals from the same dataset (Figure 4.3C). There was significantly higher *ZFHX4* expression in pan-cancer recurrent tumour samples comparative to normal samples. Furthermore, GTEx-normals had significantly higher *ZFHX4* expression compared to TCGA-normals ($p<2.2x10^{-16}$), -primary tumours ($p<2.2x10^{-16}$) and -metastatic tumours ($p<2.2x10^{-16}$) following non-parametric wilcoxon testing. Focusing on colon tissue specifically, GTEx-normals again had significantly higher *ZFHX4* expression than TCGA-normals ($p=1.959x10^{-10}$) as well as TCGA-primary tumours ($p<2.2x10^{-16}$). However, when comparing TCGA-normals to primary tumours, *ZFHX4* expression was not significantly different (p=0.3075). Furthermore, this was consistent when focusing on normal-tumour matching pairs (n=28) following paired wilcoxon testing (p=0.57). This suggests that *ZFHX4* expression in CC is not significantly different compared to matching normal tissue, however expression in general is relatively low and does vary dramatically between different colon tumours, conferring significant prognoses.

Given the strong prognostic value of *ZFHX4* across cancers, I next assessed whether this pattern was consistent in COAD specifically. Across five independent COAD cohorts patients were stratified into high and low *ZFHX4* expression groups and consistently, high expression was associated with poorer outcomes (Figure 4.4). Although two of the five datasets didn't reach statistical significance, all cohorts showed a clear trend whereby low *ZFHX4* expressors had improved recurrence/relapse-free survival, progression-free survival, and overall survival. For example, in the Nunes cohort $ZFHX4^{low}$ patients had significantly better recurrence-free survival (log-rank (LR) p=0.00445), with a significant 57% reduced risk of recurrence over 20 years (hazard ratio (HR)=0.4306, 95%CI:0.2368-0.7829, p=0.00574). This trend was also significant in the Marisa and Guinney cohorts and was evident though not statistically significant in the Sidra cohort, likely due to limited sample size. Overall, low *ZFHX4* expression was associated with a mean 34% reduced risk of recurrence (HR=0.658) across all cohorts.

Consistent trends were also observed for overall survival. In the Guinney and Nunes cohorts, low *ZFHX4* expression conferred a significant 17–34% reduction in risk of death (p=0.038 and p=0.01 respectively). Comparable but non-significant trends were observed in the TCGA cohort, again likely reflecting limited sample numbers. Pooled across cohorts low *ZFHX4* expression was associated with a mean 24% reduction in risk of death (mean HR=0.763). It should be noted that these cohorts differ in follow-up duration, stage distribution and treatment exposure.

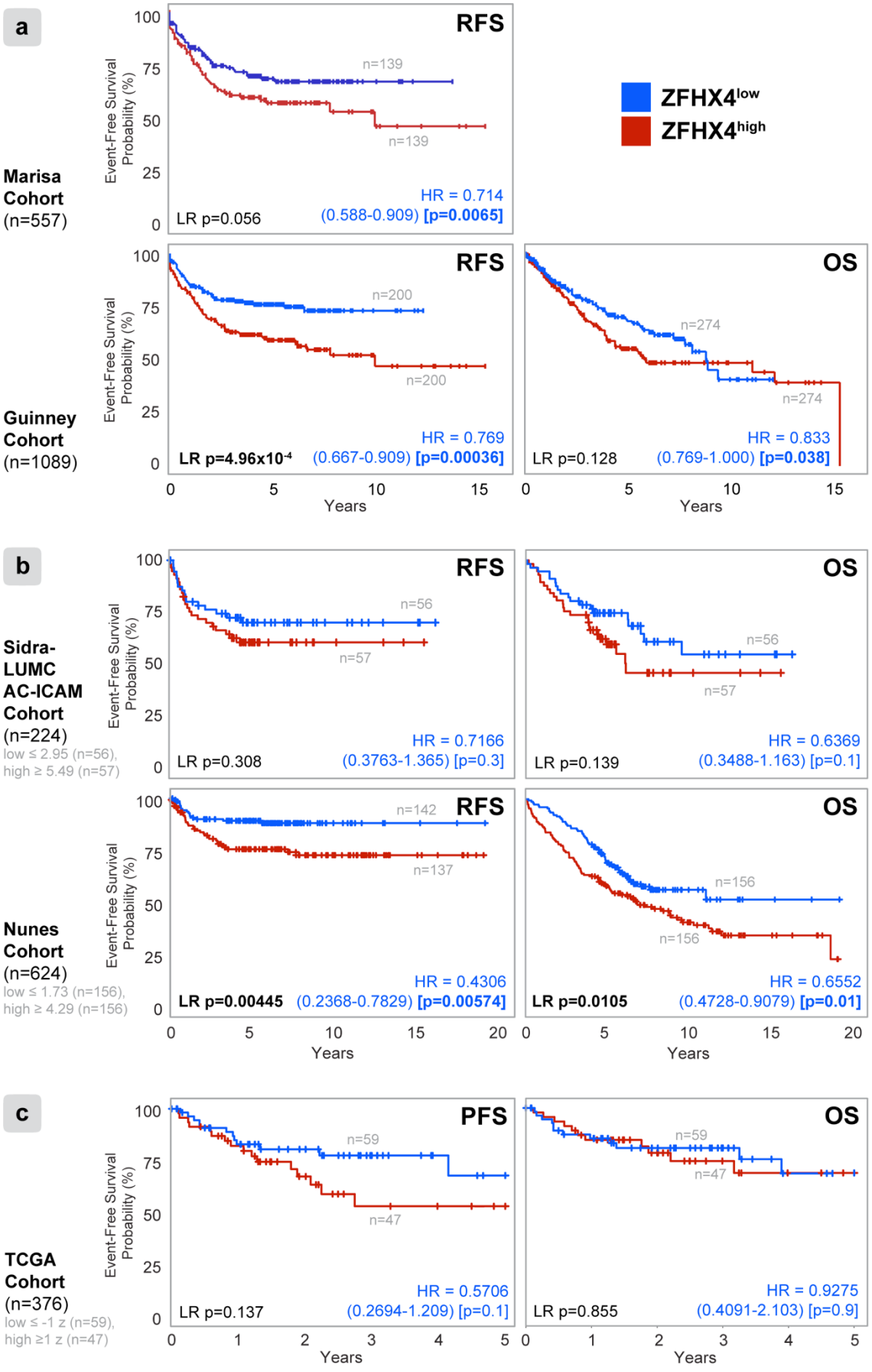

a
Marisa Cohort (n=557)
RFS
n=139
n=139
HR = 0.714
(0.588-0.909) [p=0.0065]
LR p=0.056
ZFHX4low
ZFHX4high
Guinney Cohort (n=1089)
RFS
n=200
n=200
HR = 0.769
(0.667-0.909) [p=0.00036]
LR p=4.96x10-4
OS
n=274
n=274
HR = 0.833
(0.769-1.000) [p=0.038]
LR p=0.128
Event-Free Survival Probability (%)
Years
b
Sidra-LUMC AC-ICAM Cohort (n=224)
low ≤ 2.95 (n=56), high ≥ 5.49 (n=57)
RFS
n=56
n=57
HR = 0.7166
(0.3763-1.365) [p=0.3]
LR p=0.308
OS
n=56
n=57
HR = 0.6369
(0.3488-1.163) [p=0.1]
LR p=0.139
Nunes Cohort (n=624)
low ≤ 1.73 (n=156), high ≥ 4.29 (n=156)
RFS
n=142
n=137
HR = 0.4306
(0.2368-0.7829) [p=0.00574]
LR p=0.00445
OS
n=156
n=156
HR = 0.6552
(0.4728-0.9079) [p=0.01]
LR p=0.0105
c
TCGA Cohort (n=376)
low ≤ -1 z (n=59), high ≥1 z (n=47)
PFS
n=59
n=47
HR = 0.5706
(0.2694-1.209) [p=0.1]
LR p=0.137
OS
n=59
n=47
HR = 0.9275
(0.4091-2.103) [p=0.9]
LR p=0.855

Figure 4.4 | Stratifying COAD patients by ZFHX4 expression confers significant prognoses across multiple cohorts. Expression stratification was by upper and lower quartile designation for plots in A-B. In C, patients were stratified by zscore expression with cut offs -1 to 1. Groups are highlighted by the blue and red colours denoted in the key. LR and HR results are annotated with respect to ZFHX4^{low} in blue and all signficiant p-values are highlighted in bold. **(A)** The Marisa (Marisa et al., 2021) and Guinney (Guinney et al., 2015) cohorts were accessed through the 'R2 Genomics Analysis and Visualisation' online portal (Koster, 2025) where KM, LR and coxPH analyses were also conducted. Due to platform limitations, all COAD types were utilised (i.e. non-removal of mucinous samples). **(B-C)** The Sidra (Roelands et al., 2023), Nunes (Nunes et al., 2024) and TCGA-COAD (Hoadley et al., 2018) cohorts were analysed in-house using original data downloads in R pertaining to only primary colon adenocarcinoma histology. In C, survival data was clipped and patients were censored from 5 years for more representative plots. This data was obtained from a TCGA-clinical data resource (Liu et al., 2018). KM, LR and coxPH analyses were performed in R. *[KM = Kaplan Meier, HR = hazard ratio, LR = log-rank, coxPH = cox proportional hazards, RFS = relapse/recurrence-free survival, PFS = progression-free survival, OS = overall survival]*

As *ZFHX4* expression didn't distinguish primary tumours from matched normal tissue, and high expression consistently predicted poor survival, the next step was to examine the associations with tumour stage. Interestingly, *ZFHX4* expression was significantly higher in stage IV tumours than stage I tumours in most datasets (Figure 4.5A; TCGA-COAD/Nunes/SIDRA/CPTAC2). When combining all four datasets, its expression positively correlated with increasing tumour stage when stage was treated as both a numerical variable (spearman's $\rho=0.153$, $p=2.8\times10^{-8}$) and ordinal variable (linear regression ß=0.177 per stage, $p=1.14\times10^{-8}$) (Figure 4.5B). As increased stage reflects local invasion, lymph node involvement, metastasis and an aggressive tumour phenotype, this signifies ZFHX4 is more likely implicated in tumour progression rather than initiation. Further adjustment for stage in multivariable models would clarify whether ZFHX4 functions as an independent prognostic biomarker or instead represents a molecular correlate of advanced disease. Additionally, investigation of matched metastatic samples, limited for these cohorts, would help to further define its role in late-stage tumour biology.

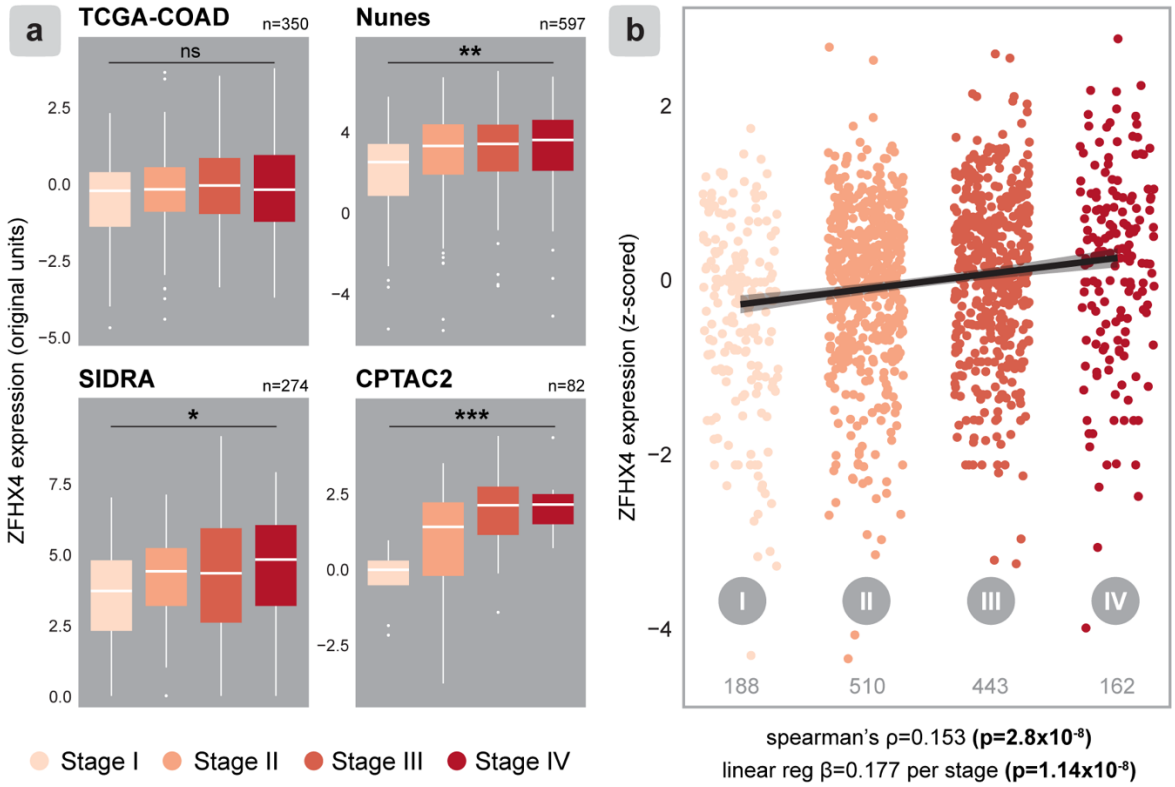


Figure 4.5 | ZFHX4 expression increases with colon cancer stage. (A) ZFHX4 expression in four independent colon cancer cohorts segregated by tumour-node-metastasis stage. Y axes are with reference to expression units of each dataset. Statistical testing was performed with non-parametric wilcoxon comparing ZFHX4 expression between stage I and IV. **(B)** Scatter plot of z-score within-cohort-scaled ZFHX4 expression versus stage (I-IV) as an ordinal parameter across all four cohorts (n=1303). Each point represents a single patient. The black line indicates the linear regression trend with 95% confidence interval. Spearman correlation and linear regression tests are denoted. N numbers per stage are written in grey.

Given the potential for *ZFHX4* expression in both epithelial and stromal compartments as observed in the healthy colon, scRNAseq data from COAD tissue was analysed to determine whether *ZFHX4's* association with poor prognosis reflects pro-tumour stromal biology and/or tumour cell-intrinsic biology. One of the largest scRNAseq datasets ('Multicellular Hubs') encompassing normal and COAD cells was interrogated (Pelka et al., 2021). To note, pre-cancerous and COAD samples were also provided by the HTAN Set1 and Set2 datasets but were not further explored due to low COAD cell numbers. In matched normal cells, 0.12% (n=135) of cells expressed *ZFHX4* and this was predominantly stromal in origin (85%), pertaining to fibroblasts which secrete stem-niche factors (Figure 4.6).

ZFHX4 was also expressed in neuronal (Schwann) and endothelial (LEC) cells but in numbers <10. Interestingly, COAD tumours exhibited a significantly higher proportion of *ZFHX4*$^{+}$ cells (n=1635) compared with normal tissue (0.64% vs 0.12%). Fisher's exact testing indicated that tumours were significantly 5.3-fold more likely to contain *ZFHX4*$^{+}$ cells than matched-normals (odds ratio=5.34, 95%CI:4.48–6.42, $p<2.2x10^{-16}$). This association was confirmed by Pearson's X^2 test ($X^2=437.7$, df=1, $p<2.2x10^{-16}$). *ZFHX4* was primarily expressed in epithelium pertaining to stem/TA-like cells and the majority of other *ZFHX4* expressing cell types were different CAF subsets. Interestingly, *ZFHX4* expression levels in *ZFHX4*-expressing epithelial cells from normal tissue were unusually low (max 0.5 counts per 10k, 6000-fold lower relative to *ACTB*) but were 27-fold higher in epithelial cells from COAD tissue relative to *ACTB* (max 38 counts per 10k, 226-fold lower relative to *ACTB*) suggesting a potential tumour-specific induction of *ZFHX4* expression in epithelial cells.

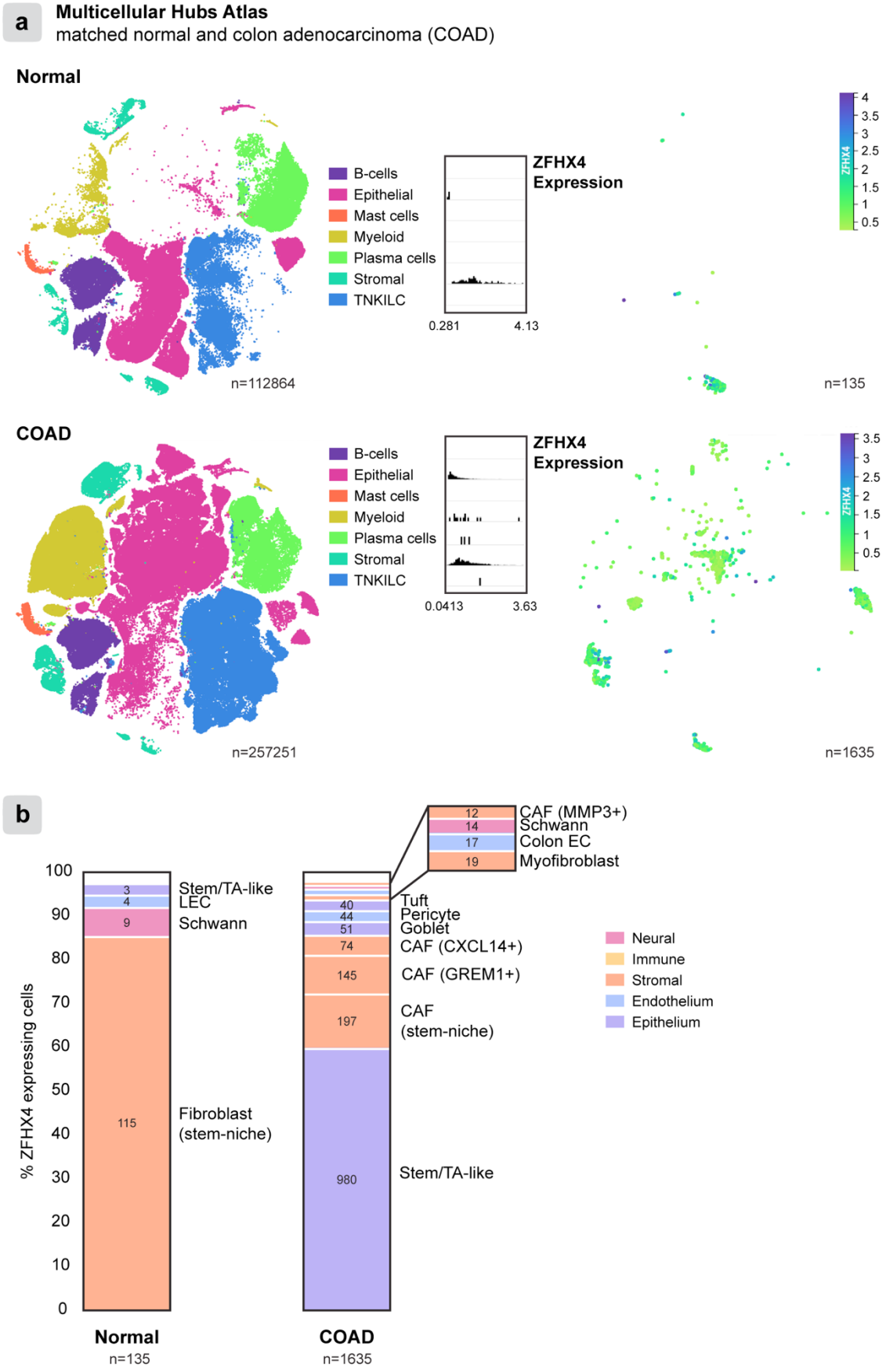

a
Multicellular Hubs Atlas
matched normal and colon adenocarcinoma (COAD)
Normal
B-cells
Epithelial
Mast cells
Myeloid
Plasma cells
Stromal
TNKILC
ZFHX4 Expression
0.281
4.13
n=112864
n=135
COAD
0.0413
3.63
n=257251
n=1635
b
% ZFHX4 expressing cells
Stem/TA-like
LEC
Schwann
Fibroblast (stem-niche)
CAF (MMP3+)
Schwann
Colon EC
Myofibroblast
Tuft
Pericyte
Goblet
CAF (CXCL14+)
CAF (GREM1+)
CAF (stem-niche)
Stem/TA-like
Neural
Immune
Stromal
Endothelium
Epithelium
Normal
n=135
COAD
n=1635

Figure 4.6 | ZFHX4 expression at the single cell level in COAD and matched normal tissue. (A) tSNE plots on the left are annotated by the lowest level of cell type annotation. Matching colour keys also depict mini density plots of ZFHX4 expression (x axis also denoted with scales of expression). tSNE plots on the right depict cells that express ZFHX4 (non-zero filtering) to degrees respective to the viridis colour scale on each plot. N numbers represent the number of ZFHX4-expressing cells. **(B)** Proportional stacked bar charts showing the contribution of cell types expressing ZFHX4 in normal vs COAD tissue. Colours are coded by the key of general lineage annotation. This dataset was obtained from (Pelka et al., 2021).

4.4.1.3 ZFHX4 protein expression in normal and cancer intestinal tissues

Presenting with low consistency to mRNA expression data, ZFHX4 protein expression in the Human Protein Atlas (Uhlen et al., 2015) appeared to be high in all healthy adult tissues (Supplementary Figure S4.7A), demonstrating ubiquitous nuclear expression in the colon and small intestine in epithelial and stromal compartments (Supplementary Figure S4.7B). In cancer including COAD, staining was moderate to strong and largely epithelial, consistent across 16 COAD samples (Supplementary Figure 4.8A/B). This contrasts with scRNA-seq which suggests rarer expression, and with no published ZFHX4-IHC, I performed my own staining. ZFHX4-IHC was initially optimised using small intestine tissue (patient #11563) following verbal communication (Chudnovsky et al., 2014). All five commercially available ZFHX4 antibodies displayed divergent patterns irrespective of concentration (Supplementary Note 4.1, Table 4.3) but overall, the HPA antibody at 1:100 offered the best balance of specificity and reproducibility, avoiding M02/M11 background while confirming nuclear localisation and an interesting subset also identified by abc.

Table 4.3 | Staining profiles of ZFHX4 on small intestine by antibody, irrespective of conc.

Ab. name	Epithelium			Stromal		Evidence of non-stained cells (Hmtxl only)
	Nuclei – general	Nuclei – high in very specific cells	Cytoplasmic	Nuclei	Cytoplasmic	
HPA^{old}	☑	☑	weak-mid	☑	weak	☑
HPA^{new}	☑	☑	mid-high	☑	mid	☑
PA5	☑	☒	high	☑	mid	☑

Abc	☑	☑	weak	☑	weak	☑
M02	☑	☒	very high	☑	high	very few
M11	☑	☒	very high	☑	high	very few

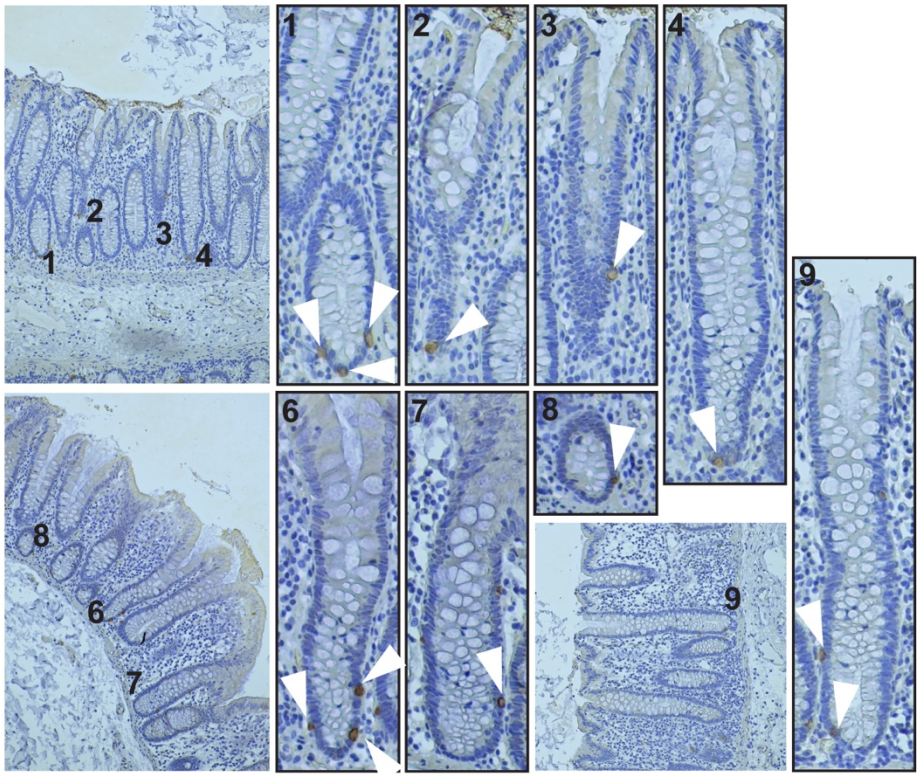
HPAnew = newer Ab batch, HPAold = older Ab batch, Hmtxl = haematoxylin

Resulting from previous optimisation work, the HPA antibody was selected for staining matched normal colon, small intestine and COAD tissue from two patients, alongside no-primary, β-catenin, villin and ZFHX4-M02 controls. No-primary controls conducted on small intestine were successful, demonstrating only blue Haematoxylin staining, and β-catenin and villin controls were representative of the Human Protein Atlas across normal, small intestine and COAD tissue (Supplementary Figure S4.9). Normal colon tissue structure was slightly different between patients, with #11563 displaying thin crypts and #16525 displaying wider and longer crypts, likely a result of tissue procurement from different colonic regions where crypt structures differ. Also, in general staining patterns irrespective of antibody are consistent between both patients, however #16525 staining appears to be much weaker/lighter in comparison, including haematoxylin. Even though IHC staining was performed at the same time, this might have been due to slight differences in fixation, embedding or processing as patient samples were obtained at different times. ZFHX4-M02 was additionally tested as a backup for ZFHX4-HPA and again showed poor nuclear specificity in all tissues, however intriguingly for COAD tissue in both patients most epithelial nuclei seem to be devoid of expression, as this seems to be restricted to stromal nuclei.

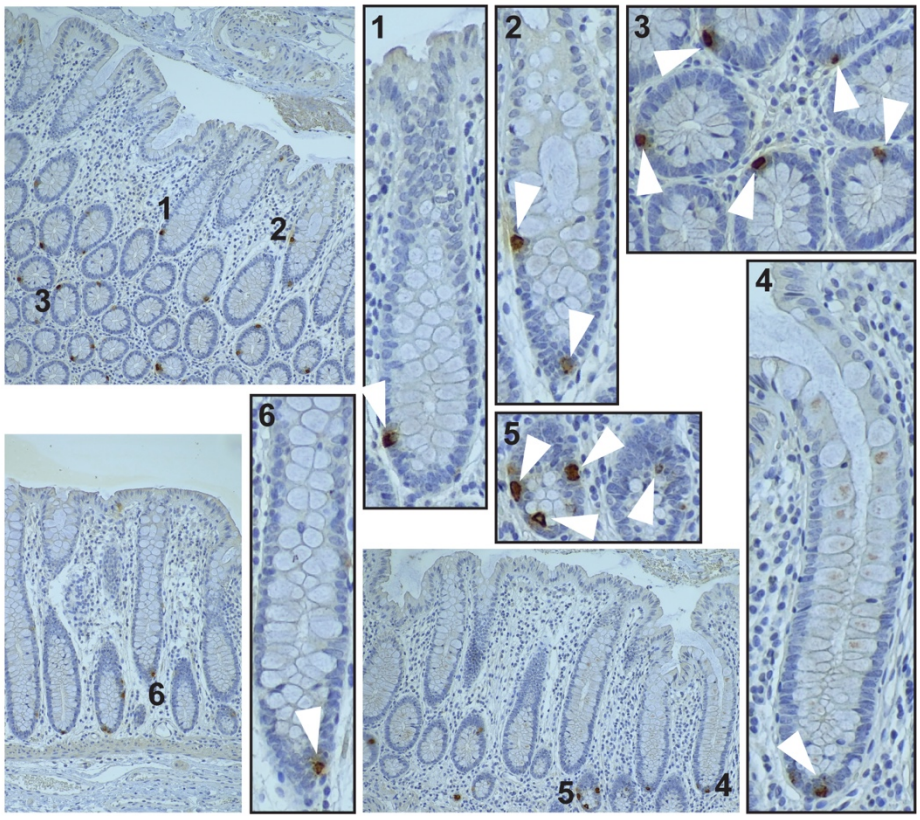
Using the ZFHX4-HPA antibody on two different normal colon tissues, staining revealed very specific strong brown/black signal in individual cells consistently found at the crypt bottom, with 1 per whole crypt on average (Figure 4.7). There was no apparent staining anywhere else, suggesting specific epithelial cell subsets residing towards the bottom of normal colonic crypts express high levels of ZFHX4 protein. In small intestine tissue, the same pattern was observed however detection was also in cells in villi as well as crypts,

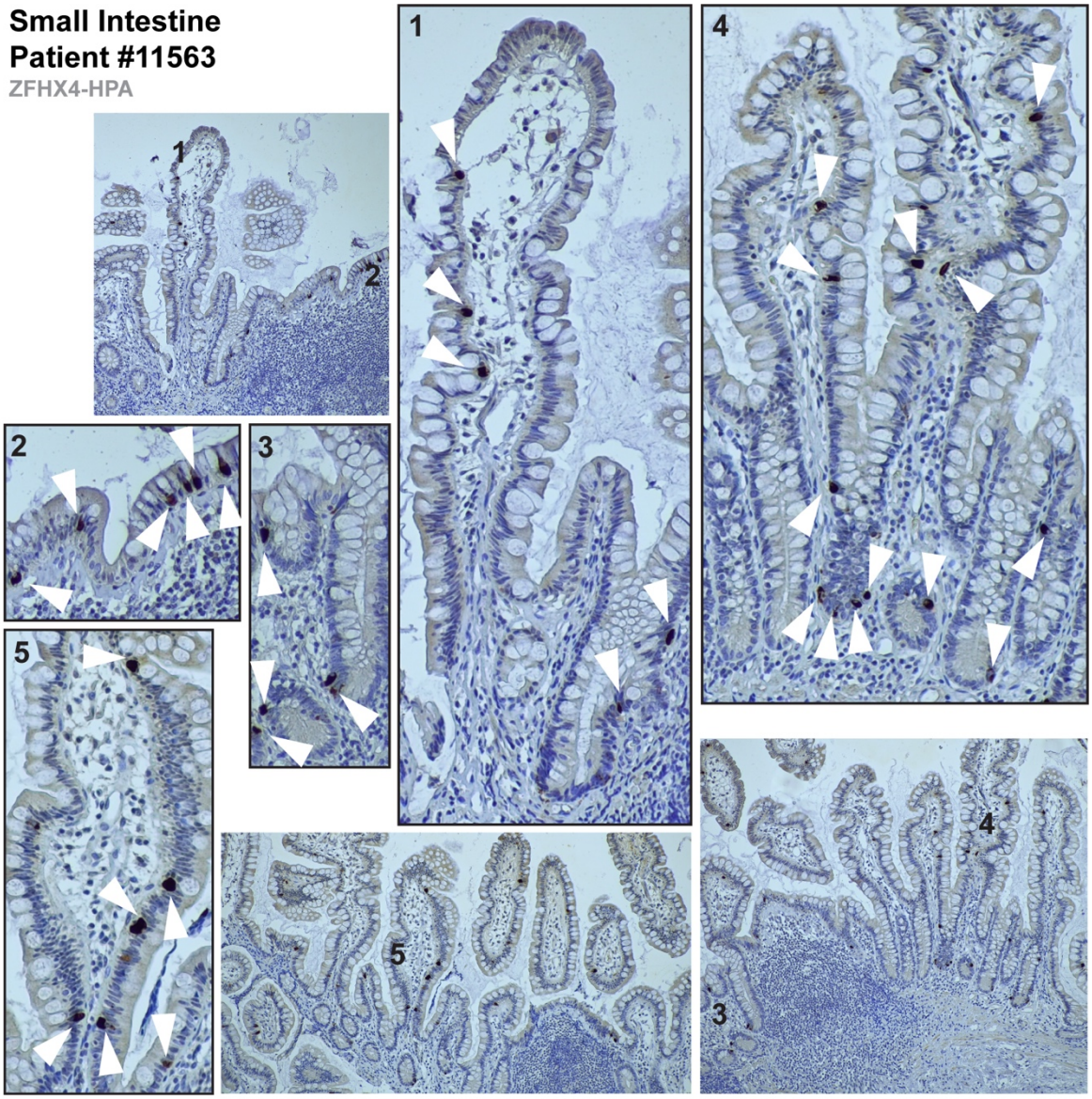
with up to 5 detected per villus. This staining demonstrated slightly different patterns to the previous optimisation run on the same tissue from one of the same patients (#11563), displaying less background cytoplasmic epithelial staining which likely refined this positive and specific signal. In COAD tissue, these cells were still present at regions of retained organised structure, but in one patient (#11563) expression intensity in both epithelial nuclei and cytoplasm was heightened in some areas towards the tissue edge, likely indicating an invasive edge. Furthermore, high expression in some stromal cells particularly surrounding endothelium were found in another patient (#16525).

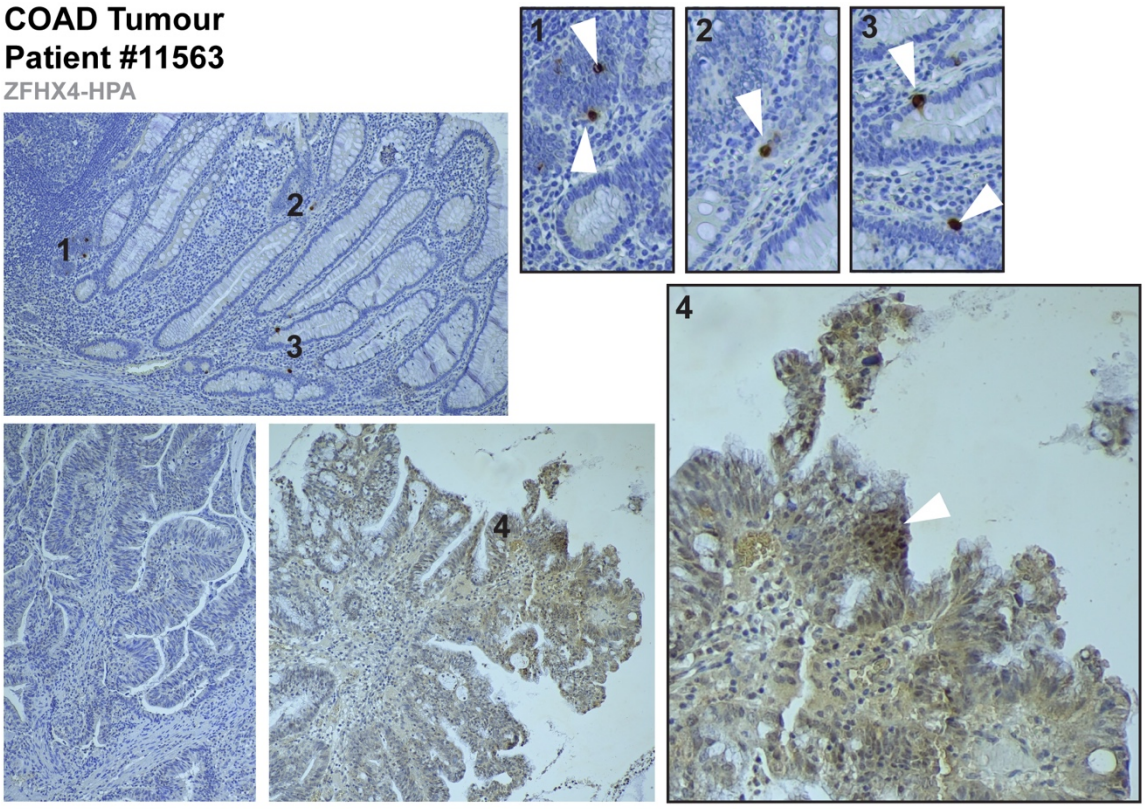
Normal Colon
Patient #11563
ZFHX4-HPA



Normal Colon
Patient #16525
ZFHX4-HPA




Small Intestine
Patient #11563
ZFHX4-HPA
1
2
3
4
5
Small Intestine
Patient #16525
ZFHX4-HPA
1
2
3
4
5


COAD Tumour
Patient #11563
ZFHX4-HPA
1
2
3
4

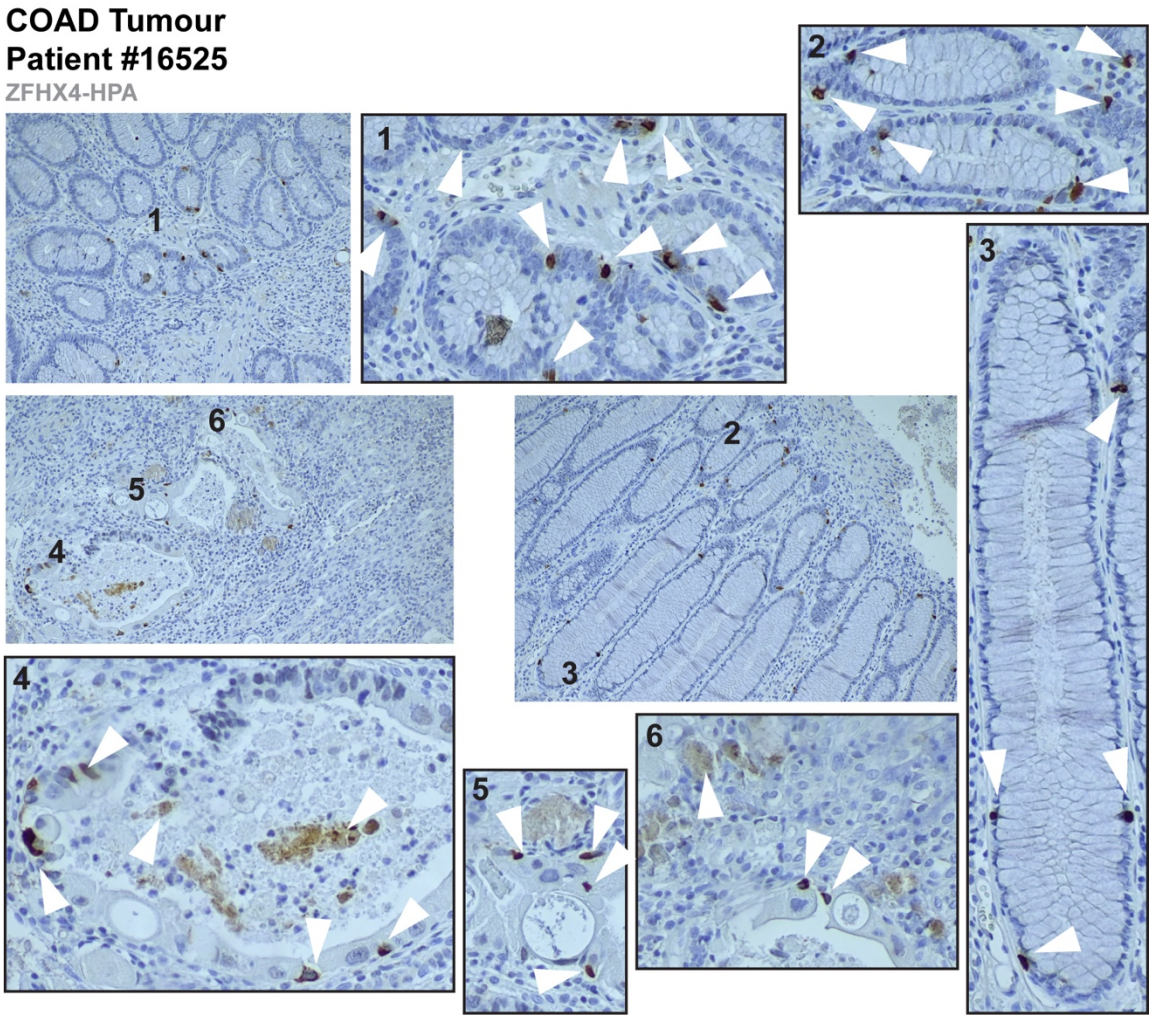

COAD Tumour
Patient #16525
ZFHX4-HPA
1
2
3
4
5
6

Figure 4.7 | ZFHX4 protein expression across matched normal colon, small intestine and COAD tissue from two patients. Slides were counterstained with Haematoxylin (blue signal) and DAB reagent produced brown signal for antibody binding. All images were taken at 10X and zoomed images highlighted with black boxes and annotated with respective number represent 30X. White triangles depict distinct dark brown/black staining.

In summary, despite showing a different staining pattern to that seen in the Human Protein Atlas, IHC here using the ZFHX4-HPA antibody consistently highlighted a specific epithelial subset residing in the bottom of crypts and in villi across all matched tissue types of both patients, indicating that this result is likely biologically significant and not a technical artefact. Furthermore, this is consistent with scRNAseq data of normal and cancer intestinal tissue. Specific cell subsets that exist in both crypts and villi include enteroendocrine and tuft cells, but alongside the scRNAseq data which suggests ZFHX4 can be expressed by multiple epithelial cell types, those identified in crypts may indicate a stem cell/transit-amplifying cell.

4.4.2 Generation of *ZFHX4*$^{-/-}$ cell line and organoid models

Building on Chapter 3 and the preceding literature, *ZFHX4* is frequently altered by missense mutations that associate with enhanced immunogenicity, whereas high *ZFHX4* expression correlates with increased tumorigenicity. In principle, modelling clinically observed missense variants would be the most representative way to interrogate ZFHX4's role in immunogenicity however in practice, precise missense modelling is technically demanding and achieved only with Cas9-directed homology directed repair (HDR), base editing or prime editing which are further discussed in Supplementary Note 4.2. Given that my work spans both cell lines and organoids, a standardised, translatable editing strategy was required. The Buczacki lab already has optimised, high-efficiency multi-guide CRISPR/Cas9-KO pipelines in colon organoids (Chan et al., 2023b). Therefore, I applied traditional CRISPR/Cas9-KO across both model systems as a pragmatic compromise to interrogate tumorigenicity in this chapter and immunogenicity in Chapter 5: complete functional abrogation and subsequent phenotyping are well-suited to reveal mechanisms by which ZFHX4 regulates both processes.

COLO320DM and NCIH716 colon cell lines were chosen based upon TMBL and *ZFHX4*-wildtype (WT) status as well as the highest *ZFHX4* expressing colon lines in both mRNA (MANE) and protein (Q86UP3) expression in the DepMap database (n=263, Figure 4.8A, (DepMap Broad Institute, 2025)). The T-cell leukaemia (Jurkat) and Ewing Sarcoma (SKNMC) lines were selected as negative and positive controls respectively based upon mRNA and protein expression, with SKNMCs expressing the third highest *ZFHX4* MANE mRNA of all known cell lines (n=1684). Furthermore, the ZFHX4-MANE transcript was the most expressed variant by proportion, representing >80% *ZFHX4* total mRNA expression across COLO320DM, NCIH716 and SKNMC cell lines (Supplementary Figure S4.10).

For organoid model selection, considerable time was spent working on deriving and culturing normal colon organoids from healthy colon tissue obtained from COAD patients

undergoing surgical resection. These organoids were prioritised because they provide a genetically “clean” background, enabling clearer biological interpretation of KO effects without the additional confounding variants typically present in COAD-derived organoids. However, several challenges hindered progress including inefficient CRISPR editing, poor organoid growth, recurrent bacterial and fungal contamination, limited availability of DNA/RNA/protein material for downstream experiments and sudden cell death or senescence. Although such issues are common in routine normal organoid culture and consistently optimised in the Buczacki group, they ultimately prompted a shift in approach. Instead, I utilised pre-established and cryopreserved organoid lines generated by Dr Collins whom had also performed matched RNAseq and WES on healthy, primary tumour and metastatic tumour derived organoids from several different patients as part of a biobank collection project. Among these, the line with the highest *ZFHX4* mRNA expression, TMBL and *ZFHX4*-WT was a metastatic organoid line (Figure 4.8B). Tumour-derived organoids such as these grow extremely well and despite a mutated background, this line was selected to enable rapid inference of ZFHX4 biology following CRISPR-mediated KO.

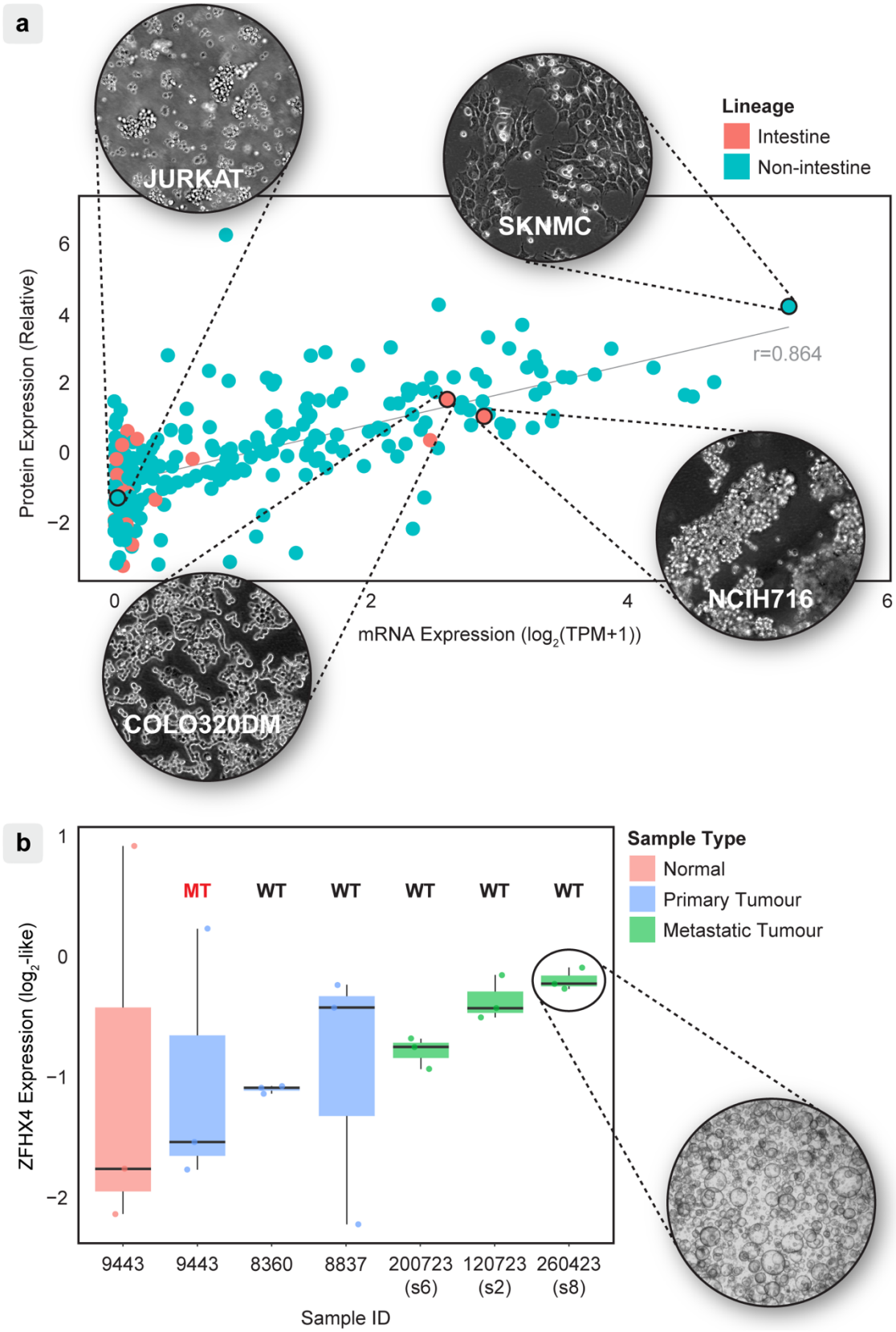


Figure 4.8 | Model selection for studying ZFHX4 in cell lines and organoids. (A) Matched mRNA-protein expression of ZFHX4 in 263 cell lines, highlighting those selected for further study. Data obtained from DepMap (DepMap Broad Institute, 2025). Microscope images are 10X magnification and upscaled 1.75x. **(B)** mRNA expression of ZFHX4 in a biobank of colon organoid lines, annotated mutation status of ZFHX4, organoid type and selection of organoid line for study. Data obtained from the Buczacki lab. Microscope image is 1.25X and downscaled 1.7x.

CRISPR/Cas9 was employed in these models to generate homozygous *ZFHX4*$^{-/-}$ knockouts (KOs). A multi-guide RNA (mgRNA) approach targeting the 5'-end of exon 2 consisted of three single-guide RNAs (sgRNAs) binding 20bp regions and spaced roughly 40bp apart (Figure 4.9A). These are designed to generate DSBs at multiple sites, producing >21bp fragment deletions at the target locus. This approach is more efficient than using one sgRNA and increases the likelihood of complete gene inactivation (Synthego and EditCo, USA). Ultimately, this should disrupt mRNA transcription and translation into protein and thus deplete or truncate total ZFHX4 protein as a result, causing LoF.

Equivalent CRISPR strategy was applied across all models. In cell lines, this was extended to incorporate single-cell cloning to generate monoclonal populations with uniform *ZFHX4* disruption (Figure 4.9B). Here, polyclonal KOs (polyKOs) were expanded and single-cell cloned by limiting dilution following KO confirmation by sanger sequencing. Clones were named and characterised through downstream experiments. Nomenclature was based on the cell line of origin (e.g. 'C' for COLO320DM) with a suffix denoting clone identity or *ZFHX4* status (Figure 4.9C). For enhanced controls, 'nosg' cells (Cas9 electroporation only) were used as WT comparators in all KO analyses. Parental lines were cryopreserved and excluded from subsequent experiments unless otherwise specified.

In organoid models, single-cell cloning wasn't feasible, therefore polyKOs were used as representative KOs when KO efficiency exceeded 80% by sanger sequencing. Both 'noCRISPR' and 'nosg' controls were included to extrapolate potential electroporated-related effects.

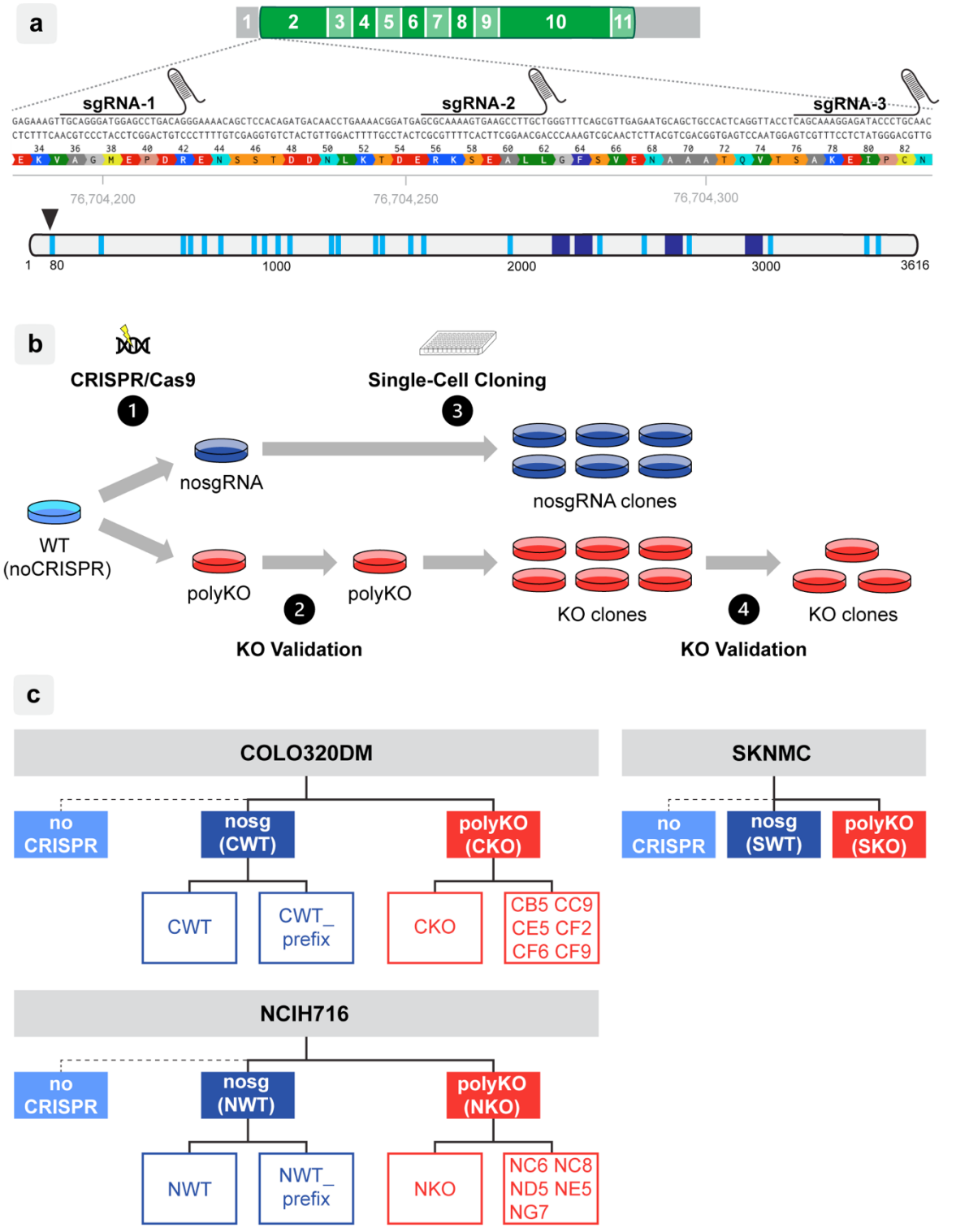


Figure 4.9 | CRISPR/Cas9 strategy employed to generate *ZFHX4*$^{-/-}$ models and additional pipelines in cell lines to obtain monoclonal populations. (A) Schematic showing binding regions of three single-guide RNAs (sgRNAs) enabling a multi-guide approach. Regions are accurate to the bp and sequences were aligned on benchling. Sequences in colour blocks represent amino acids and numbers in grey represent chromosomal location according to GRCh38. The *ZFHX4*-MANE mRNA transcript variant and predominant protein isoform Q86UP3 are also shown with respect to Exon 2 targeting. **(B)** Schematic describing the approach to obtain monoclonal populations of *ZFHX4*$^{-/-}$ cell lines. **(C)** Nomenclature classification used throughout the project, based on cell line lineage and ZFHX4 mutation status. Colours in B and C are representative throughout this chapter.

4.4.3 Transcriptomic profiling of *ZFHX4*$^{-/-}$ organoids

The metastatic organoid line described above (#260423-S8) was cultured over a month from P5 to P10, with growth documented by imaging (Figure 4.10A). CRISPR was performed at P9 and P10 using four domes per reaction, kept separate across successive CRISPRs to achieve biological replicates of polyclonal WTs and KOs (n=3 each). One week post-CRISPR of each attempt, some KO replicates displayed fewer and smaller organoids (Figure 4.10B). However, it's difficult to ascertain whether this reflects a true KO effect or variability in plating density, as organoid passaging isn't cell count-based and can be subjective. This observation requires validation through controlled seeding densities, AI-based growth quantification from imaging and viability assays such as CTG. Notably, independent of CRISPR, organoid growth slowed and morphology appeared more differentiated from P5 to P10, suggesting reduced culture fitness over time.

Sanger sequencing was performed on extracted gDNA and trace files were aligned to ZFHX4-GRCh38 on Benchling to estimate genomic disruption (Figure 4.10C). Relative to noCRISPR and nosgRNA controls, a clear 108bp deletion between sgRNA-1 and sgRNA-3 was identified. Trace files were also analysed by Synthego's Inference of CRISPR Edits (ICE) platform which applies statistical modelling to estimate indel spectra and frequencies. ICE reported 0% WT sequence, with an estimated KO efficiency of 81% (frameshift-causing or ≥21bp in length). Multiple small indels were detected at low confidence (48%) likely due to primer proximity to sgRNA-3's region. Given this high editing efficiency, organoids were further expanded for downstream RNAseq analysis.

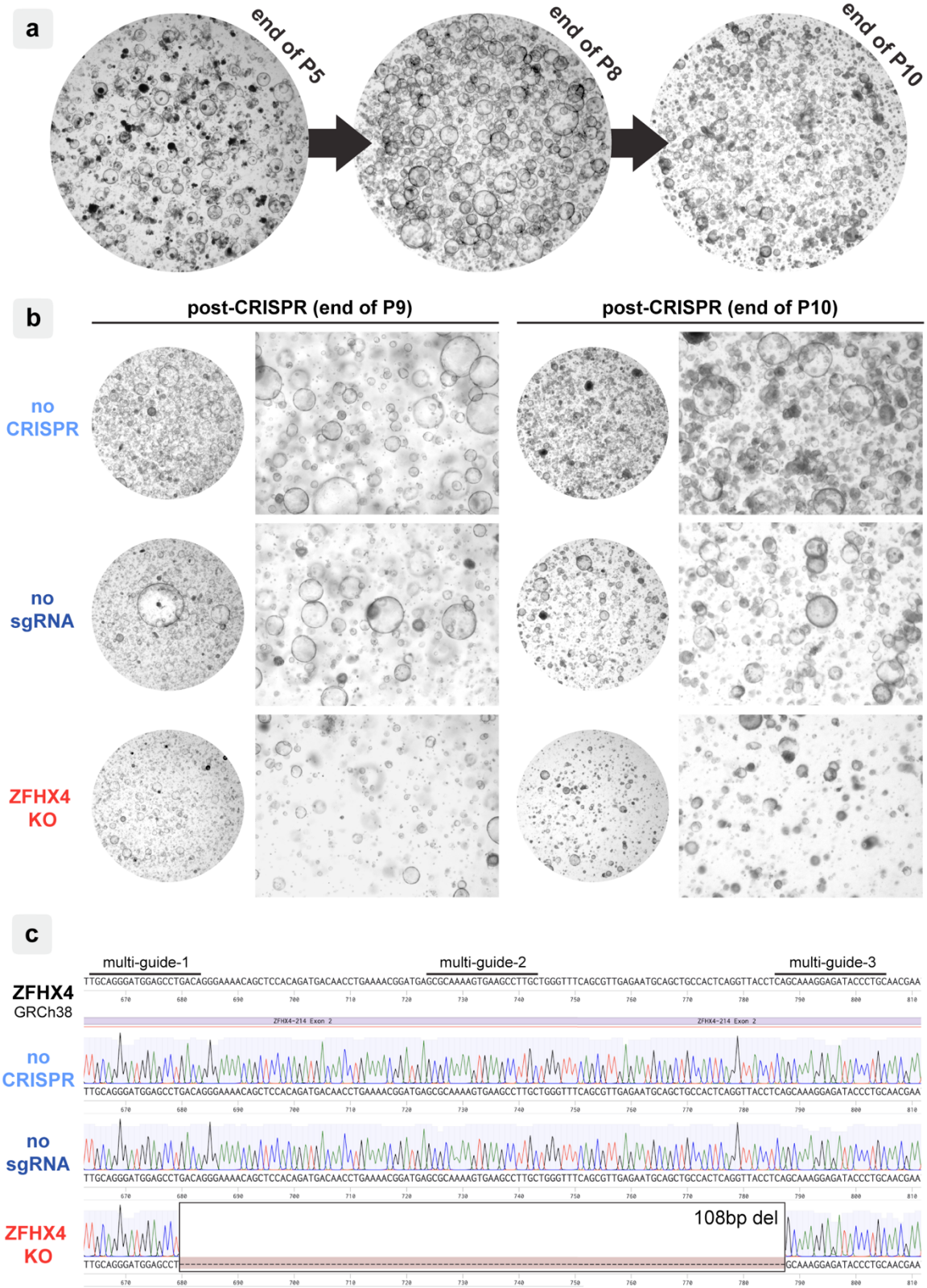


Figure 4.10 | Expansion and CRISPR-mediated KO of ZFHX4 in a metastatic colon organoid line. (A) Images of organoids thawed at P5 which recovered successfully and were expanded up to P10. Images are at 1.25X magnification. **(B)** Images of organoids 7 days post-CRISPR attempt 1 (left) and 7 days post-CRISPR attempt 2 (right), representative of multiple dome technical replicates. Circular images are 1.25X and rectangular images are 4X. **(C)** Trace files of sanger sequencing results aligned to the ZFHX4 Exon 2 region on Benchling software.

RNAseq was performed following approximately 2 weeks of expansion to ensure enough material. Three biological replicates were submitted for each condition (noCRISPR, nosgRNA, *ZFHX4*$^{-/-}$) and these were maintained throughout expansion following the second CRISPR attempt. Initial analysis of RNAseq data demonstrated unexpected clustering of *ZFHX4*$^{-/-}$ replicates with nosgRNA replicates which were separate from noCRISPR replicates on the PCA plot (Figure 4.11A). This separation was observably greater than between *ZFHX4*$^{-/-}$ replicates with nosgRNA replicates, suggesting that electroporation/Cas9-related effects were a greater driving force for gene expression differences than *ZFHX4* perturbations. Additionally, the noCRISPR replicates were segregated along PC1 which reiterates the importance of selecting multiple replicates per condition. Furthermore, *ZFHX4* expression levels were low overall and no significant differences in *ZFHX4* mRNA expression were found between the three conditions (unpaired t-tests) (Figure 4.11A).

These unexpected results prompted sanger sequencing of gDNA extracted at the same time point as RNA for RNAseq (post-expansion) to confirm the high KO percentage obtained two weeks prior (pre-expansion). Sanger sequencing traces showed that the 108bp deletion was no longer visible in any *ZFHX4*$^{-/-}$ replicate, rather subtle disruptions in DNA sequence comparative to noCRISPR and nosgRNA controls overwhelmed with an overall match to wild-type sequence (Figure 4.11B). Synthego-ICE results validated this dramatic increase of wild-type sequences across all three *ZFHX4*$^{-/-}$ replicates comparative to the pre-expansion *ZFHX4*$^{-/-}$, with a confidence of >94% (Figure 4.11C). Each replicate consisted of >75% wild-type populations which had been undetectable two weeks earlier before expansion. These results were also consistent with KO-score estimation. Collectively, this evidence suggests that over the two week expansion period, *ZFHX4*$^{-/-}$ organoid populations were lost and overtaken by selected *ZFHX4*-WT populations, potentially hinting at an interesting phenotype of lethality and/or senescence in the metastatic setting.

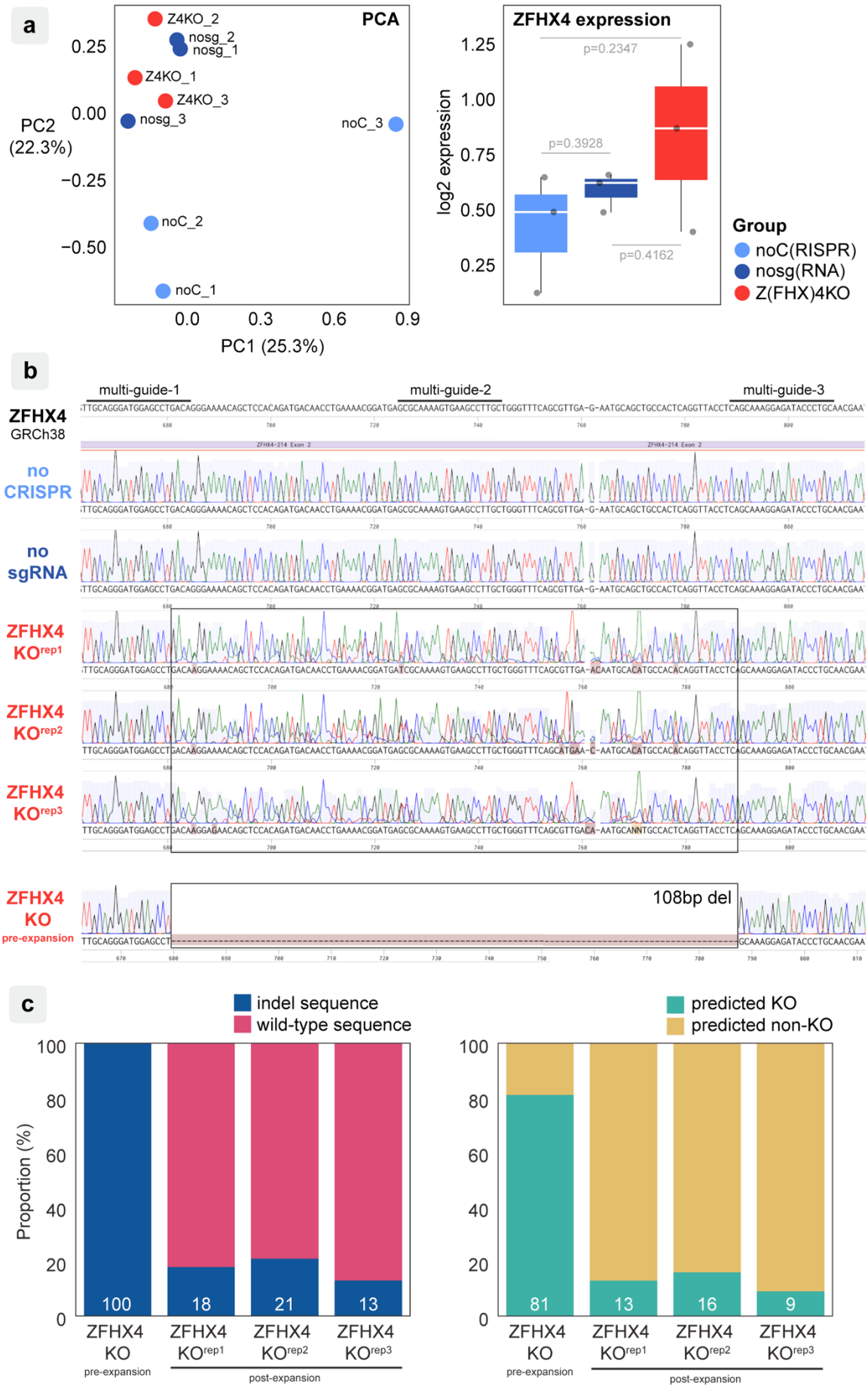
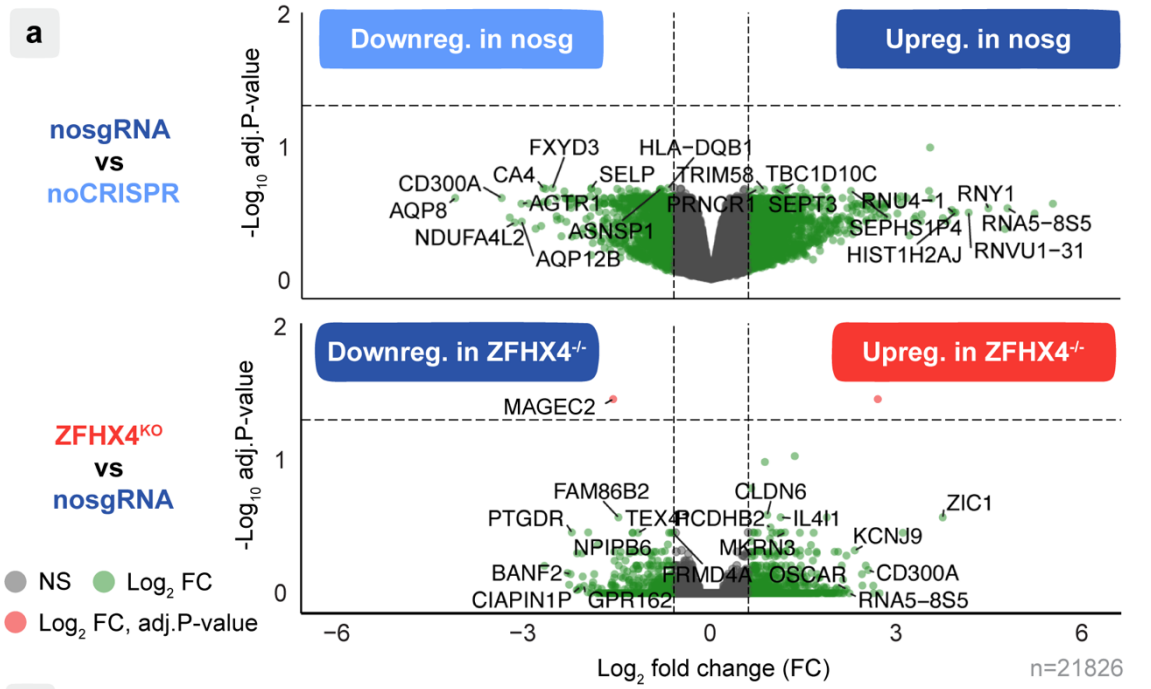

a
PCA
Z4KO_2
nosg_2
nosg_1
Z4KO_1
Z4KO_3
nosg_3
noC_3
noC_2
noC_1
PC2 (22.3%)
PC1 (25.3%)
ZFHX4 expression
log2 expression
p=0.2347
p=0.3928
p=0.4162
Group
noC(RISPR)
nosg(RNA)
Z(FHX)4KO
b
multi-guide-1
multi-guide-2
multi-guide-3
ZFHX4
GRCh38
no CRISPR
no sgRNA
ZFHX4 KOrep1
ZFHX4 KOrep2
ZFHX4 KOrep3
ZFHX4 KO pre-expansion
108bp del
c
indel sequence
wild-type sequence
predicted KO
predicted non-KO
Proportion (%)
100
18
21
13
81
13
16
9
ZFHX4 KO
ZFHX4 KOrep1
ZFHX4 KOrep2
ZFHX4 KOrep3
pre-expansion
post-expansion

Figure 4.11 | RNAseq and further sanger sequencing results of *ZFHX4*$^{-/-}$ metastatic organoids. **(A)** PCA plot of three replicates per condition indicated with a _ suffix and colour coded by the key. ZFHX4 expression was plotted with boxplots representing medians, overlaying dots of specific outlier values. P-values in grey were produced following unpaired t-tests. **(B)** Trace files of post-expansion sanger sequencing results aligned to the ZFHX4 Exon 2 region on Benchling software. The pre-expansion ZFHX4^{KO} validation trace has been added for comparison. **(C)** Synthego-ICE results of proportions of indel vs wild-type estimations as well as KO and non-KO predictions for both pre-expansion and post-expansion samples. Values overlayed in white represent % of indel or KO according to plot.

Despite a low KO efficiency, DEG analysis was still performed as fold change differences could still indicate a potential phenotype irrespective of significance. To elucidate potential electroporation/Cas9-related effects, nosgRNA samples were compared with noCRISPR samples and revealed diverse transcriptomic changes (Figure 4.12A/B, Supplementary Table 4.1), highlighting the importance of employing nosgRNA controls for accurate KO comparisons. To elucidate potential *ZFHX4*-related gene expression effects, *ZFHX4*$^{-/-}$ samples were compared with nosgRNA samples, revealing 1 significantly upregulated gene in *ZFHX4*$^{-/-}$ (*LINC00871*) and 1 significantly downregulated gene in *ZFHX4*$^{-/-}$ (*MAGEC2*) (Figure 4.12A). Comparing upregulated and downregulated genes following t-statistic ranking (Supplementary Table 4.2) may be tedious considering the insufficient KO efficiency, thus pathway-level analysis was additionally performed with GSEA (Figure 4.12B). Upon *ZFHX4*$^{-/-}$, there was significant upregulation of EMT, intestinal development and adhesion-related pathways in comparison to a significant downregulation of metabolic and catabolic processes as well as DNA replication-related pathways. The flattening of the volcano plot in the *ZFHX4*$^{-/-}$ vs nosgRNA comparison reiterates gene expression differences due to electroporation and Cas9 incorporation were greater than gene expression differences due to KO, likely reflecting insufficient KO efficiency.


a
nosgRNA
vs
noCRISPR
Downreg. in nosg
Upreg. in nosg
-Log10 adj.P-value
FXYD3
HLA−DQB1
CA4
SELP
TRIM58
TBC1D10C
CD300A
AQP8
AGTR1
PRNCR1
SEPT3
RNU4−1
RNY1
SEPHS1P4
RNA5−8S5
NDUFA4L2
ASNSP1
HIST1H2AJ
RNVU1−31
AQP12B
ZFHX4KO
vs
nosgRNA
Downreg. in ZFHX4-/-
Upreg. in ZFHX4-/-
MAGEC2
FAM86B2
CLDN6
ZIC1
PTGDR
TEX41
CDHB2
IL4I1
KCNJ9
NPIPB6
MKRN3
BANF2
FRMD4A
OSCAR
CD300A
CIAPIN1P
GPR162
RNA5−8S5
NS
Log2 FC
Log2 FC, adj.P-value
Log2 fold change (FC)
n=21826

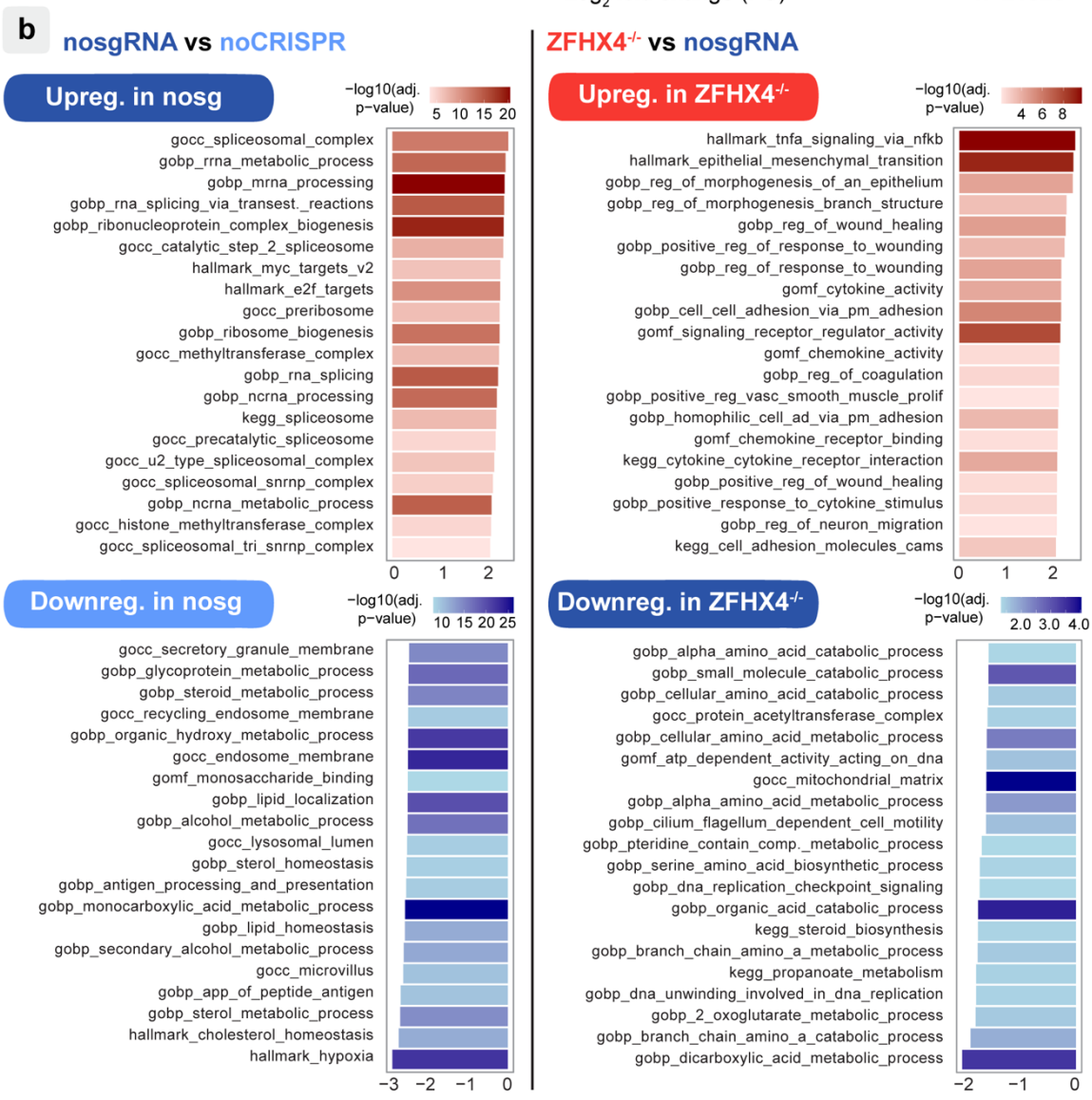

b
nosgRNA vs noCRISPR
ZFHX4-/- vs nosgRNA
Upreg. in nosg
−log10(adj. p−value)
gocc_spliceosomal_complex
gobp_rrna_metabolic_process
gobp_mrna_processing
gobp_rna_splicing_via_transest._reactions
gobp_ribonucleoprotein_complex_biogenesis
gocc_catalytic_step_2_spliceosome
hallmark_myc_targets_v2
hallmark_e2f_targets
gocc_preribosome
gobp_ribosome_biogenesis
gocc_methyltransferase_complex
gobp_rna_splicing
gobp_ncrna_processing
kegg_spliceosome
gocc_precatalytic_spliceosome
gocc_u2_type_spliceosomal_complex
gocc_spliceosomal_snrnp_complex
gobp_ncrna_metabolic_process
gocc_histone_methyltransferase_complex
gocc_spliceosomal_tri_snrnp_complex
Upreg. in ZFHX4-/-
hallmark_tnfa_signaling_via_nfkb
hallmark_epithelial_mesenchymal_transition
gobp_reg_of_morphogenesis_of_an_epithelium
gobp_reg_of_morphogenesis_branch_structure
gobp_reg_of_wound_healing
gobp_positive_reg_of_response_to_wounding
gobp_reg_of_response_to_wounding
gomf_cytokine_activity
gobp_cell_cell_adhesion_via_pm_adhesion
gomf_signaling_receptor_regulator_activity
gomf_chemokine_activity
gobp_reg_of_coagulation
gobp_positive_reg_vasc_smooth_muscle_prolif
gobp_homophilic_cell_ad_via_pm_adhesion
gomf_chemokine_receptor_binding
kegg_cytokine_cytokine_receptor_interaction
gobp_positive_reg_of_wound_healing
gobp_positive_response_to_cytokine_stimulus
gobp_reg_of_neuron_migration
kegg_cell_adhesion_molecules_cams
Downreg. in nosg
gocc_secretory_granule_membrane
gobp_glycoprotein_metabolic_process
gobp_steroid_metabolic_process
gocc_recycling_endosome_membrane
gobp_organic_hydroxy_metabolic_process
gocc_endosome_membrane
gomf_monosaccharide_binding
gobp_lipid_localization
gobp_alcohol_metabolic_process
gocc_lysosomal_lumen
gobp_sterol_homeostasis
gobp_antigen_processing_and_presentation
gobp_monocarboxylic_acid_metabolic_process
gobp_lipid_homeostasis
gobp_secondary_alcohol_metabolic_process
gocc_microvillus
gobp_app_of_peptide_antigen
gobp_sterol_metabolic_process
hallmark_cholesterol_homeostasis
hallmark_hypoxia
Downreg. in ZFHX4-/-
gobp_alpha_amino_acid_catabolic_process
gobp_small_molecule_catabolic_process
gobp_cellular_amino_acid_catabolic_process
gocc_protein_acetyltransferase_complex
gobp_cellular_amino_acid_metabolic_process
gomf_atp_dependent_activity_acting_on_dna
gocc_mitochondrial_matrix
gobp_alpha_amino_acid_metabolic_process
gobp_cilium_flagellum_dependent_cell_motility
gobp_pteridine_contain_comp._metabolic_process
gobp_serine_amino_acid_biosynthetic_process
gobp_dna_replication_checkpoint_signaling
gobp_organic_acid_catabolic_process
kegg_steroid_biosynthesis
gobp_branch_chain_amino_a_metabolic_process
kegg_propanoate_metabolism
gobp_dna_unwinding_involved_in_dna_replication
gobp_2_oxoglutarate_metabolic_process
gobp_branch_chain_amino_a_catabolic_process
gobp_dicarboxylic_acid_metabolic_process

Figure 4.12 | Differential gene and pathway analysis of ZFHX4$^{-/-}$ metastatic organoids. (A) Volcano plots demonstrating both nosgRNA vs noCRISPR and ZFHX4$^{-/-}$ vs nosgRNA comparisons. Colours are representative of comparison and group. **(B)** Bar plots demonstrating normalised enrichment scores (NES) of pathway enrichment following GSEA on the differential genes obtained from each comparison ranked by t-statistic. Gene sets were obtained from msigdbr and included hallmark, biocarta, kegg, go-bp, go-mf and go-cc subsets. Significant pathways were selected and then ranked by NES. Each bar colour intensity refers to significance of adj.p-value.

One disadvantage to the *ZFHX4*$^{-/-}$ vs nosgRNA comparison is that it assumes a complete knockout, and so with low editing efficacy the effect on expression is diluted which is why fold changes are present but are not statistically significant. One way around this is to use the INDELscore (i.e. % non-wild-type sequence) derived from the sanger sequencing results of each replicate as a continuous covariate i.e. model gene expression as a linear function of editing % ($model.matrix(\sim INDELscore, data = meta)$). This approach will capture dose-dependent changes, increasing sensitivity to detect subtle transcriptional effects and ultimately identify genes whom expression correlates with % editing efficiency. Performing limma-voom differential gene expression between *ZFHX4*$^{-/-}$ and nosgRNA using this new design (noCRISPR replicates removed), log fold-change (FC) output represented the estimated change in log2(expression) per 1% increase in INDELscore, and the top 20 positive and negatively correlated genes with % editing efficiency were ranked by t-statistic (Table 4.4). Positive logFC indicates a gene's expression increases as INDELscore increases and vice versa.

Table 4.4 | Top 20 positively and negatively correlated genes with % editing efficiency, potentially highlighting genes regulated by ZFHX4 in a dose-dependent manner. Genes in bold and highlighted in green were also identified in the top 30 upregulated and downregulated genes respectively in the ZFHX4^{KO} vs nosgRNA comparison in Supplementary Table 4.2.

Top 20 Positive						
Gene	**t-stat.**	**logFC**	**adj.p-val**	**Z4KO_1FC**	**Z4KO_2FC**	**Z4KO_3FC**
CXCL8	7.96	0.04	0.061	1.68	1.84	1.46
FGF9	6.34	0.04	0.200	1.75	1.92	1.50
TNFAIP2	6.22	0.03	0.200	1.50	1.60	1.34
LCT	5.96	0.05	0.201	1.91	2.12	1.59
NKD1	5.70	0.02	0.258	1.23	1.28	1.16
KRT17	5.58	0.05	0.283	1.81	2.00	1.53
NES	5.42	0.02	0.283	1.26	1.31	1.18
MMP2	5.40	0.04	0.283	1.57	1.69	1.38
CXCL5	5.36	0.06	0.487	2.06	2.33	1.69
LBH	5.25	0.04	0.326	1.71	1.87	1.48

NAV3	5.25	0.06	0.309	2.24	2.57	1.79
H19	5.23	0.04	0.358	1.72	1.88	1.48
WIF1	5.21	0.02	0.314	1.30	1.36	1.21
IGDCC3	5.08	0.04	0.326	1.55	1.67	1.37
CDH6	5.01	0.04	0.334	1.56	1.68	1.38
HEY2	5.00	0.04	0.334	1.55	1.67	1.37
NPR3	4.92	0.08	0.334	2.72	3.21	2.06
ZIC1	4.89	0.19	0.334	10.42	15.40	5.43
MKRN3	4.86	0.07	0.339	2.36	2.72	1.86
PPDPF	4.81	0.02	0.347	1.24	1.28	1.16
Top 20 Negative						
Gene	**t-stat.**	**logFC**	**adj.p-val**	**Z4KO_1FC**	**Z4KO_2FC**	**Z4KO_3FC**
MAGEC2	-8.89	-0.09	0.058	0.33	0.28	0.45
CPVL	-8.33	-0.03	0.058	0.69	0.65	0.76
PCSK9	-6.54	-0.03	0.200	0.73	0.69	0.80
DUSP27	-6.21	-0.07	0.200	0.42	0.36	0.54
CDCA7L	-6.06	-0.02	0.200	0.76	0.72	0.82
NT5E	-5.93	-0.02	0.201	0.75	0.71	0.81
PKDCC	-5.51	-0.02	0.283	0.80	0.77	0.85
PROM2	-5.37	-0.02	0.283	0.74	0.70	0.80
SOD3	-5.37	-0.03	0.283	0.66	0.61	0.74
KLF15	-5.34	-0.04	0.283	0.63	0.58	0.71
MARC2	-5.09	-0.03	0.326	0.66	0.62	0.74
PTPN7	-5.09	-0.07	0.326	0.40	0.35	0.52
PTGDR	-5.07	-0.13	0.326	0.20	0.15	0.31
DMD	-4.96	-0.03	0.334	0.72	0.68	0.79
SPON1	-4.94	-0.04	0.334	0.60	0.55	0.69
IDH2	-4.91	-0.02	0.334	0.80	0.78	0.85
NAV2	-4.84	-0.03	0.340	0.68	0.63	0.75
MCCC1	-4.75	-0.02	0.350	0.81	0.79	0.86
SCPEP1	-4.63	-0.02	0.375	0.74	0.71	0.81
ALDH1L1	-4.63	-0.03	0.375	0.67	0.62	0.75

The three last columns are with respect to each ZFHX4$^{-/-}$ replicate (18%, 21% and 13% editing efficiencies) where logFC has been multiplied by editing efficacy before inverting log2

For example, the top-ranked gene by t-statistic is *CXCL8* who's expression shows the strongest positive correlation with *ZFHX4* editing efficiency. The estimated $\log_2$FC per 1% *ZFHX4* editing was 0.04, indicating that a full 100% KO would be expected to increase expression by 4 log2 units. Within the mosaic replicates (18%, 21% and 13% editing respectively), the expected log2 changes corresponded to fold changes of 1.68, 1.84 and 1.46 respectively. These results suggest *CXCL8* expression increases in a dose-dependent manner with *ZFHX4* perturbations, although the small sample size and mosaic editing indicates this should be interpreted cautiously. Other interesting genes identified here were plotted to visualise expression and *ZFHX4* editing efficiency correlations (Figure 4.13).

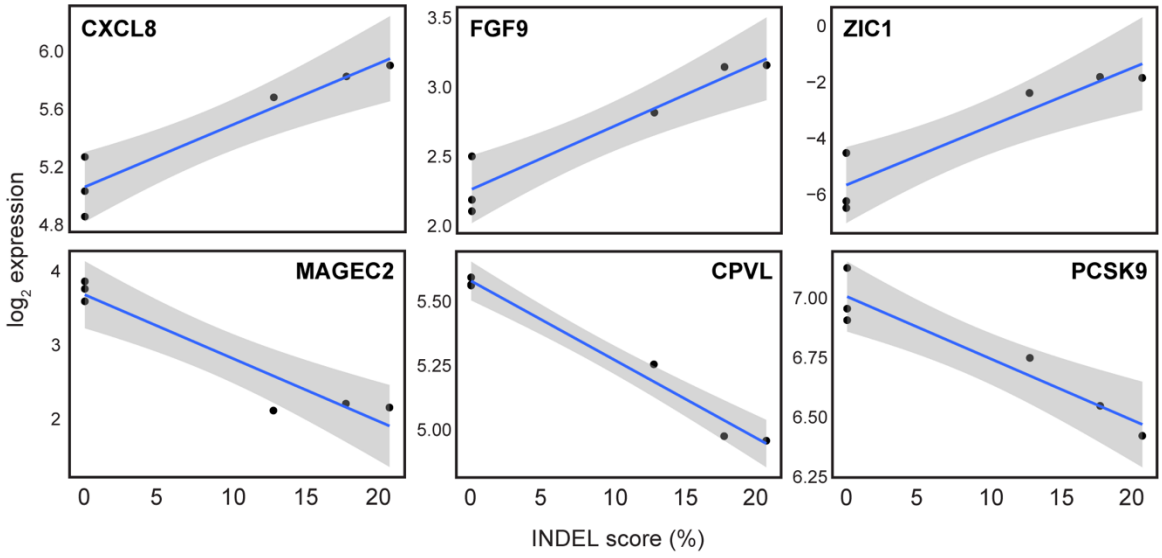


Figure 4.13 | Scatter plots depicting expression of top genes positively or negatively correlated with ZFHX4 editing. Blue lines represent linear regression lines. 0% indel score represents the three sgRNA replicates, and the dots to the right represent the ZFHX4^{KO} replicates (Z4KO_rep3 13%, Z4KO_rep1 18%, Z4KO_rep2 21%).

Furthermore, some of the top genes identified in the standard *ZFHX4*$^{-/-}$ vs nosgRNA comparison were also identified here providing further confidence and complementarity in the approach. Additionally, this approach identified other genes that were not as highly upregulated or downregulated in the standard *ZFHX4*$^{-/-}$ vs nosgRNA comparison suggesting a significant advantage of using both approaches to look into interesting genes. Those with expression positively correlating with *ZFHX4* editing efficiency comprised developmental and Wnt/Notch signalling (*NES*, *KRT17*, *NKD1*, *HEY2*, *ZIC1*, *LBH*, *H19*, *WIF1*), EMT-related factors (*CDH6*, *NAV3*) and ECM signalling and invasion (*MMP2*, *CXCL8*, *CXCL5*, *TNFAIP2*, *FGF9*) themes suggesting these could be negatively regulated by *ZFHX4*. Those with expression negatively correlating with *ZFHX4* editing efficiency comprised metabolic and oxidative stress-related (*IDH2*, *ALDH1L1*, *MCCC1*, *MARC2*, *SOD3*), tumour-immune signalling (*MAGEC2*, *NT5E*, *PTGDR*), ECM-related (*PROM2*, *SPON1*, *DMD*) and transcription (*CDCA7L*, *KLF15*) suggesting these could be positively regulated by *ZFHX4*, assuming that editing causes LoF.

To explore reasons for *ZFHX4*-WT selection over *ZFHX4*$^{-/-}$, senescence (SenMayo, n=91/125 genes) and apoptosis (Hallmark Apoptosis, n=149/183 genes) signatures were interrogated (Figure 4.14). *ZFHX4*$^{-/-}$ organoids demonstrated a significant enrichment of a senescent phenotype (FDR=1.44x10^{-9}) comparative to nosgRNA controls. Furthermore, a substantial enrichment in apoptosis was observed though this was just under significance following multiple testing correction (FDR=0.087).

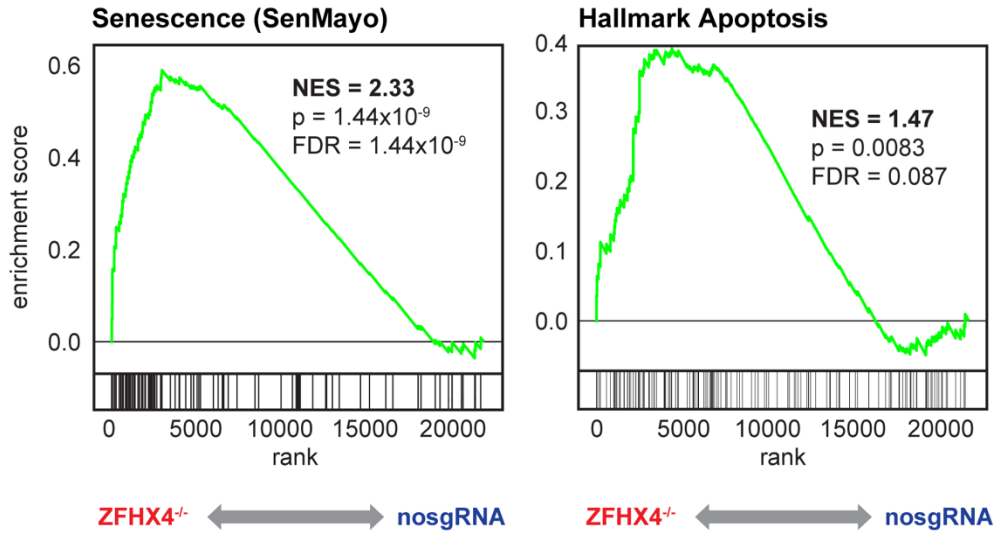


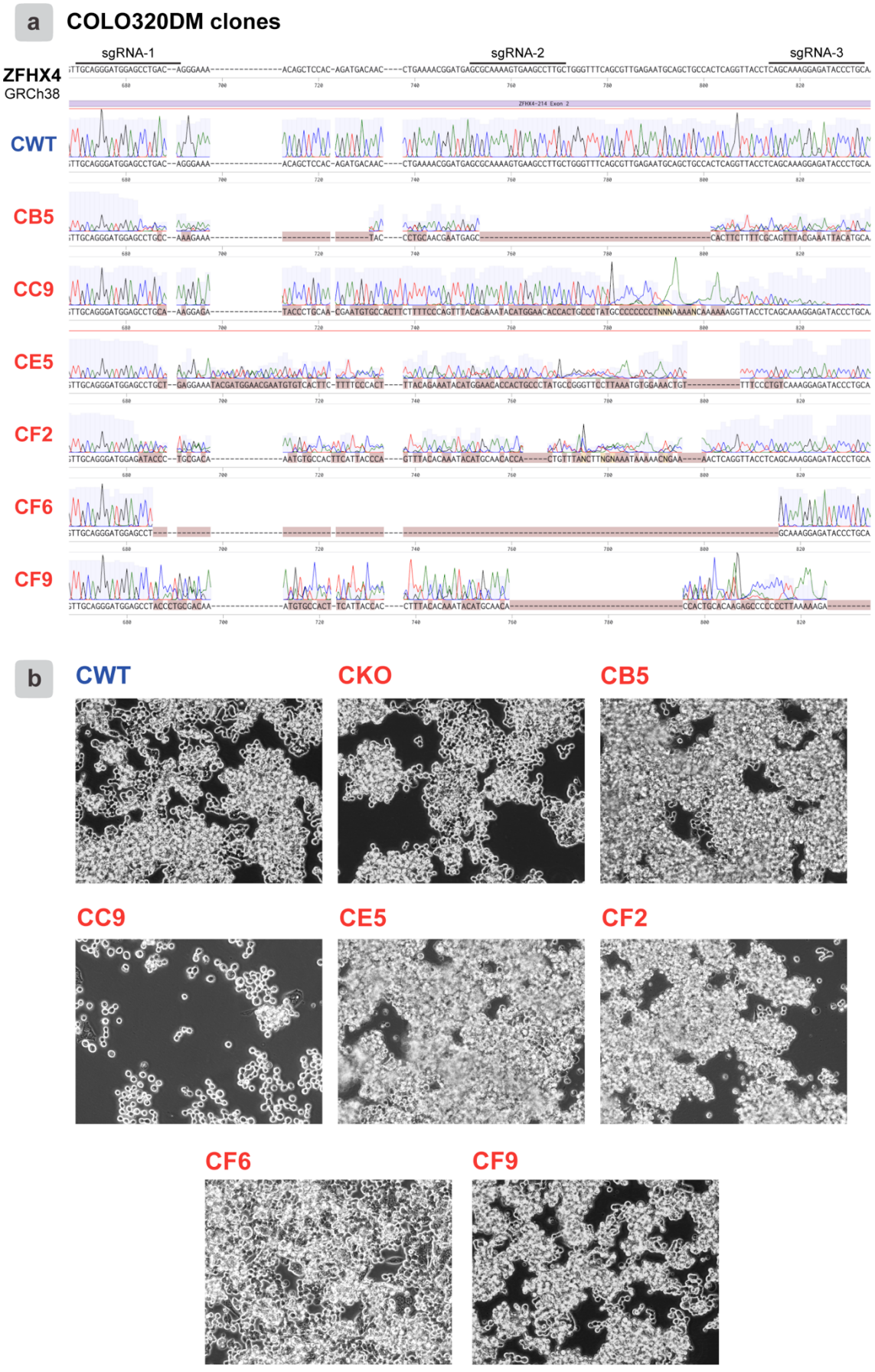
Figure 4.14 | Gene set enrichment analysis (GSEA) results of senescence and apoptosis in *ZFHX4*$^{-/-}$ organoids and nosgRNA controls. The SenMayo gene set was obtained from the literature (Saul et al., 2022) and the Hallmark Apoptosis gene set was obtained from MSigdbr. *[NES = normalised enrichment score, FDR = false discovery rate]*

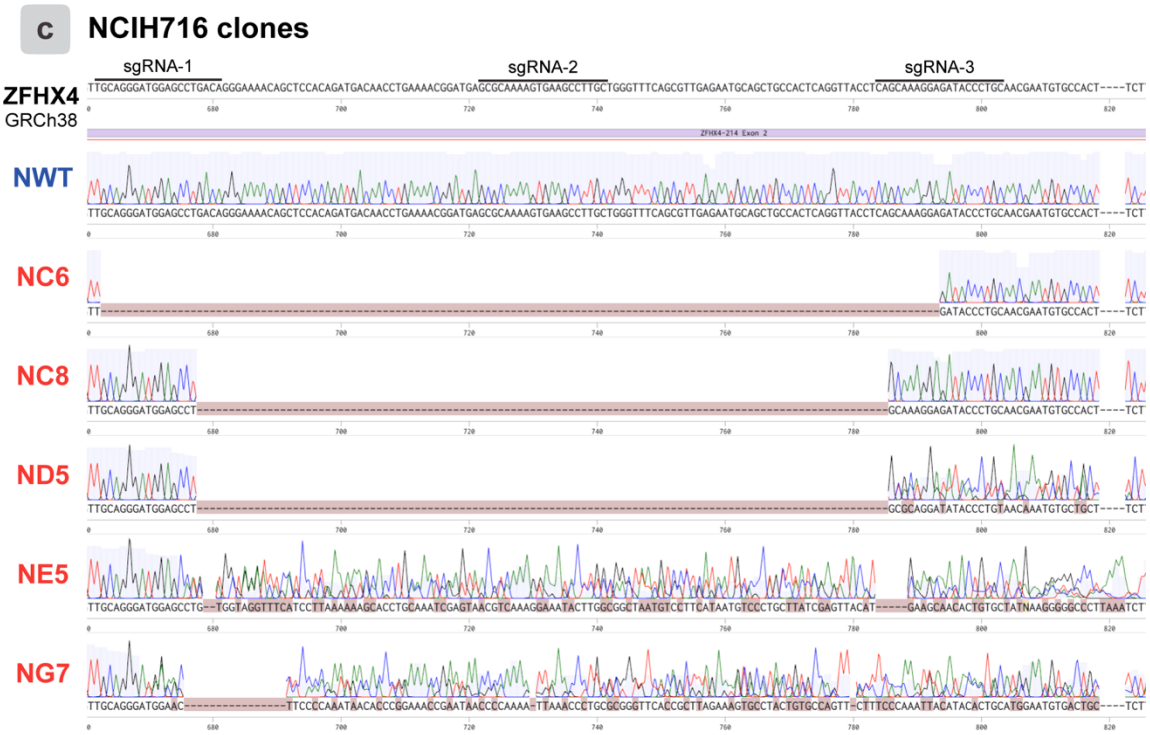
Taken together, these data suggest interesting hints into potential roles for ZFHX4 in the metastatic setting. However, further experimental repeats on more organoids from different patients and colon tissue types with higher CRISPR editing efficiencies are required to fully elucidate these roles in the colon organoid setting.

4.4.4 Genomic, transcriptomic and proteomic validation of *ZFHX4*$^{-/-}$ cell lines

4.4.4.1 Genomic validation of ZFHX4$^{-/-}$ *cell lines*

Polyclonal *ZFXH4*$^{-/-}$ lines were first established, validated by sanger sequencing and then single-cell cloned to generate monoclonal *ZFHX4*$^{-/-}$ (monoKO) lines (Figure 4.9). These were then characterised further at the gDNA level. CRISPR of SKNMC and NCIH716 lines was highly efficient, prompting mostly 108bp deletions representing >96% gDNA (Supplementary Figure S4.11). Due to large frameshifts, these were classed as predicted KOs and thus SKNMC and NCIH716 polyKOs were assumed to be *ZFHX4*$^{-/-}$. Conversely, editing efficiency for COLO320DM was much lower (67%). Whilst a disrupted trace is observed, ICE analysis again predicted all indels as 108bp deletions. As well as maintaining these three polyKO populations throughout all experiments, they were single-cell cloned with aims to obtain 100% monoKO populations. COLO320DM polyKO provided 16 clones and NCIH716 provided 20 clones of varying editing efficiencies. Based on ICE analysis, only those with 100% indels and 0% wild-type sequence were selected (n=6 and n=5 respectively). These have been named according to Figure 4.9 (KOs coloured by red) and all 'WT' lines represent nosgRNA populations (coloured by dark blue). All monoKOs demonstrated significant disruption, exhibiting insertions and deletions at the trace level comparative to WT (Figure 4.15A/C). However, most *ZFHX4*$^{-/-}$ clones displayed no overt differences in morphology or growth compared to WT (Figure 4.15B/D). Notable exceptions were CC9 and NG7 which were notoriously slow growing post-CRISPR, with the former demonstrating a sparse, isolated morphology. NG7's confluence in Figure 4.15D reflects extended culture time to obtain sufficient RNAseq material.


a
COLO320DM clones
sgRNA-1
sgRNA-2
sgRNA-3
ZFHX4
GRCh38
CWT
CB5
CC9
CE5
CF2
CF6
CF9
b
CWT
CKO
CB5
CC9
CE5
CF2
CF6
CF9


c
NCIH716 clones
sgRNA-1
sgRNA-2
sgRNA-3
ZFHX4
GRCh38
NWT
NC6
NC8
ND5
NE5
NG7

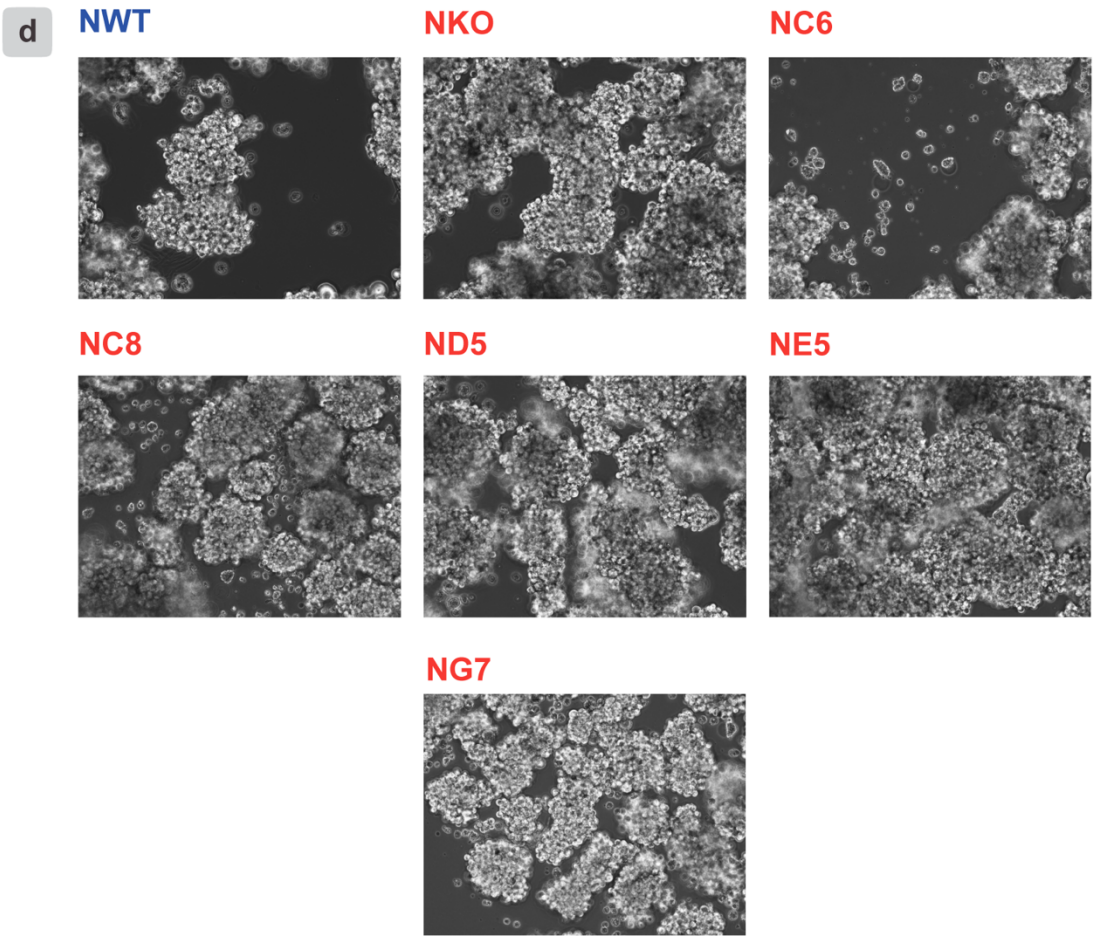

d
NWT
NKO
NC6
NC8
ND5
NE5
NG7

Figure 4.15 | Genomic validation of monoclonal *ZFHX4*$^{-/-}$ cell lines by sanger sequencing and their morphology. (A/C) Trace files of each cell line's sanger sequencing results were aligned to the ZFHX4 Exon 2 region on Benchling software. Wild-types (i.e. nosgRNA populations) are coloured in dark blue, and KO clones are coloured in red. **(B/D)** Microscope images at 20X magnification depict morphology of all clones, including the polyKO lines. These images are representative of a technical replicate that was submitted for RNAseq (2 month-old cells).

ICE analysis provided further quantitation on the types of most preponderant indels based upon statistical modelling of trace files (Table 4.5). The ZFHX4 multi-guide approach appears to preferentially generate 108bp deletions but is also capable of generating many other indel types of various sizes.

Table 4.5 | Predictions of KO scores and top 3 indel types by ICE analysis of polyKO and monoKO clones at month 0

Line	KO %	Non-KO %	Top 3 indel type distributions
SKO	100	0	100% 108bp del
CKO	67	33	67% 108bp del
CB5	87	13	24% 17bp del, 11% 65bp del, 11% 1bp ins
CC9	100	0	100% 108bp del
CE5	97	3	71% 108bp del, 5% 13bp ins, 4% 62bp del
CF2	98	2	58% 120bp del, 26% 1bp ins, 11% 120bp del
CF6	100	0	100% 108bp del
CF9	99	1	35% 2bp ins, 9% 10bp del, 6% 71bp del
NKO	96	4	96% 108bp del
NC6	100	0	45% 147bp del, 34% 147bp del, 13% 136bp del
NC8*	72	28	9% 1bp ins, 9% 25bp del, 8% 65bp del
ND5*	77	23	8% 1bp ins, 7% 73bp del, 7% 20bp ins
NE5	85	15	24% 2bp del, 18% 2bp del, 13% 63bp del
NG7	84	16	35% 17bp del, 33% 16bp del, 9% 15bp del

** denote poor model fit (<50%) so results need interpreting with caution, del = deletion, ins = insertion*

All KO populations were routinely tested for genetic drift, with predictions that presence of wild-type cells which are undetectable initially may eventually overtake. Furthermore, even though the single-cell cloning method should have selected 1 cell to begin with, by chance this may have been more and thus the indel type distributions were also compared across three different time points (month 0, month 2 and month 3). To note, all lines were consistently grown since the initial CRISPR which equated to a total of 11 months in culture, and references throughout this thesis to 'x'-m cells refer to cells of 'x'-months since CRISPR as a surrogate for age. This is an approximate time scale as cells were frozen down at multiple time points and reanimated later. A further validation check at these later time

points would be necessary, however data from the first three months could give clues into clonal behaviour. Over the first three months, monoKOs rarely varied in % WT sequence detected suggesting little genetic drift (Figure 4.16). However, % predicted KO did fluctuate in some clones, namely NC8, ND5, NE5 and NG7. Furthermore, the most drift observed pertained to the polyKO CKO, demonstrating a 1.8-fold increase in WT (indel rates of 67% to 38% by month 3) which was expected considering a low editing efficiency to start with.

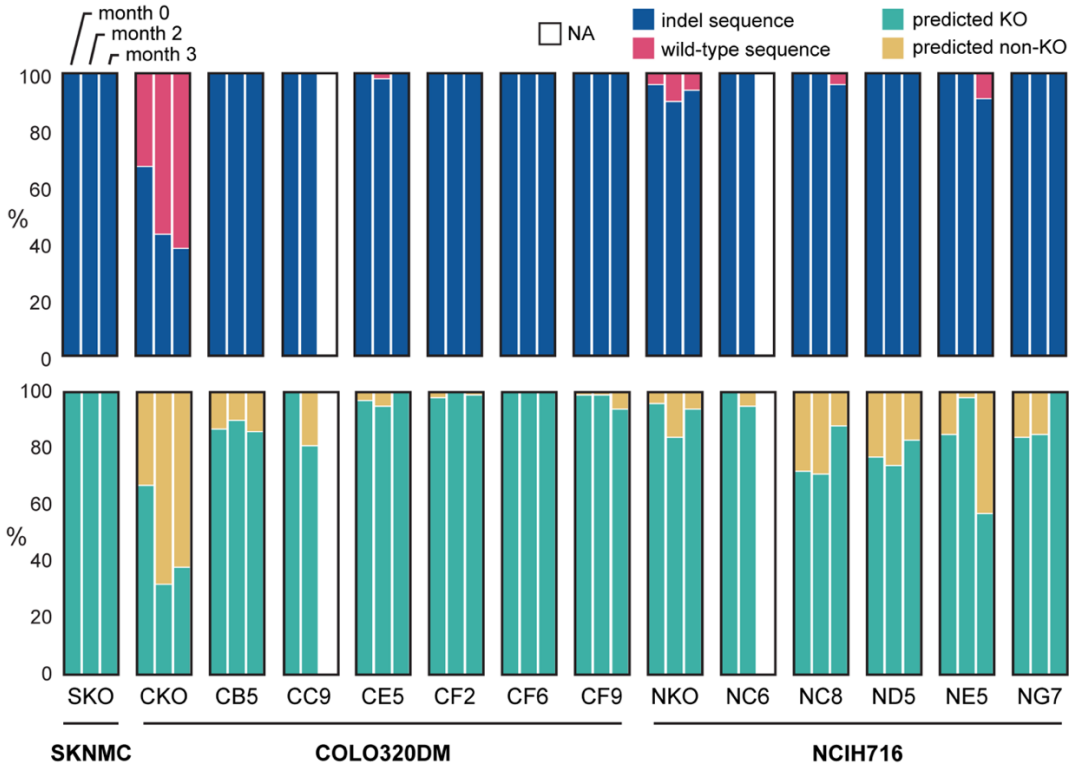


Figure 4.16 | Genomic validation of polyclonal and monoclonal *ZFHX4*$^{-/-}$ cell lines by sanger sequencing over the first three months of culture. Each individual bar represents a time point indicated by the key at the top, and three bars are grouped into respective cell lines. Bars are proportional, with the top representing indel vs wild-type, and bottom representing predicted KO vs predicted non-KO values produced following ICE analysis. White bars represent failure to fit a model by ICE on the trace.

Delving further into types of indel in each clone and how these differed over time, indel prediction by ICE was investigated across the three time points (Table 4.6). Clones which comprised indels that remained stable over three months included SKO (108bp del), CB5 17bp del), CE5 (108bp del/13bp ins), CF2 (120bp del), CF6 (108bp del), CF9 (2bp ins), NC8 (1bp ins), ND5 (1bp ins), NE5 (2bp del) and NG7 (16bp and 17bp del). Without further time points, it's difficult to predict how accurate these representations are in terms of

drawing interpretation on selection or if some are artifacts of modelling by ICE. For example, in 2m NE5 cells, a 63bp del was detected in 43% of sequences, however at month 3 this deletion represented 4% of sequences, hinting at potential negative selection which may have progressed over time.

Table 4.6 | Predictions of KO scores and top 3 indel types by ICE analysis of polyKO and monoKO clones over time

Line	Top 3 indel type distributions			Most represented indels for each clone
	Month 0	Month 2*	Month 3	
SKO	**100% 108bp del**	**100% 108bp del**	**100% 108bp del**	108bp del
CKO	40% 108bp del 20% 108bp del 7% 108bp del	7% 15bp ins 7% 16bp del 6% 32bp del	31% 108bp del 7% 108bp del	na
CB5	**24% 17bp del** 11% 65bp del 11% 1bp ins	**27% 17bp del** 25% 44bp del 8% 25bp del	**25% 17bp del** 8% 17bp del 7% 11bp ins	17bp del
CC9	99% 108bp del 1% 108bp del	13% 1bp ins 13% 16bp del 12% 32bp del	NA	na
CE5	**71% 108bp del** **5% 13bp ins** 4% 62bp del	**72% 108bp del** **4% 13bp ins** 3% 63bp del	**71% 108bp del** 7% 62bp del **4% 13bp ins**	108bp del and 13bp ins
CF2	**58% 120bp del** 26% 1bp ins **11% 120bp del**	**58% 120bp del** 25% 1bp ins **9% 120bp del**	**61% 120bp del** 25% 1bp ins **6% 120bp del**	120bp del
CF6	**100% 108bp del**	**100% 108bp del**	**100% 108bp del**	108bp del
CF9	**35% 2bp ins** 9% 10bp del 6% 71bp del	**33% 2bp ins** 12% 10bp del 12% 8bp del	**31% 2bp ins** 11% 8bp del 8% 11bp ins	2bp ins
NKO	89% 108bp del 5% 108bp del 1% 108bp del	40% 108bp del 23% 108bp del 12% 2bp del	56% 124bp del 23% 2bp del 10% 124bp del	na
NC6	45% 147bp del 34% 147bp del 13% 136bp del	16% 80bp del 12% 16bp ins 11% 63bp del	NA	na
NC8*	**9% 1bp ins** 9% 25bp del 8% 65bp del	9% 65bp del **9% 1bp ins** 9% 25bp del	16% 16bp del **12% 1bp ins** 10% 63bp del	1bp ins
ND5*	**8% 1bp ins** 7% 73bp del 7% 20bp ins	**10% 1bp ins** 8% 73bp del 8% 24bp del	13% 16bp del **12% 1bp ins** 10% 12bp ins	1bp ins
NE5	**24% 2bp del** **18% 2bp del** 13% 63bp del	43% 63bp del **22% 2bp del** 9% 2bp del	**16% 2bp del** 14% 12bp ins 14% 1bp del	2bp del
NG7	**35% 17bp del** **33% 16bp del** 9% 15bp del	**33% 17bp del** 21% 45bp del **18% 16bp del**	**38% 17bp del** **34% 16bp del** **7% 16bp del**	16bp and 17bp del

** month 2 is around the same time RNAseq was performed. Red text indicates that the model fit was <0.50 and might not be accurate. Bold indicates indel types detected in all timepoints.*

4.4.4.2 Transcriptomic validation of ZFHX4$^{-/-}$ cell lines

For additional clonal characterisation it was of interest to see if these mutations affected *ZFHX4* gene expression through RNAseq. Here, gene expression was quantified by gene-level counts, specifically STAR which counts all reads mapping any exon of *ZFHX4*. Three biological replicates of WT and polyKOs grown separately were submitted, alongside one replicate for each monoKO to represent a further 6 biological replicates for *ZFHX4*$^{-/-}$. One NCIH716 clone (NG8) was included in the original analyses and RNAseq however this clone eventually died in culture which appeared to be extremely slow growing. NG8 was originally validated with a 100% KO of 14bp deletion and kept in the RNAseq analysis to see if anything could be unpicked regarding a 14bp deletion.

Interestingly, every polyKO group irrespective of cell line had higher *ZFHX4* gene expression comparative to its respective WT, though this wasn't significant following non-parametric wilcoxon testing (Figure 4.17A). Furthermore, monoKOs demonstrated varied *ZFHX4* expression changes as expected with different mutation profiles. However, most monoKOs didn't show significant differences in *ZFHX4* expression relative to WT, except for CC9 and NG7 which appeared to have dramatically 6.14-fold and 1.18-fold reduced *ZFHX4* gene expression respectively. In summary, despite evidence at the gDNA level suggesting all monoKOs should have lost *ZFHX4* gene expression, CC9 appeared the only clone demonstrating bona-fide *ZFHX4*$^{-/-}$ at the mRNA level.

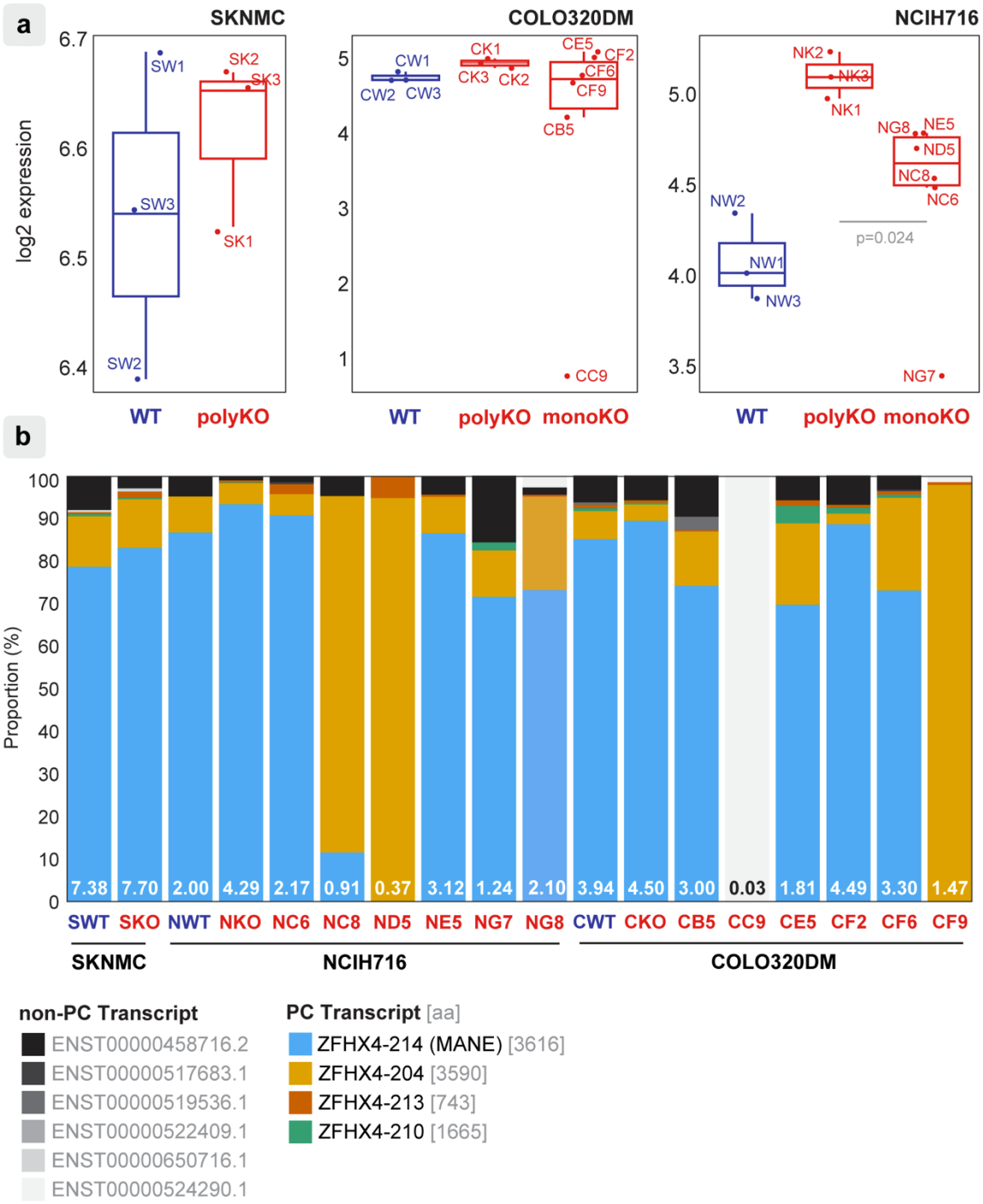


Figure 4.17 | RNAseq data demonstrating ZFHX4 expression in WT, polyKO and monoKO cell lines. (A) Derived from gene-level count-based expression quantification, box plots represent log2 expression of ZFHX4. Individual replicates are additionally labelled. Note there are differences with each y-axis. The p-value represents the only significant pairwise comparison by non-parametric wilcoxon testing. **(B)** Derived from transcript-level estimation-based expression quantification, proportional bar charts depict contributions of different non-protein coding (PC) and protein-coding (PC) transcript isoforms of ZFHX4. Values at the bottom of each bar represent the summed TPM expression from all isoforms as a metric of isoform-collapsed ZFHX4 expression. All WT and polyKO proportions were averaged to give a single representative bar. *[aa = amino acid]*

Additionally, we explored the expression of mRNA isoforms to see if these were different across different *ZFHX4*$^{-/-}$ backgrounds. RNAseq data was quantified by means of estimation-based method StringTie that looks at read depth distribution along exon-intron boundaries to calculate expression from specific mRNA isoforms or variants, typically in TPM format. Interestingly, the relative proportions of *ZFHX4* transcripts differed over varying *ZFHX4*$^{-/-}$ backgrounds (Figure 4.17B). For example, the slight increase in gene-level *ZFHX4* expression observed in polyKOs appears to result from increased expression of the MANE transcript, contributing an additional 4.27–6.66% relative to WT across the three cell lines. This is consistent with the primary use of MANE to explain the majority of a gene's expression. One reason for the increase in polyKOs could be a potential feedback mechanism if ZFHX4 regulates its own promoter.

monoKO clones that appeared to lose the majority if not all MANE transcript expression included NC8, ND5, CC9 and CF9 presenting a reduction in isoform-collapsed expression by 2.2-, 5.41-, 131- 2.68-fold respectively comparative to their averaged WT. Clearly the absence of ZFHX4 expression in CC9 is reflected in both gene-level and transcript-level quantifications, however interestingly the loss of MANE transcript and reduced isoform-collapsed expression in NC8, ND5 and CF9 is discordant with gene-level quantifications where they appear to have the same if not higher expression of ZFHX4 (Figure 4.17A). This emphasises the importance of investigating *ZFHX4* mRNA expression using both quantification metrics and also implies a potential phenomenon of degraded *ZFHX4* transcripts leading to an artificially-inflated expression in gene-level quantification despite a reduction in functional isoforms. Given the large size of *ZFHX4* and its isoforms, transcript-level quantification provides a more accurate measure for predicting populations in which *ZFHX4* could be functional or non-functional, as CRISPR-induced gDNA disruptions could produce large partial transcripts that escape NMD and remain detected at the gene level.

When taking the data from Figure 4.17B and plotting isoform-collapsed expression from either all isoforms or protein-coding isoforms only, there was an observable separation of

monoKOs that was not evident when observing gene-level *ZFHX4* expression as previously discussed (Figure 4.18). Furthermore, this separation appeared to roughly relate to both indel size (i.e. of the most common) and the likelihood of frameshift mutations (i.e. not a multiple of 3), as inferred from month 2 Sanger sequencing (Table 4.6) performed around the same time as RNA-seq. Formal testing across all polyKOs and monoKOs revealed a significant positive correlation between indel size and *ZFHX4* expression fold change relative to WT. This was true whether quantifying across all isoforms (Pearson r=0.335, p=0.037) or protein-coding isoforms only (r=0.343, p=0.033) (Supplementary Figure S4.12). Fold changes above 1 depict KOs which displayed greater transcript-level *ZFHX4* expression than their respective WT average and vice versa. Furthermore, KOs with higher expression than their WT counterparts were significantly more likely to harbour in-frame indels than frameshift indels, for both all isoforms (p=0.0308) and protein-coding isoforms (p=0.0310) based on unpaired t-testing. For example, all polyKOs irrespective of cell line were mostly comprised of large potentially-inframe indels and these consistently had the same if not higher expression than their WT. The designation of “potentially” in-frame refers to codon context, which cannot be fully resolved without precise mapping.

Together, these results suggest that KOs with small or frameshift indels tend to exhibit reduced *ZFHX4* expression at the mRNA level, whereas clones with larger in-frame indels often maintain comparable or even elevated transcript levels. While such transcripts may still be non-functional, for example due to altered protein folding or disrupted domains, the preservation of transcript structure likely explains their detection at the isoform level despite potential functional disruption at the protein level. This could be because *ZFHX4* is a large gene and generates large transcripts that can cope with changes such as these, for example an in-frame indel of 120bp may likely not disrupt much considering the MANE transcript is 13987bp in length, though codon context is required to fully predict this. After investigation at the gene and mRNA level, the gold-standard approach was to see whether ZFHX4 is affected at the protein level and thus western blotting was performed.

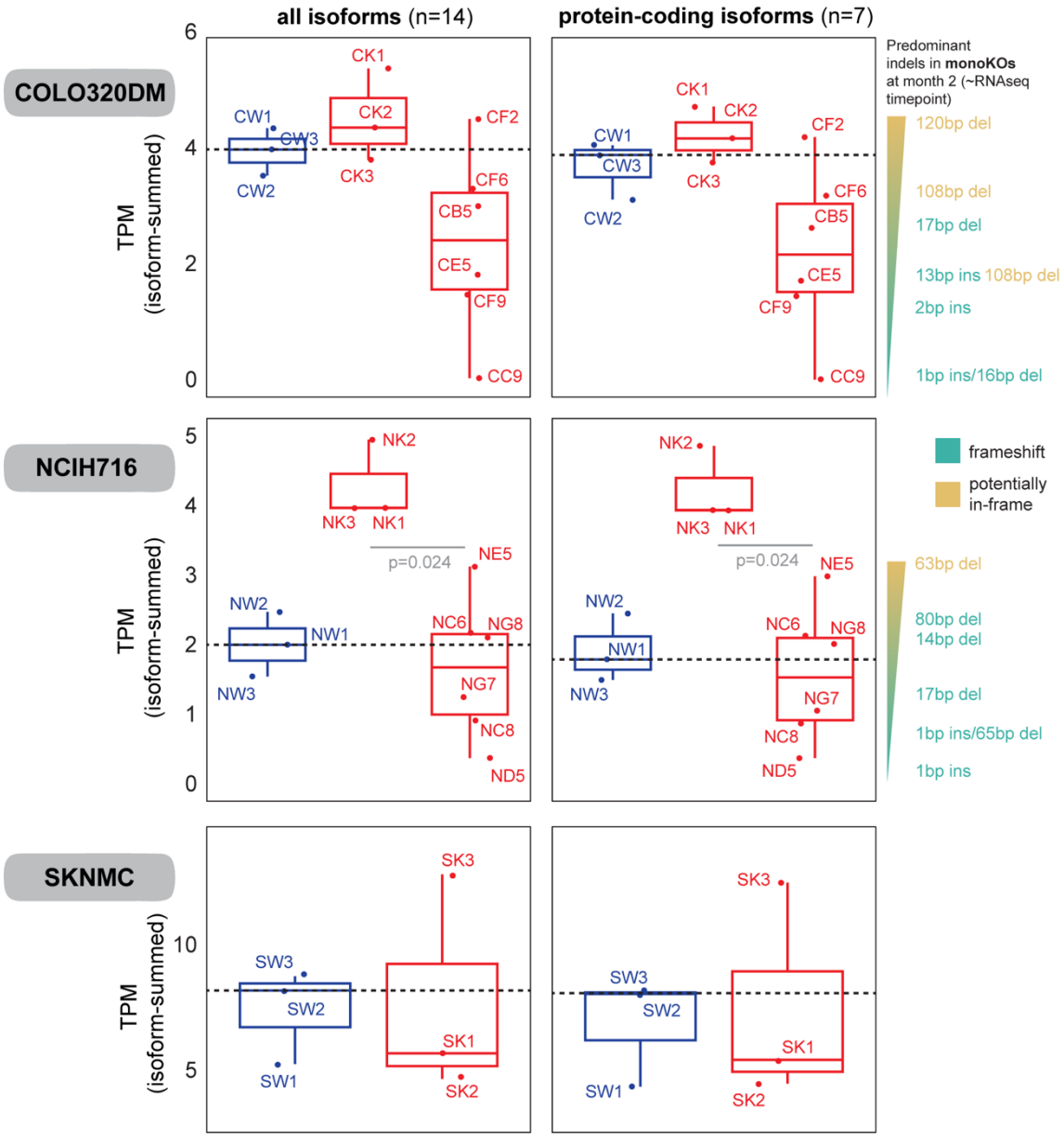


Figure 4.18 | Transcript-level ZFHX4 expression in WT, polyKO and monoKO cell lines referring to all isoforms or protein-coding isoforms only. Transcript-level expression in TPM was summed from either all isoforms (n=14) or protein-coding isoforms (n=7) of ZFHX4. RNAseq replicates are labelled. Annotation on the right of most common indel types in the respective monoKO at month 2 (a similar time to RNAseq). These are coloured by frameshift (not a multiple of 3) or potentially in-frame (a multiple of 3).

4.4.4.3 Proteomic validation of ZFHX4$^{-/-}$ cell lines

ZFHX4 protein is predicted to be ~400kDa in size and currently few publications have been able to demonstrate its expression. Considerable effort was devoted to optimising western blotting (WB) detection, employing high molecular weight (MW) techniques as existing ZFHX4 publications provide no specific methodological detail on WB procedures. WB was optimised on Jurkats and SKNMCs as negative and positive controls respectively as well as my colon lines of interest COLO320DM and NCIH716, all of which were noCRISPR origin. Vinculin (117kDa) was routinely used as the housekeeper comparative to smaller loading controls such as GAPDH to accommodate extended electrophoresis, allowing ZFHX4 to migrate further into the resolving gel and hopefully improve band separation and migration clarity. HSP60 was used additionally when testing Jurkats as they don't express Vinculin. Both HPA and M02 antibodies were tested alongside the HiMark high MW ladder.

Initially, WB consisted of 3-8% Tris-Acetate gels with protein transferred onto PVDF, blotting with the M02 primary as used by a previous publication (Chudnovsky et al., 2014). Testing two transfer methods (dry vs semi-dry), methanol percentage and sample heating, it was clear that a semi-dry 5% methanol transfer of lysate heated at 70°C for 10 minutes demonstrated the blot with most distinct and clean bands with reduced background (Figure 4.19A). Regardless of optimisation strategy, it was evident that ZFHX4 doesn't present as one discrete band, rather a spread of multiple immunoreactive species with majority staining higher than its predicted MW (i.e. >460kDa). Blotting intensity was consistent with predicted cell line expression (Figure 4.8) with absence in Jurkat and maximum in SKNMC lanes, supporting on-target reactivity by M02 but with incomplete specificity either towards different proteins, isoforms or proteolytic fragments. Despite a higher MW than expected, one prediction is that the top band is likely ZFHX4 which appears to degrade during sample heating at higher temperatures.

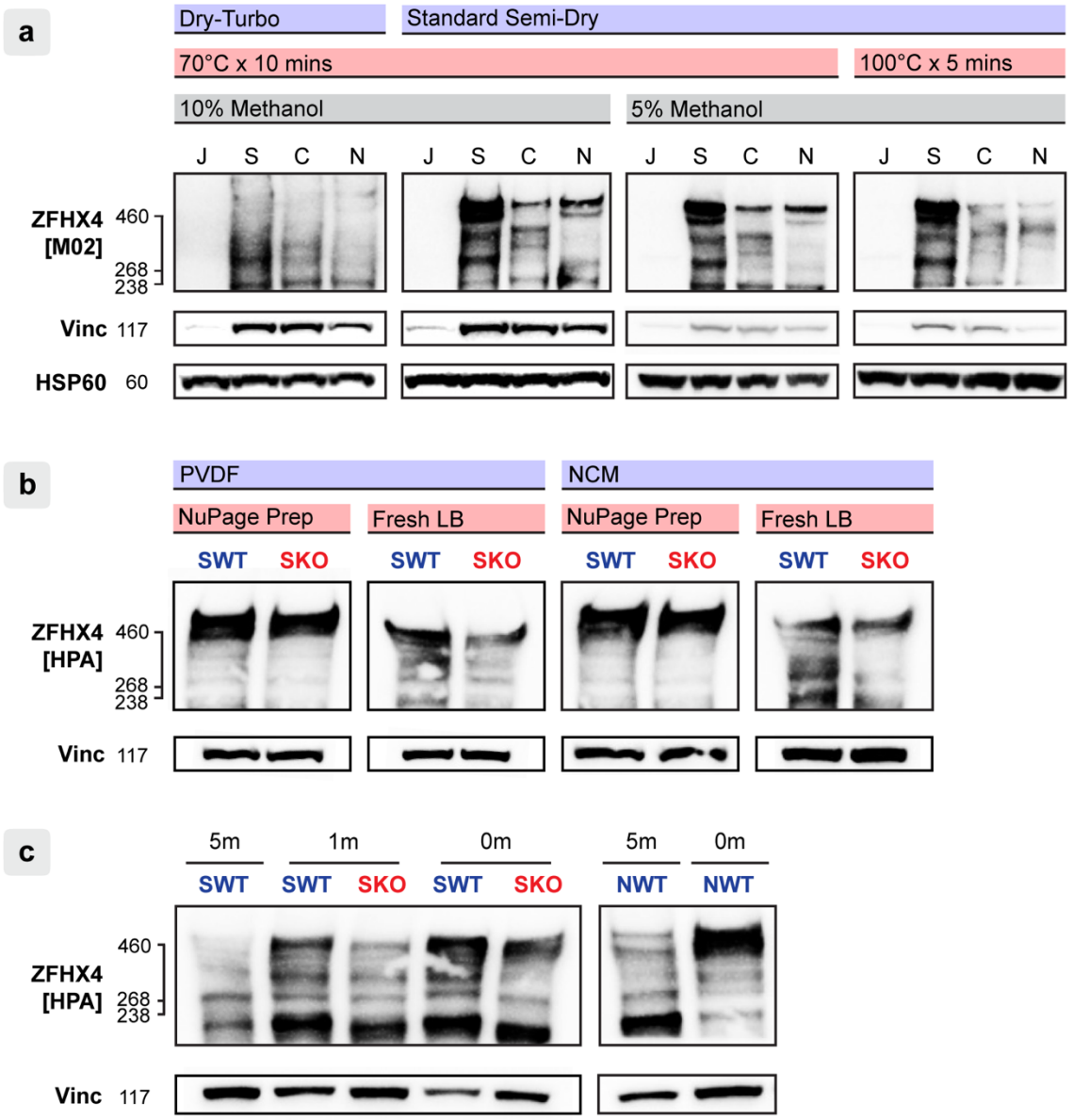


Figure 4.19 | Initial WB optimisation for ZFHX4 protein. (A) Transfer type 'Dry-Turbo' utilised fast-dry by BioRad Turbo (25V, 10min) and Semi-Dry was using the xCell Blotting Module (30V, 4°C, 24hr). Heating in red refers to sample preparation by sonication in NuPage format. Methanol % is in transfer buffer. Gels were 1mm. 1°-M02 Ab used at 1:500 (24hr, 4°C) and 2°-αM02 Ab used at 1:2000 (1hr, RT). Each black ZFHX4 square represents a different gel and the same 50µg lysate was used per sample. **(B)** Employing mostly protocol of (Perez Baca et al., 2025), 50µg of protein was lysed gently in RIPA (no sonication), prepared in respective loading buffers, heated at 95°C for 10 minutes and loaded into 1.5mm gels. Black ZFHX4 squares pertaining to PVDF or NCM represent lanes on the same gel. 1°-HPA Ab used at 1:2000 (24hr, 4°C) and 2°-αHPA Ab used at 1:10000 (1hr, RT). **(C)** Same method as B but by full-wet Biorad TransBlot method (30V 24hr) comparative to semi-dry xCell Blotting in B (100V 1hr15, 60V 2hr). Lysate obtained and frozen at -80°C at different timepoints were thawed. All housekeeper 1° and 2° Abs were used at 1:10000 (1hr, RT) and MW was annotated by running HiMark protein ladder in kDa. *[Vinc=Vinculin, PVDF=polyvinylidene fluoride, NCM=nitrocellulose membrane. LB=Laemmli buffer, RT=room temperature, J=JURKAT, S=SKNMC, C=COLO320DM, N=NCIH716, 5m=5 months old, 1m=1 month old, 0m=fresh]*

Following a publication and by personal communication, I gained access to a new ZFHX4-specific WB protocol and this became the basis of all further westerns (Perez Baca et al., 2025). Five main differences not tested previously were fast full-wet transfer (Biorad TransBlot), nitrocellulose membrane (NCM) over PVDF and thus a higher methanol percentage (20%), Laemmli buffer (LB)-based sample preparation over LDS-based NuPage, gentle RIPA lysis over sonication and the ZFHX4-HPA antibody at a more dilute concentration but supplementing blotting with a higher strength ECL (Thermo-Femto/Atto). Membrane type and sample preparation buffer were initially tested using SWT and SKO, predicting loss of a single band indicative of ZFHX4 (Figure 4.19B). Differences at the protein level between SWT and SKO were observed upon NCM and LB which still showed a top thick band but less banding between 268-460kDa, highlighting this method may reveal true ZFHX4 identity. Furthermore, in all WB experiments fresh lysates were used (i.e. processed fresh or frozen at sample-prep stage) because lysates older than 1 month demonstrated ZFHX4 degradation at the high MW end (Figure 4.19C).

Employing this new protocol, all clones were profiled for ZFHX4 protein expression using the HPA antibody (Figure 4.20). PolyKOs of all lines seemed to have slightly reduced banding intensity, though this is difficult to tell with slightly offset protein loading. Interestingly, most ZFHX4 monoKOs exhibited reduced or completely ablated protein expression but some clones appeared to have equal if not higher expression (NC8, NG7) than their respective WT. Though, this protein-level information didn't correlate with mRNA expression, for example NG7 displayed the lowest gene-level and reduced transcript-level quantified expression compared to NWT. This may exemplify potential complex feedback mechanisms linking *ZFHX4* mRNA regulation with protein translation, folding and expression. Also, this blot reiterates similar M02 observations of on-target reactivity but with incomplete specificity. Interestingly, an additional band is observed when a newer HPA antibody batch is used (see CWT vs CC9) suggesting further technical considerations when probing ZFHX4 at the protein level.

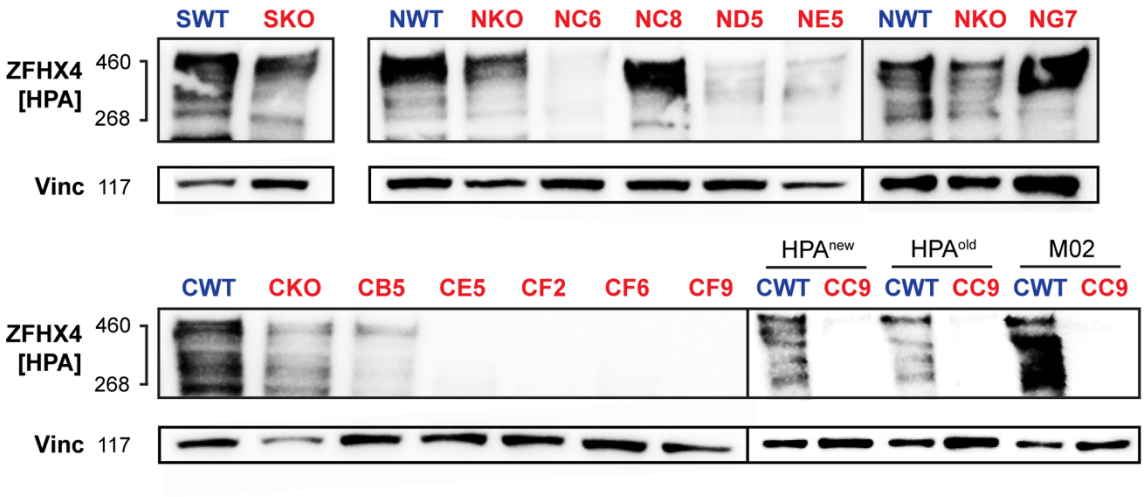


Figure 4.20 | Characterisation of ZFHX4 protein expression in WT, polyKO and monoKO lines using the HPA antibody. Following (Perez Baca et al., 2025) protocol, 1.5mm gels were transferred using the Biorad TransBlot (100V, 1hr30, 4°C) with 20% methanol. All lysates were 50µg apart from the ZFHX4 square for NG7 and CC9 where lysates were 10µg and 12µg respectively. The CC9 profiling was performed with a later passage of cells to allow for growth and enough protein to be extracted and included testing two batches of HPA antibody as well as the M02 antibody. Numbers represent kDa size by HiMark protein ladder. 1°-HPA Ab used at 1:2000 (24hr, 4°C) and 2°-αHPA Ab used at 1:10000 (1hr, RT). All lysates were extracted fresh from 3-month old cells post-CRISPR.

Across multiple optimisations and cell lines, ZFHX4 didn't resolve as a single band with HPA or M02. To test whether off-target bands were non-nuclear artefacts, I performed nuclear fractionation on all WT lines. Fractionation was clean, demonstrated by Vinculin exclusively in cytoplasmic fractions and Lamin B1 exclusively in nuclear fractions (Figure 4.21A). ZFHX4 signal was absent in cytoplasmic fractions and enriched in nuclear fractions compared to whole cell extracts with both antibodies, indicating that while multiple bands reflect incomplete specificity, they arise from nuclear proteins and likely represent ZFHX4 biology rather than technical noise.

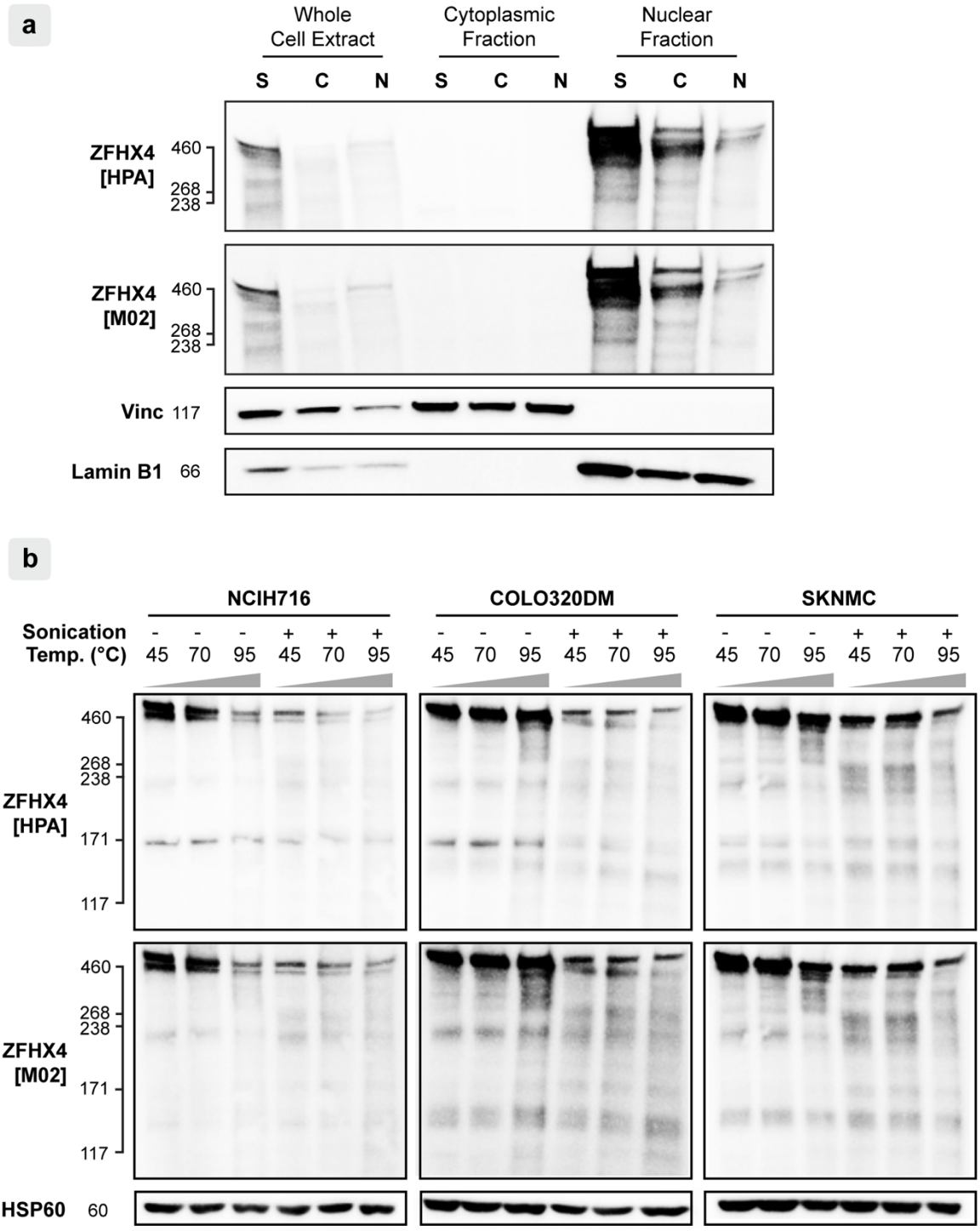


Figure 4.21 | Further ZFHX4 WB optimisation comprising nuclear fractionation and sample preparation adjustment experiments. (A) Nuclear fractionation was performed on WT SKNMC (S), COLO320DM (C) and NCIH716 (N) lysates. Blotting protocol followed (Perez Baca et al., 2025) but with longer transfer (30V, 20hr, 4°C) and 10% methanol. HPA was used first before membrane stripping and then M02 incubation. 1°-HPA/M02 Abs used at 1:2000 (24hr, 4°C) and 2°-αHPA/M02 Abs used at 1:10000 (1hr, RT). 1°-LaminB1 Ab used at 1:500 and 1°-Vinculin Ab used at 1:10000 (both 1hr, RT), with 2°s used at 1:10000 (1hr, RT). The image is representative of 1 gel. **(B)** One lysate was obtained for each cell line and then uniformly split to prepare with (+) or without (-) sonication and different heating temperatures. Each cell line was run on a separate gel, demarcated by ZFHX4 black squares. ZFHX4 probing method was identical to (A). Numbers represent kDa size by HiMark protein ladder.

As high heat appeared to alter ZFHX4 banding and promote proteolysis (Figure 4.19A), I varied lysate preparation and used longer electrophoresis for resolution (Figure 4.21B). Heating to 95°C increased low-MW bands and sonication reduced high-MW intensity, indicating lysate preparation affects ZFHX4 stability/degradation. Assuming the top band is ZFHX4 (often a doublet in NCIH716), I standardised preparation to 70 °C for 10 minutes to minimise artefactual bands and enable better WB resolution. Furthermore, the success of using two antibodies for ZFHX4 utilising a stripping approach provided extra confidence in changes at the protein level due to both being different in clonality, raised species and epitope binding (Figure 4.22).

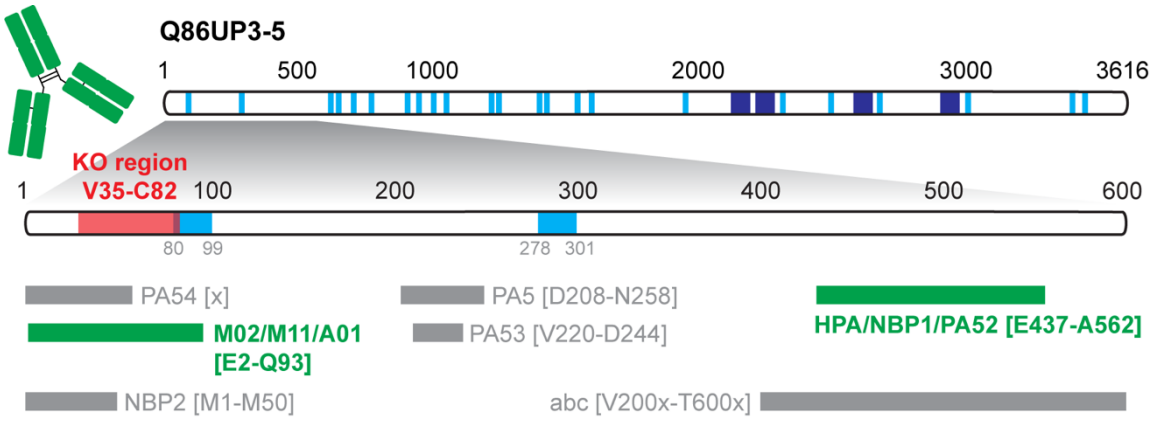


Figure 4.22 | Binding regions of all commercially available antibodies to human ZFHX4, with reference to the Q86UP3-5 isoform. Note all antibodies are N-terminal targeting. Square brackets denote epitope (i.e. sequence the antibody was raised against) and 'x' denotes unspecified by the supplier. Light blue regions are ZF motifs and dark blue regions are homeodomains. KO region displayed in red is the translated sequence spanning multi-guide RNA binding sites. The colour green denotes the two antibodies used for western blotting, HPA and M02. *[PA5=#PA5-99757, PA52=#PA5-54897, PA53=#PA5-67675, PA54=#PA5-69029, abc=#ab254654, A01=#H00079776-A01, NBP1=#NBP1-82156, NBP2=#NBP2-88615, HPA=#HPA023837, M02=#H00079776-M025E6, M11=#H00079776-M112B12]*

Therefore, all lines and clones were reprobed for ZFHX4 protein expression utilising this adapted approach and all previous WB approaches were collated and quantified (Figure 4.23). To note, HPA blots in Figure 4.23A represent 11m-old cells (i.e. lysate extracted fresh from cells 11 months in culture since CRISPR) comparative to HPA blots in Figure 4.20 which represented 3m-old cells (termed original-3m). Also, HPA blots in Figure 4.23B represent 3m-old cells which were thawed and cultured alongside 11m-old cells as an

additional passage replicate (termed current-3m). As expected, a lower temperature for sample heating demonstrated cleaner, crisper western blots on WT, polyKO and monoKO lines and 4 distinct bands irrespective of antibody in comparison to the initial attempt in Figure 4.20.

Within HPA blots, trends were consistent with original-3m cells where polyKOs had slightly reduced ZFHX4 expression and monoKOs showed dramatically reduced expression comparative to WTs. However, comparing HPA blots between original-3m, current-3m and 11m cells, there were some discrepancies in protein expression reflected by the large error bars in WB quantification (Figure 4.23C). For example, in CE5 and CF2 no expression was detected in original-3m cells but seems to reappear strongly in current-3m cells, with a slightly weaker but still apparent profile in 11m cells. This could be a reflection of clonal drift over time which can't be ascertained without sanger results in 11m cells, however I don't believe this is drift caused by an increase in WT sequence due to the explicit differences between HPA and M02 blotting explained below.

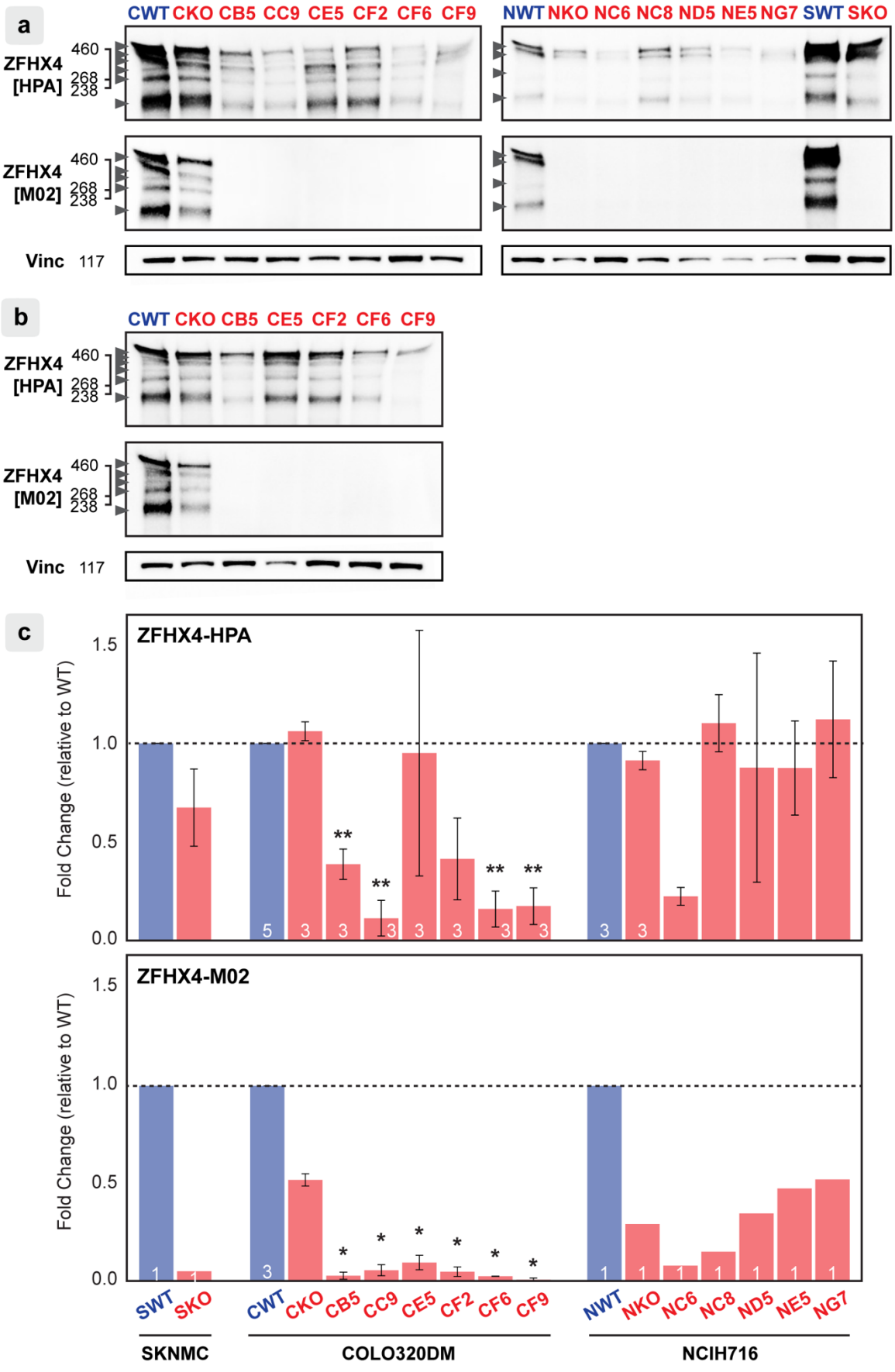

a
CWT CKO CB5 CC9 CE5 CF2 CF6 CF9
NWT NKO NC6 NC8 ND5 NE5 NG7 SWT SKO
ZFHX4 [HPA]
460
268
238
ZFHX4 [M02]
Vinc 117
b
CWT CKO CB5 CE5 CF2 CF6 CF9
c
ZFHX4-HPA
ZFHX4-M02
Fold Change (relative to WT)
1.5
1.0
0.5
0.0
SWT SKO
CWT CKO CB5 CC9 CE5 CF2 CF6 CF9
NWT NKO NC6 NC8 ND5 NE5 NG7
SKNMC
COLO320DM
NCIH716

Figure 4.23 | Quantification of ZFHX4 protein expression across all cell lines and replicates, including a newly-optimised WB approach. (A-B) Western blots of ZFHX4 expression by HPA and M02 antibodies of 11m cells (A) and current-3m cells (B). Blotting protocol followed (Perez Baca et al., 2025) but with a 70°C 10-minute sample heating, longer transfer (30V, 20hr, 4°C) and 10% methanol. HPA was used first before membrane stripping and then M02 incubation. 1°-HPA Ab (24hr, 4°C) and 1°-M02 Ab (8hr, 4°C) used at 1:2000 with 2°-αHPA/M02 Abs used at 1:10000 (1hr, RT). Each ZFHX4 square is a separate gel. Grey arrows indicate common bands. **(C)** Quantification of all western blots probing ZFHX4 in WT, polyKO and monoKO lines by HPA or M02 antibody. Fold change values represent integrated density normalised to background and housekeeper expression of KO divided by WT. Error bars are standard error of the mean (SEM) and unpaired t-tests were performed between WT and KOs if the data was normal, and wilcoxon tests if the data was not normal. Each bar represents n=2 passage replicates unless otherwise specified by a number on the bar. *[*=p<0.05, **p<0.01]*

For example, using the M02 antibody on the same membrane following stripping, it's obvious no bands were present in any monoKO, and slightly less (CKO) if not none (NKO/SKO) in polyKO comparative to WT. This finding is an important observation relating to differences in antibody-ZFHX4 binding. Figure 4.22 demonstrates M02 binds a 91aa stretch spanning the translated KO region but HPA binds 344aa's further downstream to a 125aa stretch. This implies M02 binding is a surrogate for CRISPR targeting success, as changes to translated sequence following gDNA disruption and its transcription to altered mRNA likely means the antibody can no longer bind. This is consistent with M02 only ever demonstrating bands in a WT or partial WT in the case of CKO which had 62% WT sequence detected in original-3m cells (Figure 4.16). NKO and SKO (>94%) and all monoKOs (>91%) had non-WT sequence detected respectively which is a reason for absent banding by M02 because the original aa stretch is no longer detectable.

Furthermore, this also means that HPA is the better surrogate for loss of non-altered/altered protein expression. If the CRISPR targeting causes a frameshift, the downstream aa sequence will change and thus could affect HPA binding. For example, upon quantifying all HPA western blots, CB5, CC9, CF6 and CF9 significantly exhibited reduced ZFHX4 expression comparative to WT (unpaired t-tests, $p<0.001$, Figure 4.23C) and all but CF6 primarily comprised likely-frameshift indels (i.e. non-multiples of three) (Table 4.6) and displayed reduced transcript-level gene expression (Figure 4.18). Despite its characteristic 108bp del which is a multiple of three, this discrepancy in CF6 is likely due to the location

of this 108bp del, as alignment on benchling confirms this occurs within codon 39 (GAG to GA-) disrupting the reading frame and causing a frameshift. Furthermore, the unusual overexpression of ZFHX4 in NC8 and NG7 despite their confirmed gDNA disruption and reduction in transcript-level expression suggests these clones may have interesting feedback mechanisms pertaining to ZFHX4.

Taken together, the differences in banding between HPA and M02 reflect that absence does not necessarily reflect protein loss, rather it can also reflect protein alteration and dysfunction. All polyKOs and monoKOs appear to reflect interesting ZFHX4 biology at the gDNA, mRNA and protein level and thus all were selected for differential gene expression analysis to provide a thorough selection for biological replicates of different KOs.

4.4.5 Transcriptomic profiling of *ZFHX4*$^{-/-}$ cell lines

Following RNAseq, samples and replicates were plotted by PCA (Figure 4.24). Plotting all samples on one PCA, cell lines clustered together expectedly, however it seemed that there was observable diversity between NCIH716 monoKOs and a distancing of CC9 from all other COLO320DM samples. When plotting individual PCAs by cell line, WT and polyKO replicates nicely clustered together for every cell line. However, within SKNMC, greater distance within SKO clones compared to between SWT and SKO suggests that inter-replicate variability here was a greater influence on gene expression than ZFHX4 disruption at the gDNA level. Interestingly, most monoKOs segregated independently and separately from WTs and polyKOs. However, CB5 and CE5 unusually clustered with CWT replicates and this can't be explained by characterisation at the gDNA, mRNA and protein level. For example, CB5 predominantly contained frameshift mutations, had reduced transcript-level gene expression and reduced protein expression and was thus expected to differ significantly from WT.

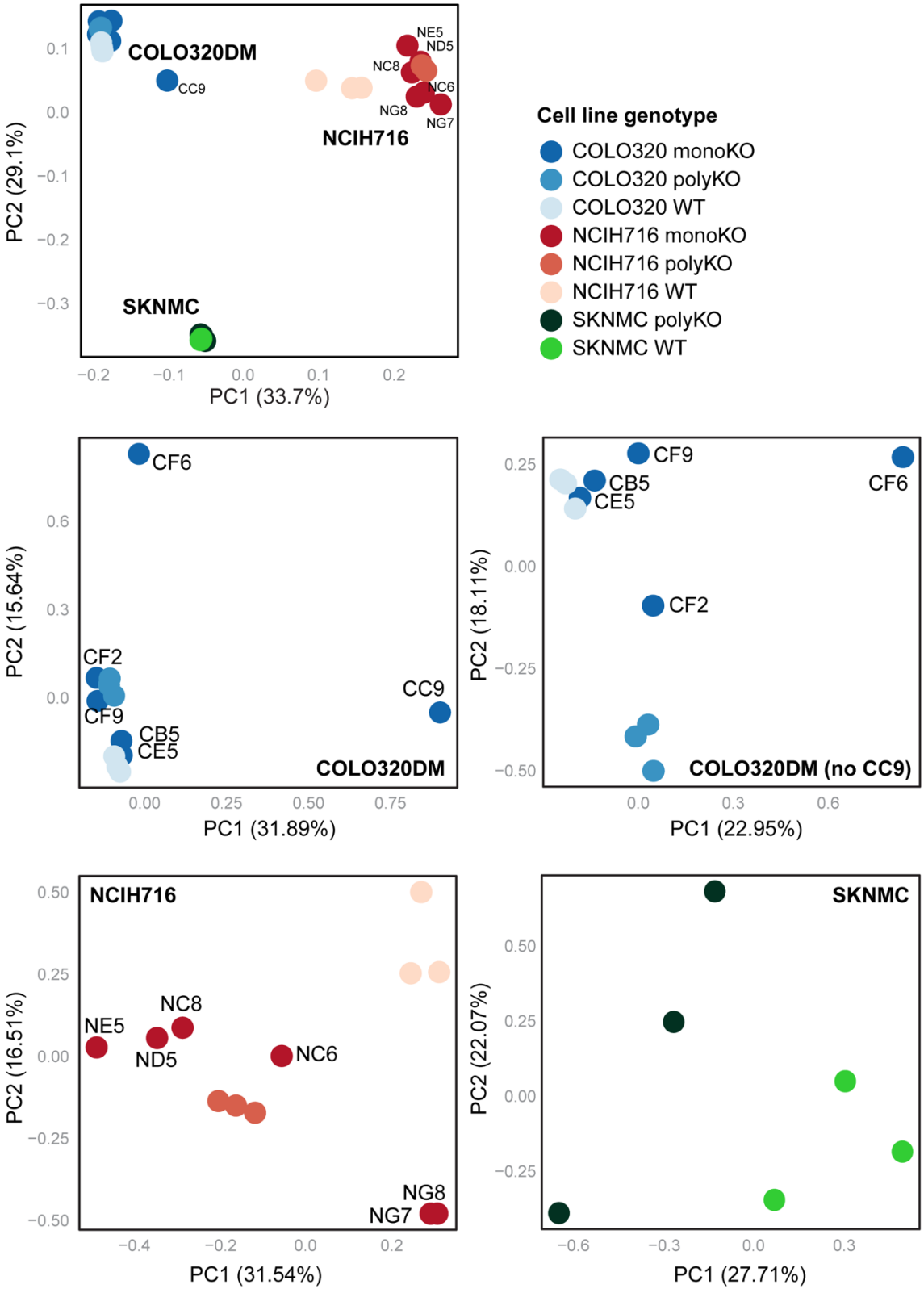


Figure 4.24 | PCA plots representing RNAseq expression profiles of WT, polyKO and monoKO cell lines. Dots are coloured respectively by the key. One COLO320DM PCA has CC9 removed to allow greater resolution in differences. WTs and polyKOs are in replicates of three, whilst monoKOs are n=1.

Firstly, differential gene expression analysis between SWT and SKO revealed a flat volcano plot with only 2 significantly downregulated (*ZNF600*, *TRIM49B*) and 1 significantly upregulated (*NEFL*) genes, as expected with consistent evidence of not much difference at the mRNA and protein level between the two (Figure 4.25A). Only 3 pathways were found significantly downregulated, pertaining to ribosomal pathways and *PRC2* (Figure 4.25B).

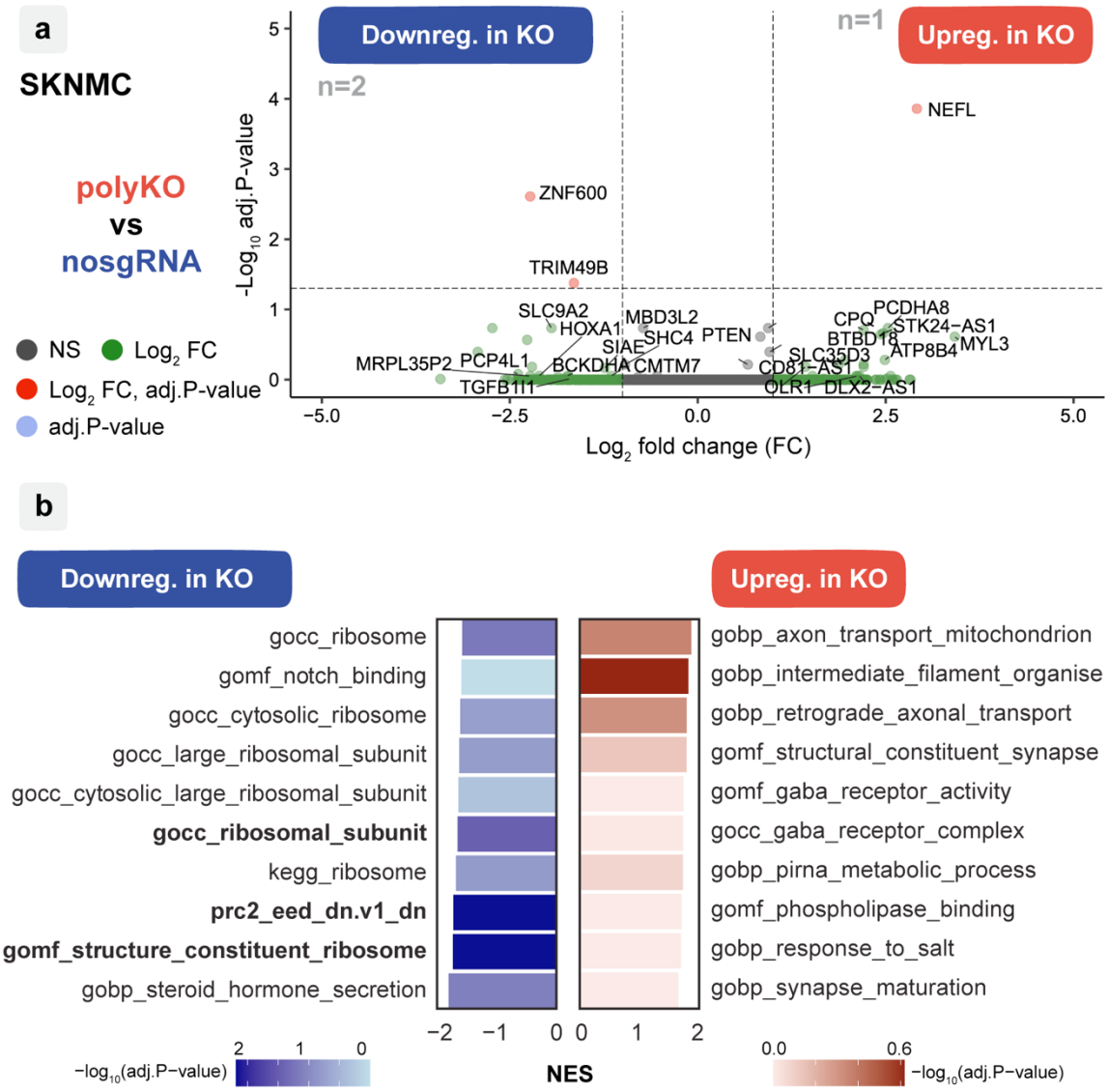
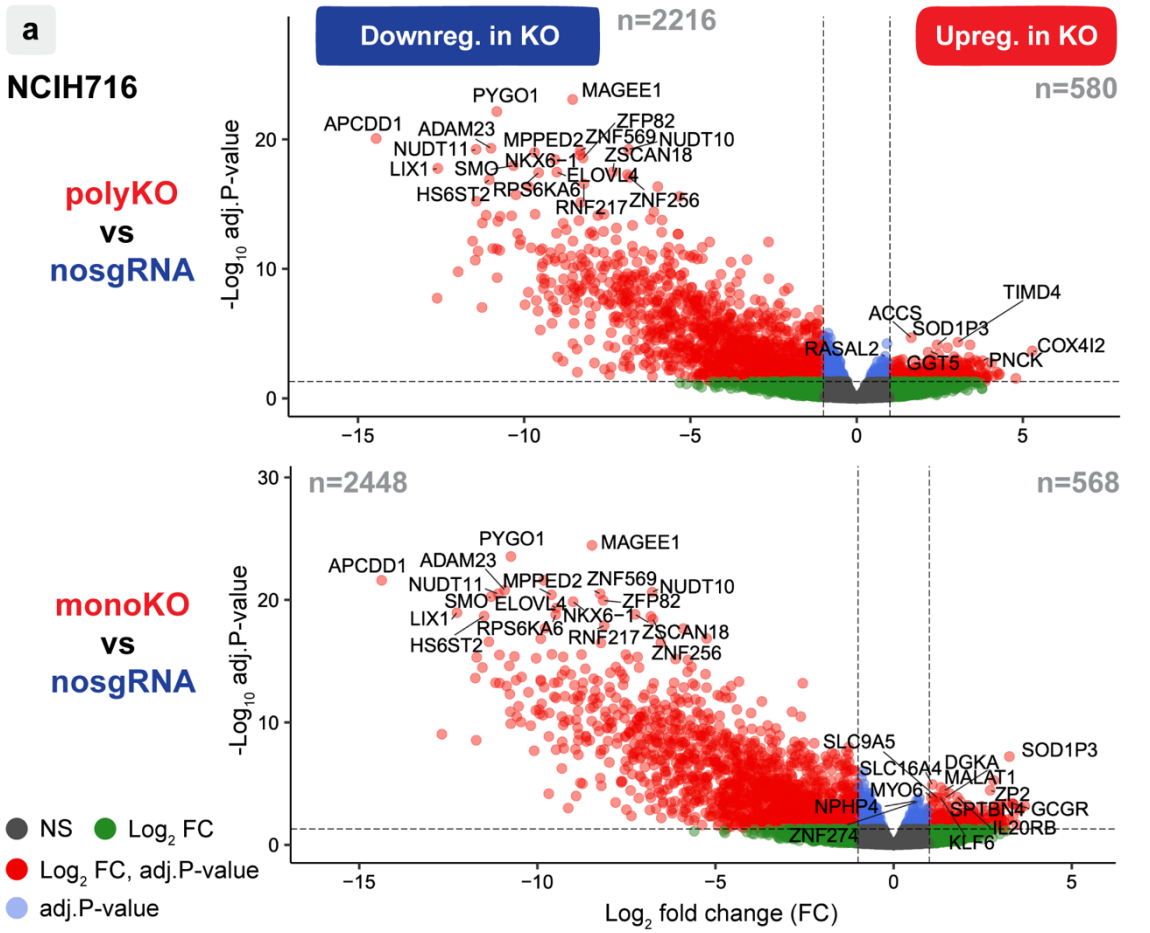
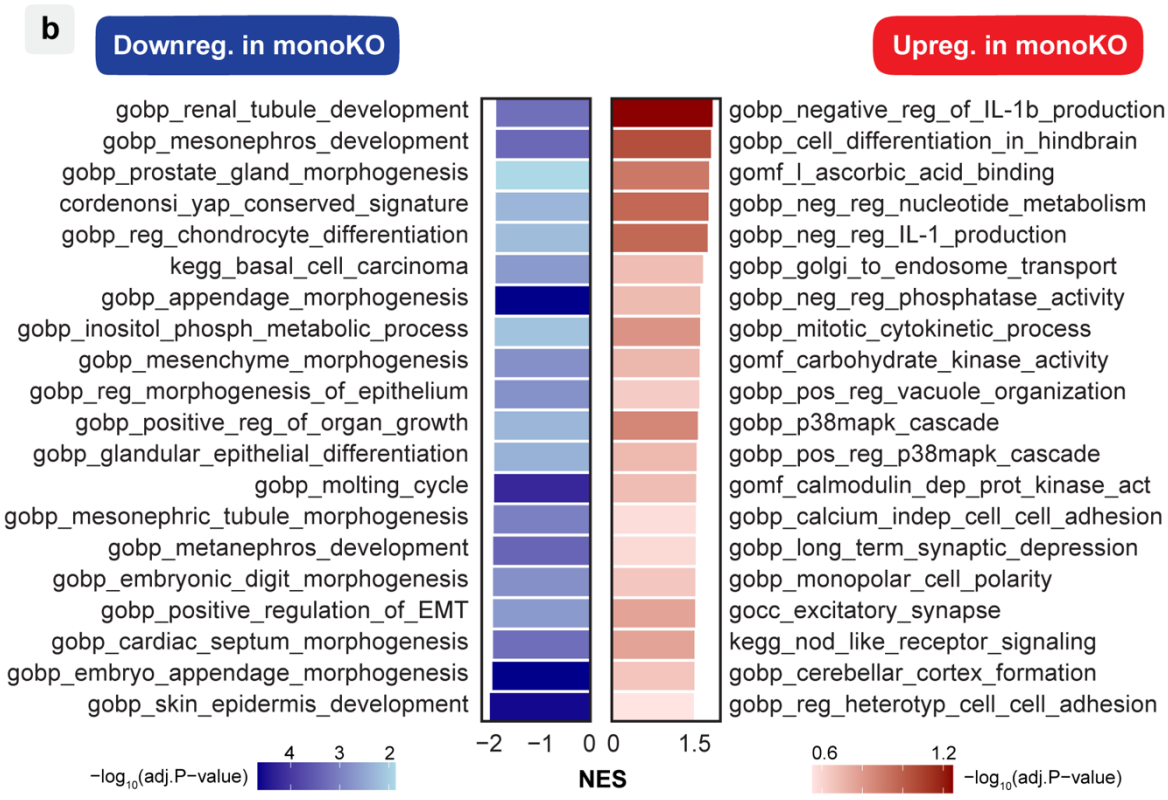


Figure 4.25 | Differential gene and pathway analysis of SKNMC ZFHX4$^{-/-}$ cells. (A) Volcano plots demonstrating the SWT and SKO comparison. Colours are representative of comparison and group. N numbers represent significant genes with respect to that group. Proteins with prefixes RP-, CTD-, AC- and LINC- had their labels removed from the plot. **(B)** Bar plots demonstrating normalised enrichment scores (NES) of pathway enrichment following GSEA on the differential genes obtained from each comparison ranked by t-statistic. Gene sets were obtained from msigdbr and included hallmark, biocarta, kegg, go-bp, go-mf, go-cc and oncogenic subsets. Top 10 pathways were selected based on NES ranking and coloured by significance, highlighting significant pathways in bold ($p<0.05$). Each bar colour intensity refers to significance of adj.p-value.

Differential gene expression regarding NCIH716 samples resulted in the most significant differences between WT and *ZFHX4*$^{-/-}$ (Figure 4.26A). Very similar volcano plots and top genes were consistent across both comparisons, demonstrating >2200 significantly downregulated genes and >500 significantly upregulated genes in *ZFHX4* KOs. This disproportionate regulation suggests that *ZFHX4* could be a transcriptional regulator and/or activator, leading to broad downregulation of many genes upon KO in NCIH716.

For a simplistic but robust approach, pathways up and downregulated in the monoKO comparison only were summarised because 6 biological replicates may provide better resolution than 3 technical replicates and more significant genes were observed (Figure 4.26B). In total, 254 pathways were significantly downregulated from utilised gene sets, and no pathway upregulated was significant by FDR. Ranking by NES and observing the top and bottom differential pathways as well as reporting other interesting significant pathways that were not in this top ranking (Supplementary Figure S4.13), there were clear trends of downregulated developmental, epithelial and morphogenetic programs relating to stemness and proliferation (WNT/YAP/EMT), suggesting that *ZFHX4* could be involved in promoting these features which are consistent with aggressive tumourigenicity and poor prognosis (Figure 4.4). Upregulated pathways in monoKOs describe cellular stress (p38 MAPK, NOD-like receptor), innate immune modulation and alterations in adhesion and polarity, consistent with a shift away from proliferative stem-like growth states toward stress and immune response states.


a
NCIH716
polyKO
vs
nosgRNA
Downreg. in KO
n=2216
Upreg. in KO
n=580
-Log10 adj.P-value
PYGO1
MAGEE1
APCDD1
ADAM23
ZFP82
MPPED2
ZNF569
NUDT10
NUDT11
ZSCAN18
LIX1
SMO
NKX6-1
ELOVL4
HS6ST2
RPS6KA6
RNF217
ZNF256
TIMD4
ACCS
SOD1P3
COX4I2
RASAL2
GGT5
PNCK
monoKO
vs
nosgRNA
n=2448
n=568
SLC9A5
SOD1P3
SLC16A4
DGKA
MALAT1
MYO6
ZP2
NPHP4
SPTBN4
GCGR
ZNF274
IL20RB
KLF6
NS
Log2 FC
Log2 FC, adj.P-value
adj.P-value
Log2 fold change (FC)


b
Downreg. in monoKO
Upreg. in monoKO
gobp_renal_tubule_development
gobp_mesonephros_development
gobp_prostate_gland_morphogenesis
cordenonsi_yap_conserved_signature
gobp_reg_chondrocyte_differentiation
kegg_basal_cell_carcinoma
gobp_appendage_morphogenesis
gobp_inositol_phosph_metabolic_process
gobp_mesenchyme_morphogenesis
gobp_reg_morphogenesis_of_epithelium
gobp_positive_reg_of_organ_growth
gobp_glandular_epithelial_differentiation
gobp_molting_cycle
gobp_mesonephric_tubule_morphogenesis
gobp_metanephros_development
gobp_embryonic_digit_morphogenesis
gobp_positive_regulation_of_EMT
gobp_cardiac_septum_morphogenesis
gobp_embryo_appendage_morphogenesis
gobp_skin_epidermis_development
gobp_negative_reg_of_IL-1b_production
gobp_cell_differentiation_in_hindbrain
gomf_l_ascorbic_acid_binding
gobp_neg_reg_nucleotide_metabolism
gobp_neg_reg_IL-1_production
gobp_golgi_to_endosome_transport
gobp_neg_reg_phosphatase_activity
gobp_mitotic_cytokinetic_process
gomf_carbohydrate_kinase_activity
gobp_pos_reg_vacuole_organization
gobp_p38mapk_cascade
gobp_pos_reg_p38mapk_cascade
gomf_calmodulin_dep_prot_kinase_act
gobp_calcium_indep_cell_cell_adhesion
gobp_long_term_synaptic_depression
gobp_monopolar_cell_polarity
gocc_excitatory_synapse
kegg_nod_like_receptor_signaling
gobp_cerebellar_cortex_formation
gobp_reg_heterotyp_cell_cell_adhesion
−log10(adj.P−value)
NES
−log10(adj.P−value)

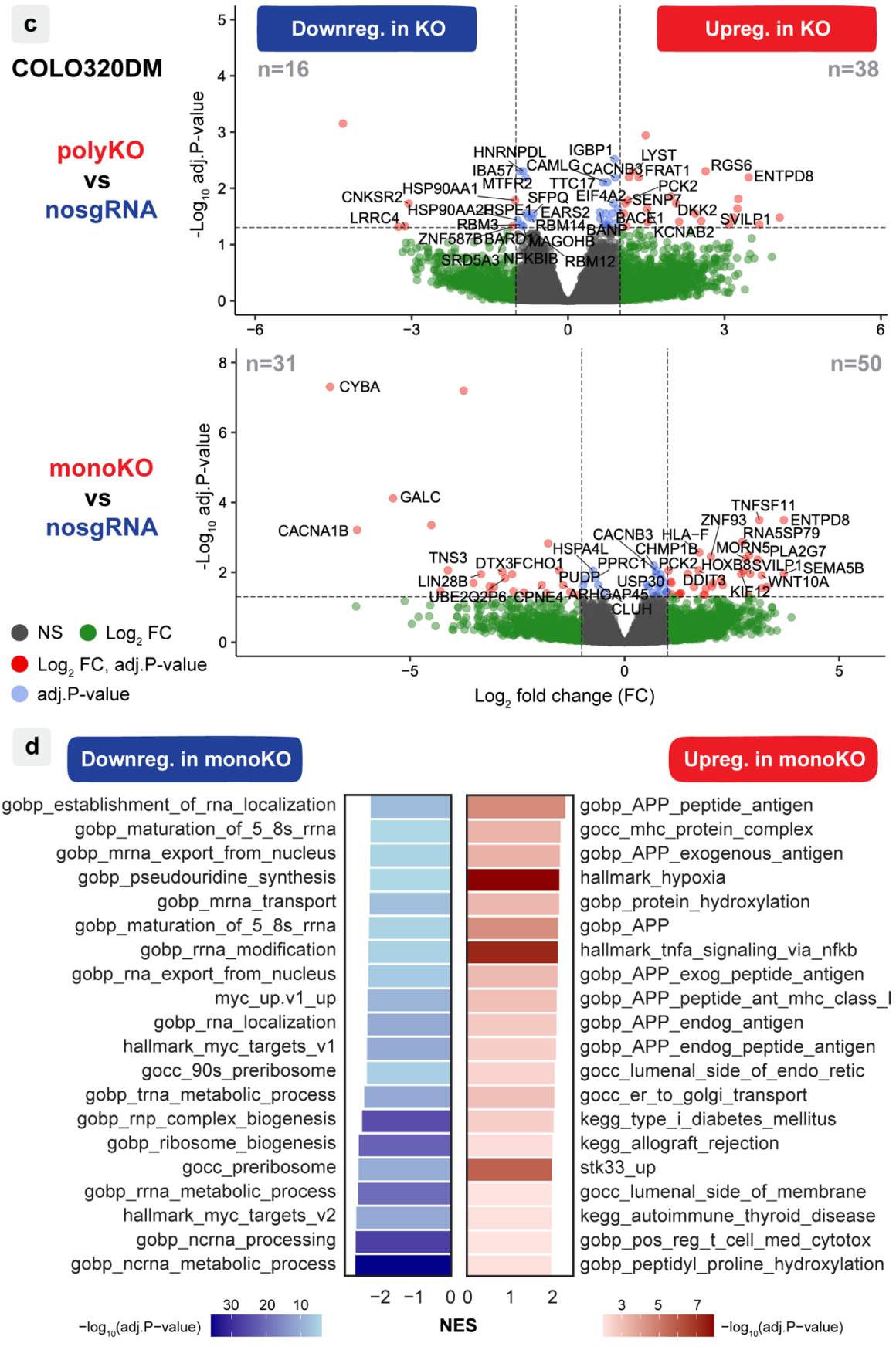

c
COLO320DM
polyKO
vs
nosgRNA
Downreg. in KO
Upreg. in KO
n=16
n=38
-Log10 adj.P-value
HNRNPDL
IGBP1
LYST
IBA57
CAMLG
CACNB3
FRAT1
RGS6
ENTPD8
MTFR2
TTC17
PCK2
CNKSR2
HSP90AA1
SFPQ
EIF4A2
SENP7
HSP90AA2
HSPE1
EARS2
DKK2
LRRC4
RBM3
RBM14
BANP
BACE1
SVILP1
KCNAB2
ZNF587B
BARD1
MAGOHB
SRD5A3
NFKBIB
RBM12
monoKO
vs
nosgRNA
n=31
n=50
CYBA
GALC
CACNA1B
TNFSF11
ZNF93
ENTPD8
CACNB3
HLA-F
RNA5SP79
HSPA4L
CHMP1B
MORN5
PLA2G7
TNS3
DTX3FCHO1
PPRC1
PCK2
HOXB8
SVILP1
SEMA5B
LIN28B
PUDP
USP30
DDIT3
WNT10A
UBE2Q2P6
CPNE4
ARHGAP45
KIF12
CLUH
NS
Log2 FC
Log2 FC, adj.P-value
adj.P-value
Log2 fold change (FC)
d
Downreg. in monoKO
Upreg. in monoKO
gobp_establishment_of_rna_localization
gobp_maturation_of_5_8s_rrna
gobp_mrna_export_from_nucleus
gobp_pseudouridine_synthesis
gobp_mrna_transport
gobp_maturation_of_5_8s_rrna
gobp_rrna_modification
gobp_rna_export_from_nucleus
myc_up.v1_up
gobp_rna_localization
hallmark_myc_targets_v1
gocc_90s_preribosome
gobp_trna_metabolic_process
gobp_rnp_complex_biogenesis
gobp_ribosome_biogenesis
gocc_preribosome
gobp_rrna_metabolic_process
hallmark_myc_targets_v2
gobp_ncrna_processing
gobp_ncrna_metabolic_process
gobp_APP_peptide_antigen
gocc_mhc_protein_complex
gobp_APP_exogenous_antigen
hallmark_hypoxia
gobp_protein_hydroxylation
gobp_APP
hallmark_tnfa_signaling_via_nfkb
gobp_APP_exog_peptide_antigen
gobp_APP_peptide_ant_mhc_class_I
gobp_APP_endog_antigen
gobp_APP_endog_peptide_antigen
gocc_lumenal_side_of_endo_retic
gocc_er_to_golgi_transport
kegg_type_i_diabetes_mellitus
kegg_allograft_rejection
stk33_up
gocc_lumenal_side_of_membrane
kegg_autoimmune_thyroid_disease
gobp_pos_reg_t_cell_med_cytotox
gobp_peptidyl_proline_hydroxylation
−log10(adj.P−value)
NES
−log10(adj.P−value)

Figure 4.26 | Differential gene and pathway analysis of NCIH716 (A-B) and COLO320DM (C-D) ZFHX4$^{-/-}$ cells. (A/C) Volcano plots demonstrating the respective comparison. Colours are representative of comparison and group. N numbers represent significant genes with respect to that group. Proteins with prefixes RP-, CTD-, AC- and LINC- had their labels removed from the plot. **(B/D)** Bar plots demonstrating normalised enrichment scores (NES) of pathway enrichment following GSEA on the differential genes obtained from the monoKO comparison only, ranked by t-statistic. Gene sets were obtained from msigdbr and included hallmark, biocarta, kegg, go-bp, go-mf, go-cc and oncogenic subsets. Top 20 pathways were selected based on NES ranking and coloured by significance. Each bar colour intensity refers to significance of the adj.p-value. *[APP=antigen processing and presentation]*

COLO320DM samples provided slightly different results following differential gene expression, ascertaining the decision to analyse these lines separately irrespective of *ZFHX4* gDNA status. Exhibiting ~13-fold less significantly upregulated and ~100-fold less significantly downregulated genes in KOs comparative to WT, the COLO320DM line was less diverse but more balanced in its phenotype upon *ZFHX4*$^{-/-}$ compared to the NCIH716 line (Figure 4.26C). Following the same approach as for NCIH716, pathway-level analysis demonstrated 160 significantly upregulated and 278 significantly downregulated pathways in *ZFHX4* monoKOs compared to WTs (Figure 4.26D). COLO320DM monoKOs demonstrated downregulation in RNA processing and trafficking, ribosome assembly and oncogenic driver pathways such as MYC, stem cell proliferation and chromatin organisation, pathways central to growth and stemness. This downregulation of ribosome production was also significantly evident in SKNMC monoKOs. Conversely, COLO320DM monoKOs exhibited upregulation of immune activation, antigen presentation, stress and inflammatory signalling pathways suggesting a shift from a proliferative state to a more ‘visible’ or immunogenic state.

In summary, both NCIH716 and COLO320DM cell lines show loss of a pro-growth program (WNT/YAP/EMT vs MYC/ribosome) and gain of a stress/immune program (p38MAPK/polarity changes vs MHC/immune recognition) upon *ZFHX4*$^{-/-}$ but through different transcriptional axes respectively. Therefore, *ZFHX4* appears to drive proliferative oncogenic circuits and limit stress/immune-visible states differently depending on context.

4.4.6 Validating ZFHX4 transcriptional targets with CUT&RUN

CUT&RUN profiles protein-DNA interactions and histone marks without crosslinking, using an in-situ antibody plus protein AG-MNase to release DNA fragments. Versus ChIP sequencing (ChIPseq), it requires fewer cells, shows lower background, avoids crosslink artefacts and requires less depth. Recent success mapping ZFHX4 in healthy human neural progenitors (Perez Baca et al., 2025) made CUT&RUN the preferred approach here. Prior ZFHX4 targets come only from mouse brainstem ChIPseq (Zhang et al., 2021a) and GBM ChIPseq (Chudnovsky et al., 2014) limiting inference to colon and underscoring the novelty of this setting. CUT&RUN was performed in nosgRNA lines in a pilot and main experiment (Table 4.7), using various kits and ZFHX4 and control (H3K4me3, IgG) antibodies. Both experiments yielded robust promoter-proximal H3K4me3 signal (<2500 peaks) across all samples, but no specific ZFHX4 peaks were detected (Table 4.7). H3K4me3 peak overlap exceeded 50% between pilot and main runs (Figure 4.27A), arguing against technical failure and pointing to antibody/protein-specific causes such as ZFHX4 not DNA-bound in these lines, a more closed chromatin context or epitope accessibility issues.

Table 4.7 | Number of called peaks with SEACR, using the negative IgG control as background

Experiment	Cell Line	Antibody/Condition	Number of called peaks post-processing (3kb, unique)
Pilot (CST kit)	**Jurkat**	H3K4me3	4106
		ZFHX4-PA5	0
		ZFHX4-abc	0
		ZFHX4-M02	58 (504 pre-processing) (Jurkats are a negative control so these peaks are likely due to MNase-accessible noise)
		ZFHX4-M11	0
	NCIH716	H3K4me3	2857
		ZFHX4-PA5	0
		ZFHX4-abc	0
		ZFHX4-M02	0
		ZFHX4-M11	0
Main (Epicypher kit) as by Baca et al.	**Jurkat**	H3K4me3	4509
		ZFHX4-HPA	0
	NCIH716	H3K4me3	2628
		ZFHX4-HPA	0
	SKNMC	H3K4me3	1506
		ZFHX4-HPA	0
	COLO320DM	H3K4me3	4479
		ZFHX4-HPA	0

Although no ZFHX4 peaks were detected, robust H3K4me3 signal enabled chromatin-state and target inference using the published CUT&RUN dataset, referred to as Baca (Perez Baca et al., 2025). As a QC, consistent with H3K4me3 modifications designating transcriptionally active genes, H3K4me3 peak genes correlated with WT-average RNAseq expression in SKNMC, COLO320DM, and NCIH716 (spearman's rho 0.198–0.265, $p<2.2x10^{-16}$; Figure 4.27B), confirming expected biology and supporting CUT&RUN in colon cancer cell lines – an approach that to my knowledge has not been previously reported. ZFHX4 was H3K4me3-marked within ±3 kb of its TSS in SKNMC and NCIH716, matching its high expression (Figure 4.8). Because peak height tracked expression better than breadth, it was used as the promoter-activity metric: COLO320DM showed the highest mean height (194), 6.3x and 4.3x higher than NCIH716 (30.8) and SKNMC (45.03) (Wilcoxon; Figure 4.27C). Gene sets of H3K4me3-marked genes were enriched among highly expressed genes in all lines, strongest in COLO320DM (NES=4.994, FDR=0.001; Figure 4.27D). Together, these findings indicate that COLO320DM harbours a globally more permissive chromatin landscape, while also highlighting distinct, cell line-specific H3K4me3 profiles.

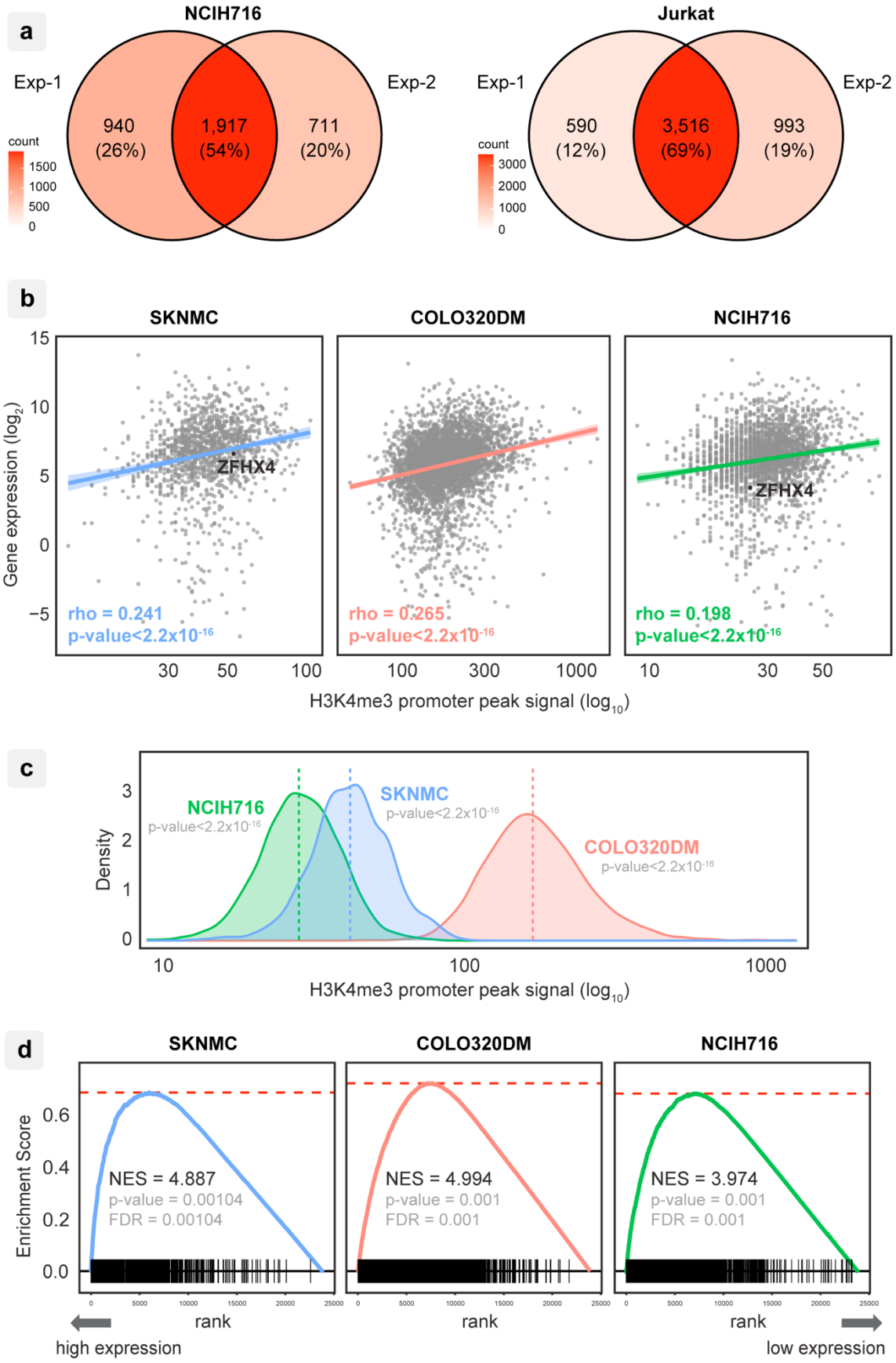
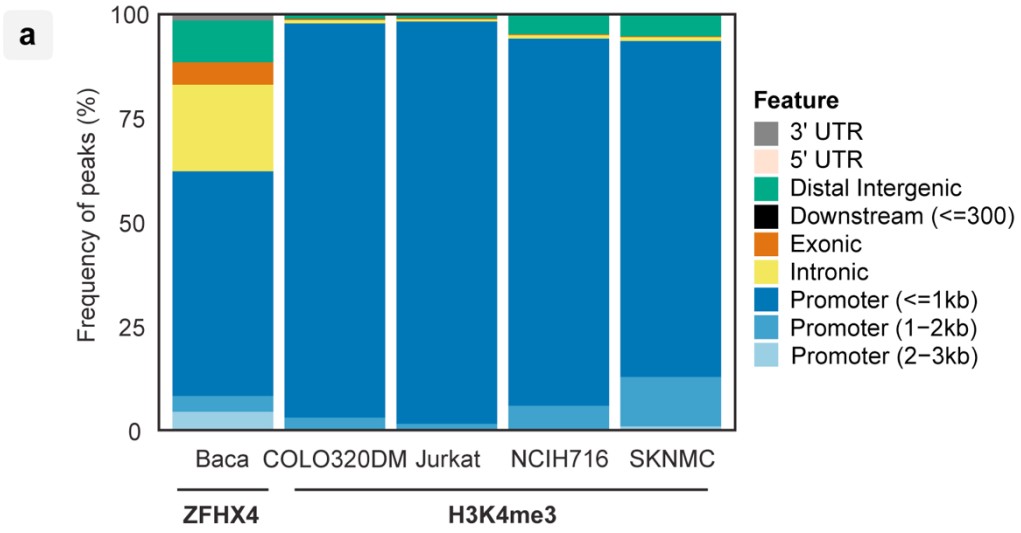

a
NCIH716
Exp-1
Exp-2
940 (26%)
1,917 (54%)
711 (20%)
count
1500
1000
500
0
Jurkat
590 (12%)
3,516 (69%)
993 (19%)
3000
2000
1000
b
SKNMC
COLO320DM
NCIH716
Gene expression (log2)
ZFHX4
rho = 0.241
p-value<2.2x10-16
rho = 0.265
p-value<2.2x10-16
rho = 0.198
p-value<2.2x10-16
H3K4me3 promoter peak signal (log10)
c
NCIH716
p-value<2.2x10-16
SKNMC
p-value<2.2x10-16
COLO320DM
p-value<2.2x10-16
Density
H3K4me3 promoter peak signal (log10)
d
SKNMC
COLO320DM
NCIH716
Enrichment Score
NES = 4.887
p-value = 0.00104
FDR = 0.00104
NES = 4.994
p-value = 0.001
FDR = 0.001
NES = 3.974
p-value = 0.001
FDR = 0.001
rank
high expression
low expression

Figure 4.27 | CUT&RUN of H3K4me3 peaks on SKNMC, NCIH716 and COLO320DM cell lines. **(A)** Concordance venn diagrams overlapping H3K4me3 peaks identified in the pilot experiment 1 (exp-1) and main run experiment-2 (exp-2). Only NCIH716 and Jurkat lines were used for both experiments. Exp-2 data contained all four cell lines and was thus used as the primary dataset. **(B)** Scatter plots showing correlation between matched RNAseq-derived gene expression and H3K4me3 peak signal, defined as signal height of H3K4me3 marks within 3kb of TSS. Spearman's Rho correlation was performed and results annotated along with a linear regression trendline. **(C)** H3K4me3 peak signal distributions within each cell line, annotating medians with dashed lines and p-values following pairwise non-parametric wilcoxon testing (only 1 p-value shown per line but were all the same). **(D)** Enrichment plots of cell line-specific H3K4me3 genes on genes ranked by order of highest to lowest expression.

With CUT&RUN validated to capture expected biology, I next assessed potential ZFHX4 activity by overlapping H3K4me3-marked promoters in each line with Baca ZFHX4 peaks (from non-cancer neuronal progenitors), assuming ZFHX4 acts as an activator. Baca peaks mapped ~62% to promoters, 21% intronic and 10% distal intergenic (Figure 4.28A). By contrast, H3K4me3 in all four lines was >80.6% promoter-proximal (<1 kb of TSS), as expected for a promoter-centric mark, and thus ZFHX4 binding in Baca appears broader and less promoter-restricted, consistent with a transcription factor-behaviour of occupying both promoter and enhancer or distal intergenic elements. Overlaps with my WT RNAseq and CUT&RUN showed ~one-third of ZFHX4 peaks at highly expressed genes, and a smaller but meaningful subset at H3K4me3-marked promoters, consistent with ZFHX4 engagement at transcriptionally permissive loci in colon-derived cells (Fig. 4.28B).


a
100
75
50
25
0
Frequency of peaks (%)
Baca
COLO320DM
Jurkat
NCIH716
SKNMC
ZFHX4
H3K4me3
Feature
3' UTR
5' UTR
Distal Intergenic
Downstream (<=300)
Exonic
Intronic
Promoter (<=1kb)
Promoter (1−2kb)
Promoter (2−3kb)

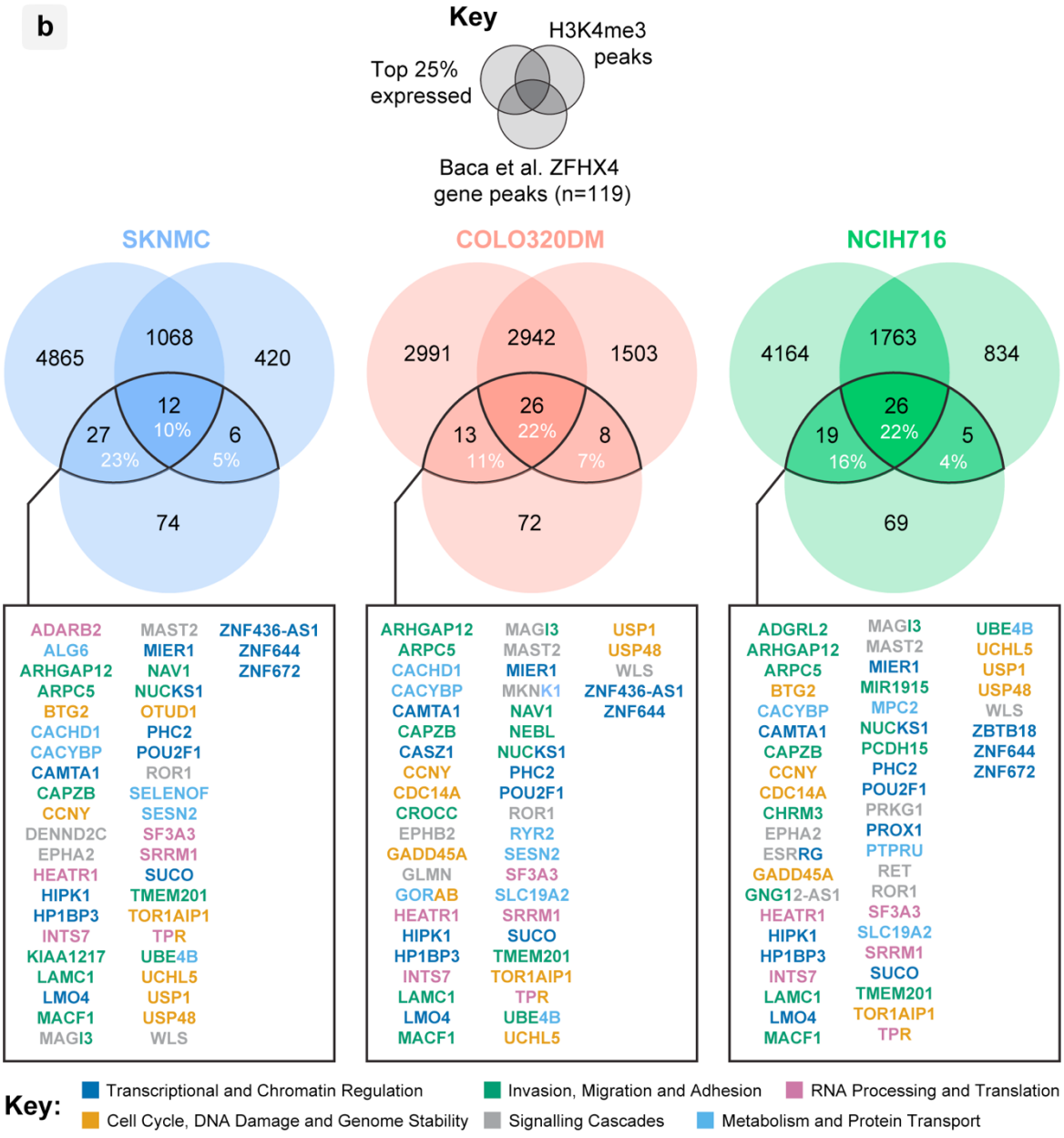

b
Key
Top 25% expressed
H3K4me3 peaks
Baca et al. ZFHX4 gene peaks (n=119)
SKNMC
4865
1068
420
12
10%
27
23%
6
5%
74
ADARB2 ALG6 ARHGAP12 ARPC5 BTG2 CACHD1 CACYBP CAMTA1 CAPZB CCNY DENND2C EPHA2 HEATR1 HIPK1 HP1BP3 INTS7 KIAA1217 LAMC1 LMO4 MACF1 MAGI3
MAST2 MIER1 NAV1 NUCKS1 OTUD1 PHC2 POU2F1 ROR1 SELENOF SESN2 SF3A3 SRRM1 SUCO TMEM201 TOR1AIP1 TPR UBE4B UCHL5 USP1 USP48 WLS
ZNF436-AS1 ZNF644 ZNF672
COLO320DM
2991
2942
1503
26
22%
13
11%
8
7%
72
ARHGAP12 ARPC5 CACHD1 CACYBP CAMTA1 CAPZB CASZ1 CCNY CDC14A CROCC EPHB2 GADD45A GLMN GORAB HEATR1 HIPK1 HP1BP3 INTS7 LAMC1 LMO4 MACF1
MAGI3 MAST2 MIER1 MKNK1 NAV1 NEBL NUCKS1 PHC2 POU2F1 ROR1 RYR2 SESN2 SF3A3 SLC19A2 SRRM1 SUCO TMEM201 TOR1AIP1 TPR UBE4B UCHL5
USP1 USP48 WLS ZNF436-AS1 ZNF644
NCIH716
4164
1763
834
26
22%
19
16%
5
4%
69
ADGRL2 ARHGAP12 ARPC5 BTG2 CACYBP CAMTA1 CAPZB CCNY CDC14A CHRM3 EPHA2 ESRRG GADD45A GNG12-AS1 HEATR1 HIPK1 HP1BP3 INTS7 LAMC1 LMO4 MACF1
MAGI3 MAST2 MIER1 MIR1915 MPC2 NUCKS1 PCDH15 PHC2 POU2F1 PRKG1 PROX1 PTPRU RET ROR1 SF3A3 SLC19A2 SRRM1 SUCO TMEM201 TOR1AIP1 TPR
UBE4B UCHL5 USP1 USP48 WLS ZBTB18 ZNF644 ZNF672
Key:
Transcriptional and Chromatin Regulation
Invasion, Migration and Adhesion
RNA Processing and Translation
Cell Cycle, DNA Damage and Genome Stability
Signalling Cascades
Metabolism and Protein Transport

Figure 4.28 | Overlap of CUT&RUN H3K4me3-centric cell line data with matched RNAseq and ZFHX4-centric CUT&RUN data from the literature. (A) Stacked bar charts representing the locational proportion of ZFHX4-peaks (from (Perez Baca et al., 2025)) and H3K4me3 peaks from our experiment 2 data. **(B)** Overlap of identified peaks with these CUT&RUN data as well as matched RNAseq of the top 25% expressed genes according to cell line. Percentages are with respect to ZFHX4-peaks (n=119). All genes identified in the overlap per cell line are listed and coloured by general cancer pathway involvement obtained from manual literature searching.

Literature-based annotation showed the overlapping genes (Figure 4.28B) map mainly to transcription/chromatin control, cell cycle, genome stability and invasion/migration/adhesion, with additional roles in RNA processing/translation, metabolism and protein transport. To infer co-regulators, TRANSFAC motif enrichment (gProfiler) was run on genes overlapping in any line with high expression and/or H3K4me3 (n=68; Table 4.8). Enriched TFs span cancer pathways including proliferation, migration, invasion, EMT, stemness and differentiation. Notably, ZNF148, SP1, MAZ, and EGR1 which bind GC-rich promoters suggest ZFHX4 operates in the same regulatory neighbourhoods, potentially co-operating or competing at GC-rich elements. The appearance of SP7/Osterix, a reported Zfhx4 partner in mice (Nakamura et al., 2021), supports this analysis.

Table 4.8 | TRANSFAC motif enrichment in the set of all overlapping genes (n=68) performed by gProfiler. Only significant results reported (BH corrected FDR)

TF	TF Motif ID	adj.p-value
ZNF148/ZBP89	TF:M09834_1	0.0003
	TF:M07397_1	0.0073
	TF:M12464_1	0.0179
SP1	TF:M02281_1	0.0009
	TF:M04953_1	0.0030
MAZ	TF:M09984_1	0.0191
	TF:M12405	0.0369
ETF	TF:M00695_1	0.0199
EGR1	TF:M01873	0.0199
	TF:M07208_1	0.0274
	TF:M07208	0.0483
PATZ1/PATZ	TF:M10026_1	0.0261
OSX/SP7	TF:M07329_1	0.0261
WT1	TF:M01118_1	0.0274
CKROX	TF:M01175_1	0.0369
ZNF281/ZBP99	TF:M12057_1	0.0389
ZNF93	TF:M12742	0.0426

However, ~60% of Baca ZFHX4 peaks didn't overlap with highly expressed or H3K4me3-marked genes in my data. These sites may denote repressed loci (closed chromatin/lack of

co-factors), context-specific targets active in neural but not colon models or enhancers distal to promoters (consistent with 38% non-promoter peaks in Baca). Without cell-matched ZFHX4 binding data in my colon models, its true regulatory activity remains unresolved, despite detectable expression.

4.4.7 Functional profiling of *ZFHX4*$^{-/-}$ cell lines

4.4.7.1 Investigating proliferation

Because ZFHX4 loss has been linked to reduced proliferation in GBM stem cells (Chudnovsky et al., 2014) and two monoKOs (CC9, NG7) grew slowly post-CRISPR, I profiled proliferation across all lines using CellTiter-Glo, repeating assays over time to control for passage effects. At 2 months post-editing (2m), most monoKOs matched WT growth (Fig. 4.29A) but CC9 (1.97x) and NG7 (1.54x) showed markedly lower day0-normalised viable cell number (VCN) at day 2 than WT. At later passages (3m, 5m, 11m), differences fluctuated, implicating clonal variation rather than ZFHX4 perturbation per se. To investigate this, I generated single-cell monoclonal WT (monoWT) clones from polyclonal WT (polyWT) lines and interestingly these displayed substantial heterogeneity in proliferation when compared to their parental polyWTs (Figure 4.29B). For example, CWT_D5 showed a 3.31-fold reduction in normalised VCN at day 6 compared to its polyWT line, whereas another monoWT clone (CWT_D11) revealed 1.55-fold increased normalised VCN at the same time point. Furthermore, these differences were also apparent when compared to the original parent noCRISPR line highlighting that this variation wasn't due to an electroporation-Cas9-mediated mechanism.

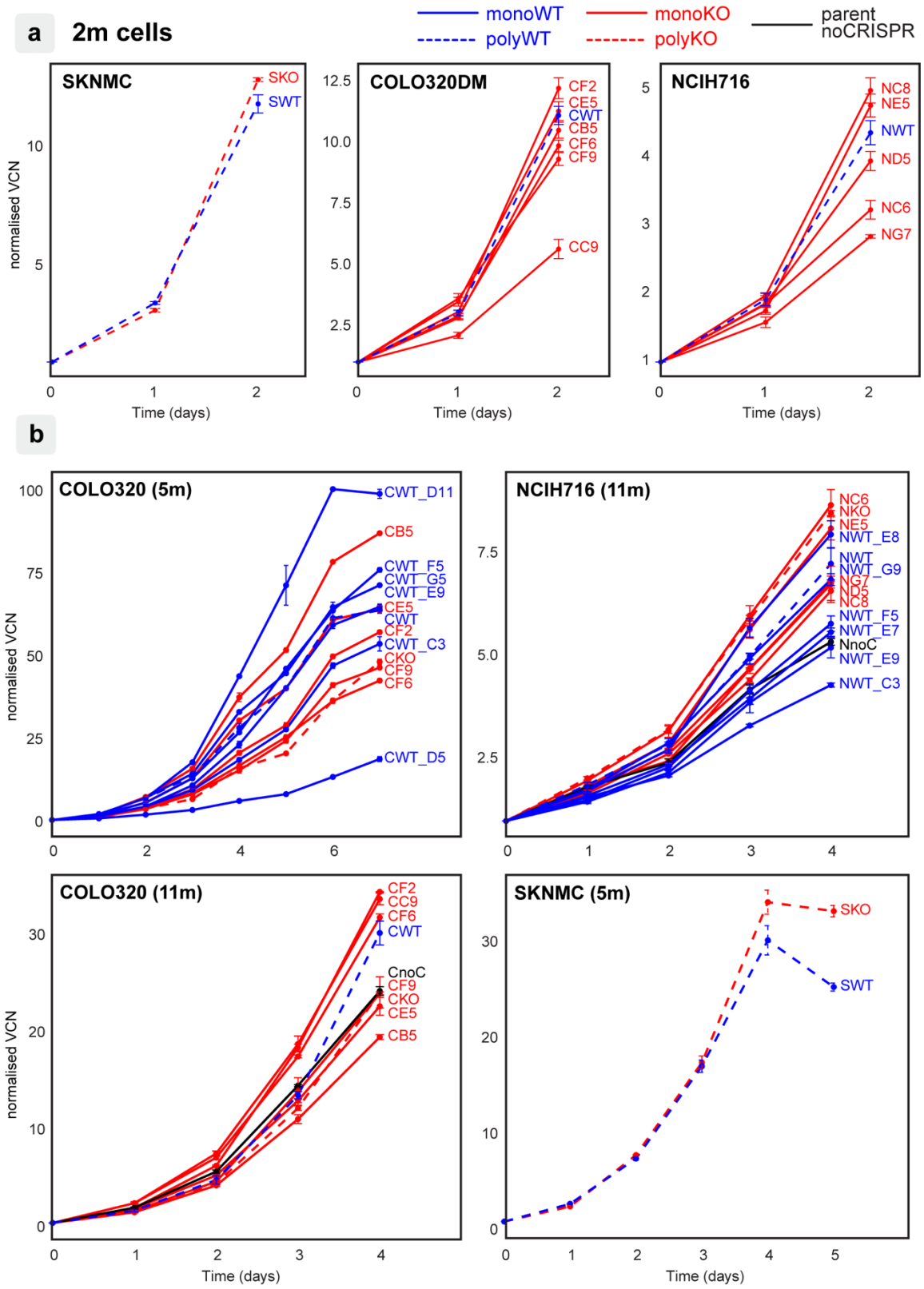


Figure 4.29 | Linear growth curves of various *ZFHX4-WT* and *ZFHX4*$^{-/-}$ cell lines, plotting normalised viable cell number (VCN) to day 0 over time. (A) Initial growth curves obtained using cells 2 months post-CRISPR (2m-cells). **(B)** Growth curves representing the last time proliferation was performed on those lines and/or a comparison with single-cell cloned WT lines. These consisted of either 5m or 11m cells. The colour key at the top is representative for all plots. Points are means with error bars representing standard error of the mean (SEM, n=3).

Circumventing the issues with comparing growth trends using linear curves and combining four experimental repeats of various timepoints and cell lines, all data was summarised and quantified to allow for statistical analysis. Doubling times were calculated based on the 'best exponential' growth phase per clone per experiment (see Supplementary Methods 2.1) and compared as a metric for proliferation (Figure 4.30). Consistent with culture, SKNMCs represented the fastest growing cells, with WT populations doubling in less than 24 hours (doubling time = 0.682), similarly to WT COLO320DMs (doubling time = 0.903). Conversely, WT NCIH716s grew 1.8-fold more slowly than COLO320DMs, doubling every ~39 hours. As expected due to the substantial growth heterogeneity observed, no significant differences were found by Wilcoxon testing when comparing doubling times of *ZFHX4*-WT cells vs *ZFHX4*$^{-/-}$ cells within all cell lines, agnostic to clonality, experiment and passage.

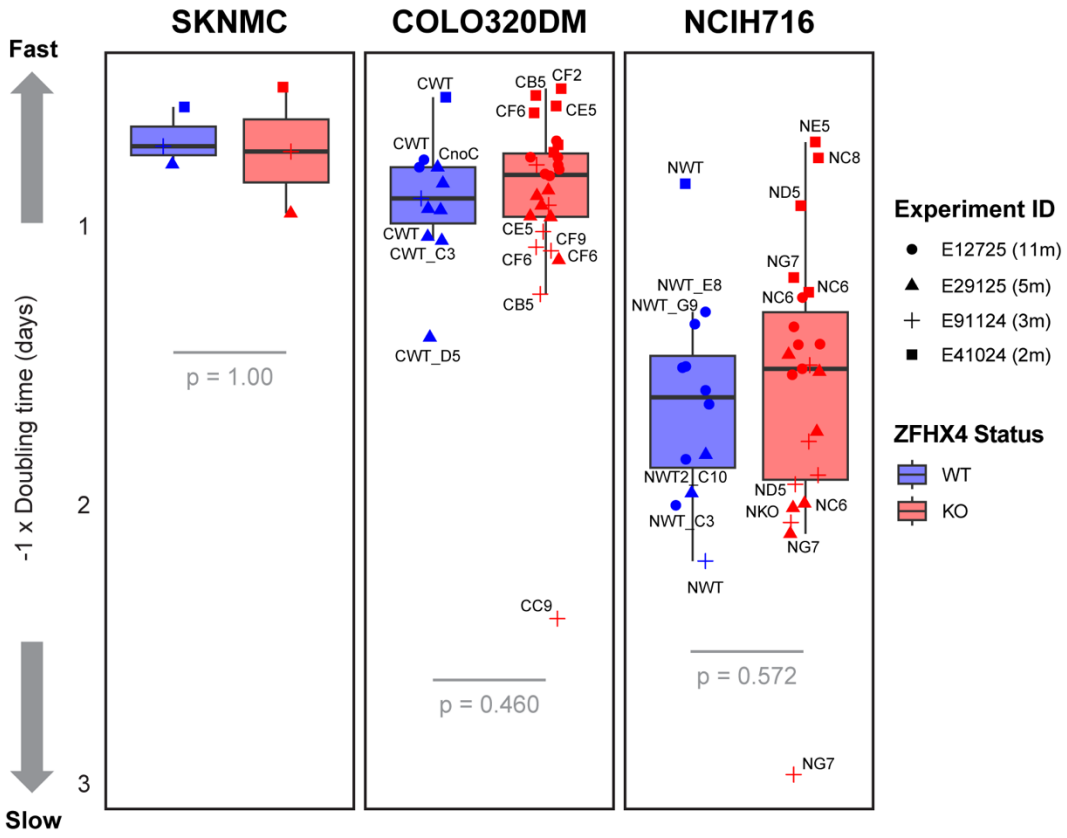


Figure 4.30 | Quantification of proliferation in SKNMC, COLO320DM and NCIH716 WT and *ZFHX4*$^{-/-}$ lines, defined using doubling time calculations. Box plots represent means of doubling times calculated from any clone irrespective of clonality of any experiment with respect to ZFHX4 status (WT vs KO). Experiment IDs relate to age of cells in months post-CRISPR. Y axis of doubling time was inverted for easier interpretation of growth speed. P-values represent results from Wilcoxon testing following non-normal confirmation by Shapiro-Wilkes. Clones furthest from the mean are indicated.

The confounding of clonal variation in investigating proliferation in cell lines was also impacted by confounding of long-term culture. For example, the change in CC9's growth rate over passages was stark, exhibiting 1.32-fold slower growth than CWT at 2m but a 1.01-fold faster growth than CWT at 11m. This was also consistent with cell culture where over time expansion for this clone became easier and eventually seemed to catch up with all other COLO320DM lines. Furthermore, this observation was also found with NG7 (1.40-fold slower at 2m, and 1.002-fold faster at 11m). Taken together, these results indicate clonal heterogeneity alone can produce artefactual differences in proliferation unrelated to *ZFHX4* status. Given this variability especially after prolonged culture (11 months), these models weren't well-suited for reliably assessing ZFHX4's impact on cell proliferation.

To alleviate confounding of clonal effects and passage-dependent drift, I performed siRNA knock-down (KD) of *ZFHX4* in nosgRNA WT COLO320DM cells utilising two different transfection reagents to enhance success (Figure 4.31A and 4.31B respectively). After 3 days, WB demonstrated KD success for all siRNAs, displaying a reduced banding pattern comparative to control cells albeit without dose-dependent reductions. In all conditions an apparent top-band remained visible, hinting incomplete KD which isn't surprising considering the nature of siRNA to reduce, not ablate, mRNA levels. siRNA-1 appeared the most efficient siRNA for reducing protein levels, however no differences in cell growth by CTG were observed after 3 days compared to control cells. siRNA-3 demonstrated significant slight reductions in proliferation across all siRNA concentrations in the first experiment but as siRNA-1 led to lower protein expression in comparison, it's difficult to correlate this with protein reduction and hints complexity in targeting specific exon-derived mRNA. This therefore suggests ZFHX4-KD doesn't alter COLO320DM proliferation and alongside *ZFHX4*$^{-/-}$ proliferation data, evidence for ZFHX4's association with proliferation remains elusive.

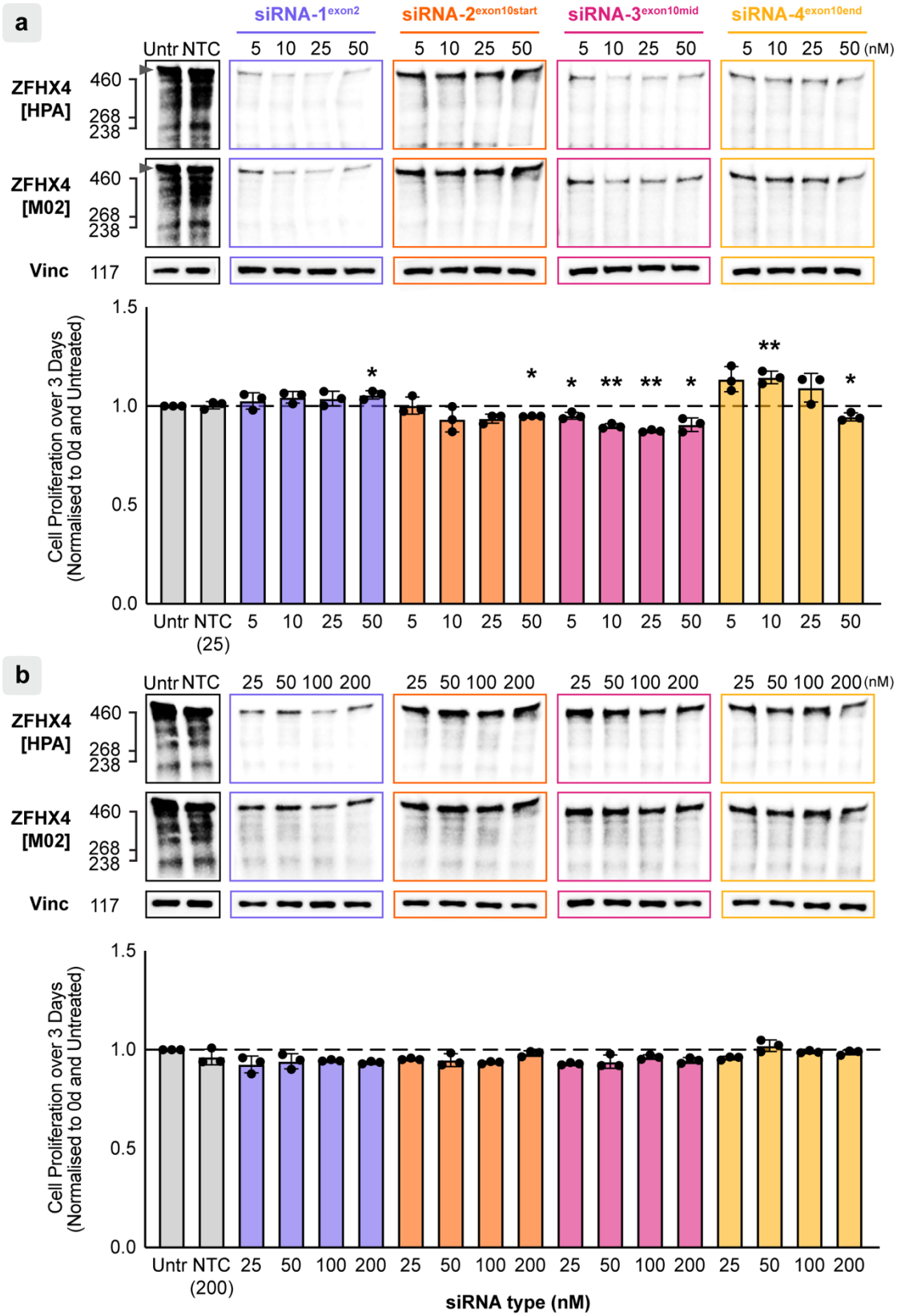


Figure 4.31 | siRNA knock-down of ZFHX4 in nosgRNA COLO320DM cells and resulting proliferation differences observed after 3 days. siRNA1/2 and siRNA3/4 samples were run on separate gels next to untreated control (Untr) and no-template control siRNA (NTC) (only one shown). HPA and M02 ZFHX4 antibodies used as usual post-stripping. Bar charts represent viable cell number at 3 days, normalised to 0d and untreated control (dashed line at y=1). Westerns and

proliferation assays were performed on the same batch of siRNA-KD cells at the same time, later separated into relevant assay plates. Statistical testing was performed comparing KD groups with NTC, using parametric unpaired t-tests if normal and nonparametric wilcoxons if not normal by Shapiro-Wilkes. **(A)** siRNA experiment using 1:100 DharmaFECT1 and siRNA concentrations 5-50nM. **(B)** siRNA experiment using 1:100 RNAiMAX and siRNA concentrations 25-200nM. *[*=p<0.05, **=p<0.01]*

4.4.7.2 Investigating cell-cycle dynamics

Following the proliferation experiments, potential differences in cell-cycle distribution were investigated. 4m cells were analysed using the ClickIT-EdU and FxCycleViolet DNA stain kit to quantify cell-cycle phase proportions and compare *ZFHX4* KOs with corresponding WTs. Cells were pulsed with 10μM EdU for 1 hour to label DNA synthesised during S-phase and then fixed and stained with FxCycle Violet (FxCVio) to measure total DNA content and distinguish sub-G1 (<2N), G1 (2N), S (2N-4N) and G2 (>4N) populations.

Only a single experiment (n=1 per condition) was performed, so results are presented descriptively as trends rather than subjected to statistical testing. Gating was established using a single-stained EdU control of CWT (Figure 4.32A) together with an unstained control (Figure 4.32B) to define EdU+ and EdU- single cell populations. This gating applied to NWT appeared consistent thus was applied on all samples irrespective of cell line. For cell-cycle phase assignment, CWT was used as a gating template (Figure 4.32C). Gates 1-4 were drawn based on DNA-FxCVio histogram peaks and expected linear distribution, before combining with EdU status using Boolean logic. EdU+ cells in gate3 were defined S-phase cells as EdU incorporation occurs only during DNA synthesis, while EdU- cells in all other gates were defined with respect to Figure 4.32C. This gating on CWT was applied to all samples, To ensure comparability, gate widths were kept identical across conditions but were anchored to the major peaks in each sample to allow for minor shifts in EdU–AF488 and DNA–FxCycle Violet signals. EdU median fluorescence intensity (MFI) was used as a surrogate marker for replication activity, as EdU intensity and uptake will be proportional to DNA replication rates.

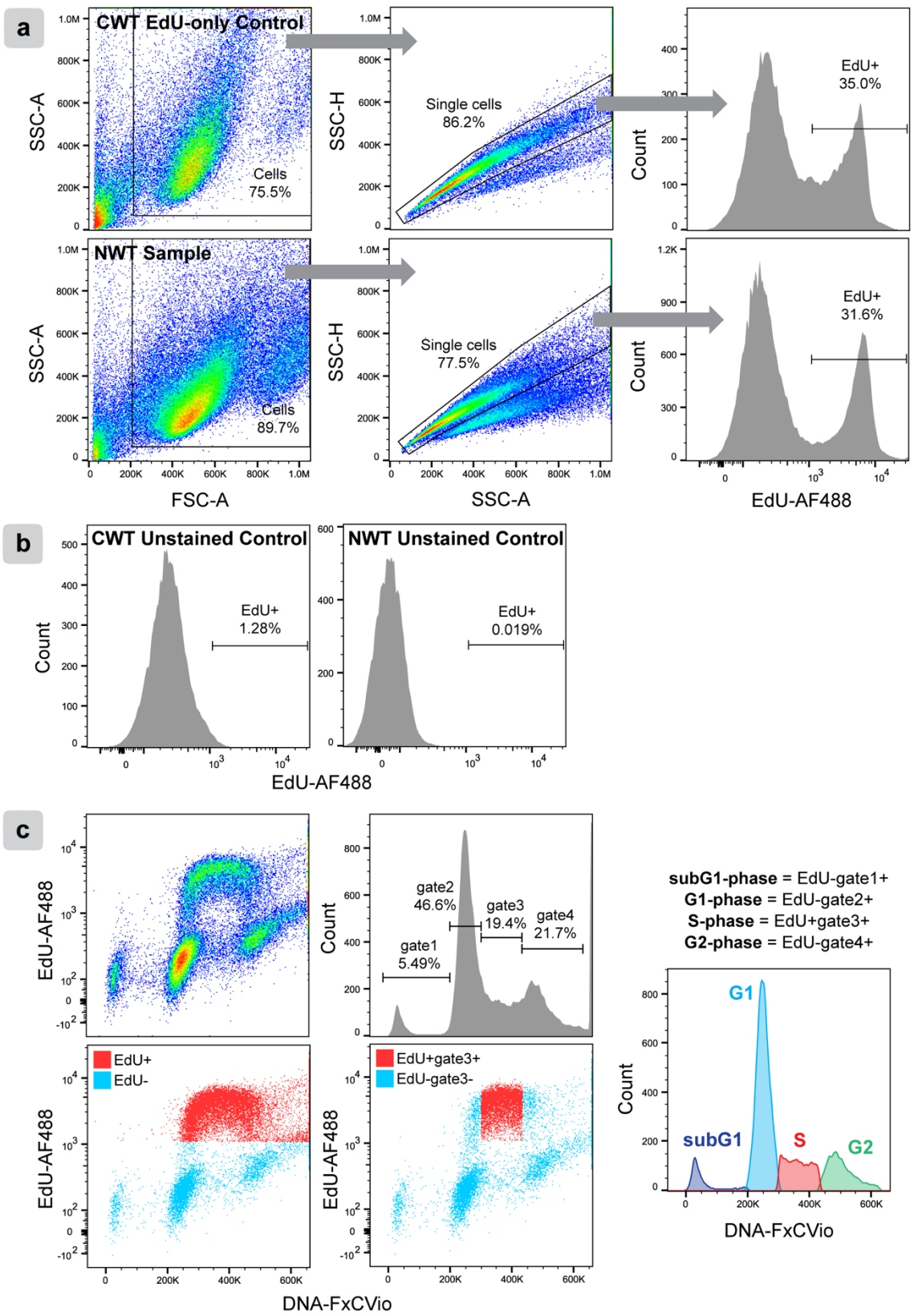


Figure 4.32 | Cell cycle gating strategy, using CWT and controls alongside confirmation of consistency with NWT. (A) Gating of single cells and EdU positivity on AF488 fluorophore using a single-stained EdU (CWT) control. NWT is shown with the same gating to show consistency between cell lines. This was applied on all samples. **(B)** Unstained controls of CWT and NWT that aided EdU+ gating above. **(C)** CWT density plots of EdU-AF488 by DNA-FxCVio. Plots show a Boolean combination strategy to obtain cell cycle phases based on DNA-FxCVio gate and EdU positivity.

For COLO320DM, proportions of cells in specific phases appeared to be variable (Figure 4.33A). Normalising these proportions to those observed in CWT, CKO and CC9 exhibited more cells in subG1 by 1.99-fold and 2-34-fold respectively, indicative of potentially more apoptotic processes and elevated cell death. Furthermore, CKO (1.59-fold) and CC9 (1.19-fold) were the only samples to display fold reduction in S-phase cells respectively, concluding that these KOs had reduced cycling capacity and proliferation relative to CWT. This was also reflected in EdU MFI shifts in S-phase cells (Figure 4.33B), showing 1.47-fold (CKO) 2.14-fold (CC9) reduced incorporation of EdU compared to CWT suggesting altered replication dynamics and slower DNA replication. These trends are consistent with 2m proliferation data, where CC9 was markedly much slower growing in culture soon after CRISPR. Conversely, CB5, CE5, CF6 and CF9 had between 1.13 and 1.44-fold more cells in S-phase which were incorporating between 1.41 and 1.92-fold more EdU and thus had faster DNA replication and enhanced proliferation. CF2 displayed an interesting phenotype of 1.2-fold more cells in G1 but high EdU incorporation, suggesting fewer cycling cells overall but those that did replicate did so robustly.

For NCIH716, NKO and NC6 showed increased fractions of cells in G2 (1.61- and 1.38-fold respectively), consistent with possible G2 checkpoint activation. NKO cells incorporated 1.2-fold more EdU per S-phase cell whereas NC6 demonstrated 1.25-fold less and thus a reduced replication rate. In contrast, ND5 and NG7 exhibited a striking >2.5-fold increase in subG1 events with >1.5-fold fewer S-phase cells representative of extensive cell stress and death with loss of proliferative capacity, something also reflected in 2m cell proliferation data. NG7 also demonstrated a 1.41-fold decrease in EdU intensity providing further evidence of weak replication rates of S-phase cells. NC8 and NE5 appeared to have similar cell-cycle progression to NWT.

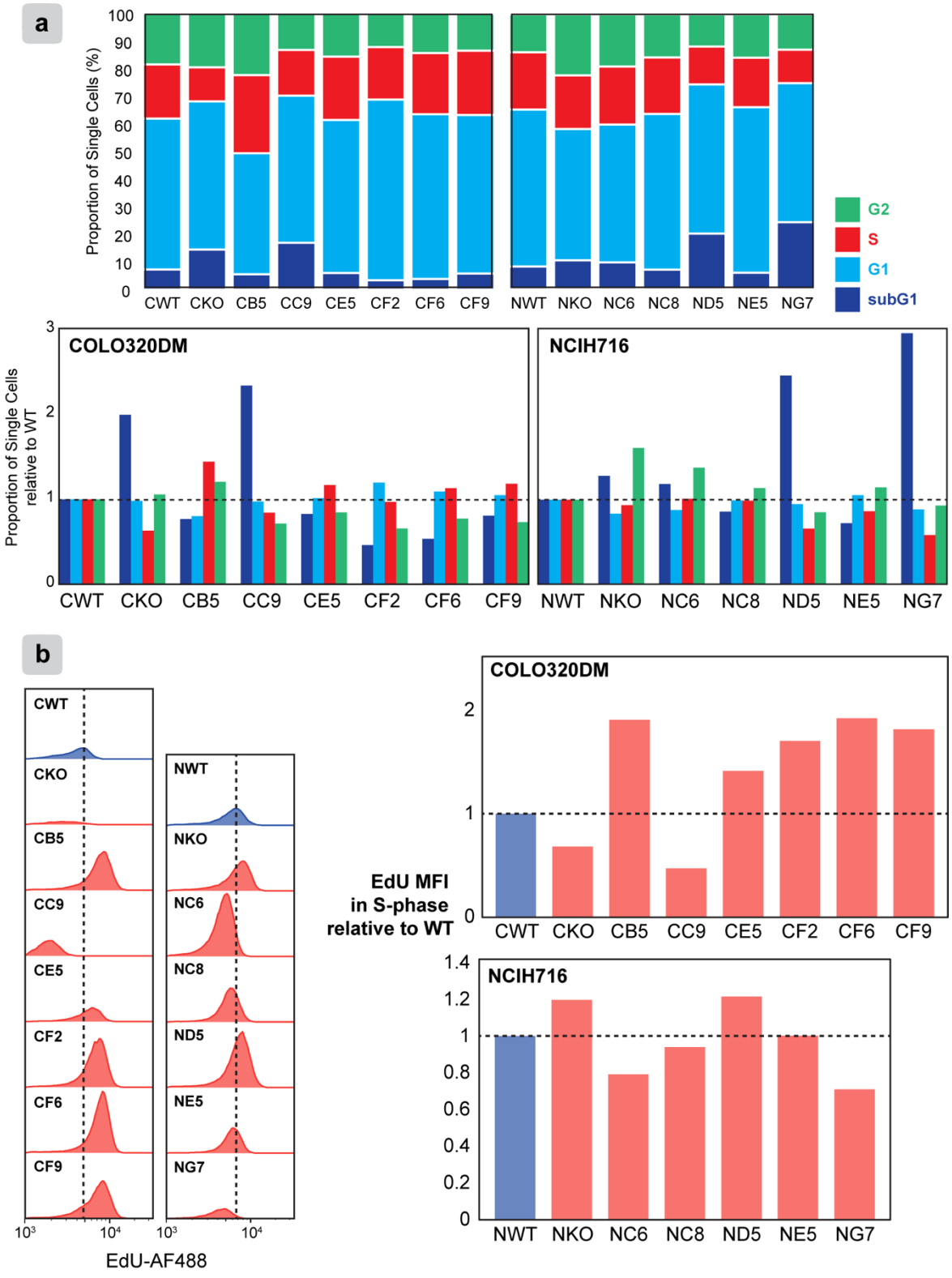


Figure 4.33 | Cell-cycle phase quantification in ZFHX4 WT and KO cells. (A) Proportions of cells in subG1, G1, S and G2 phases summarised into a stacked bar or separately following normalisation to respective WTs. **(B)** Density plots and bar charts representing EdU median fluorescence intensity in EdU+ S-phase cells. Dashed lines highlight comparisons to WTs.

These findings highlight potential roles for ZFHX4 in regulating proliferation, replication dynamics and cell-cycle progression. However, interpretation is limited by providing only one experimental repeat which prevented statistical validation. Moreover, the heterogeneous phenotypes observed across KOs hint possible clonal variation confounding attribution solely to ZFHX4 perturbations. These data suggest *ZFHX4*$^{-/-}$ can both impair and, in some contexts, enhance proliferative behaviour but additional technical replicates and functional validation will be required to distinguish consistent, ZFHX4-dependent effects from clonal heterogeneity. Furthermore, as this represents only a snapshot of the cell cycle in 4m cells, additional passage replicates are required since earlier experiments have shown marked differences in proliferation between 2m and 11m cells which would likely also translate into cell-cycle distribution differences.

4.4.7.3 Investigating the cell secretome

Alongside genomic and transcriptomic exploration, the secretome of each line was investigated for potential differences between WT and *ZFHX4* KOs under the hypothesis that ZFHX4 could regulate the expression of certain cytokines, considering interaction with the TME is important for both tumourigenicity and immunogenicity. This was hinted in organoid RNAseq (*CXCL8*, *FGF9*), NCIH716 RNAseq (*IL-1*, *IL-1b*) and COLO320DM RNAseq (*TNFα*) experiments but proteomic validation would offer greater insight, and this was employed through protein cytokine arrays and conditioned medium from each line. These were performed throughout my work on different passages (defined again by ‘m’ notation) and optimised, for example a total of five experiments were performed and three were cell-number controlled (7m-10m). This meant that despite seeding cells at the same density and extracting conditioned medium at the same time, some inherent growth differences between the different lines may have confounded the total protein measured in the media and thus the volume of media added to the arrays were normalised to each other

to account for this. Furthermore, each experiment performed on a KO included its respective WT control at the same passage. Each membrane was normalised within to generate normalised integrated area (NIA) ranging from 0 (not detected) to 1 (protein signal detected as high as the positive reference control signal) (Supplementary Methods 2.2).

Detecting up to 105 human cytokines, these arrays demonstrated that the most secreted proteins differed slightly between COLO320DM and NCIH716 cells but mostly consisted of around 10 proteins comprising angiogenic factors (Angiogenin, VEGFA) and non-immune related cytokines (Figure 4.34A). Furthermore, hierarchical clustering of the top 20 secreted proteins for both cell lines exhibited samples grouped by passage rather than biological replicate. For example, most 7m and 10m cells cluster together though in samples that don't cluster by passage they do cluster by biological replicate such as CC9 (6m, 7m) and CWT (4m and 6m). These results highlight the importance of including a matched WT control at the same passage, which helps mitigate passage-related batch effects and allows for reliable interpretation of cytokine secretion differences.

Therefore, to quantify secretome differences between WTs and KOs, delta NIA (dNIA) was calculated by subtracting passage-matched WT-NIA from KO-NIA (KO-WT) due to the difficulty of division when zeros are present. Thus dNIA values > 0 suggest the cytokine was secreted more in the KO comparative to WT and dNIA values < 0 suggest the inverse. Selecting the largest dNIA values (thresholding ±0.05 units equating to visually discernible differences) revealed predominant global downregulation of cytokines (n=18 COLO320DM, n=14 NCIH716) across most passage replicates (Figure 4.34B). These consisted mostly of pro-angiogenic factors such as Angiogenin and VEGFA and both angiogenic and ECM remodelling-related factors such as GDF-15, CD147, CXCL11 and IGFBP-2 across both cell lines in most KO cells. Examples of representative cytokine array membranes demonstrate the robustness of the NIA and dNIA approach to quantify these data as heatmaps closely reflect dot intensity as expected.

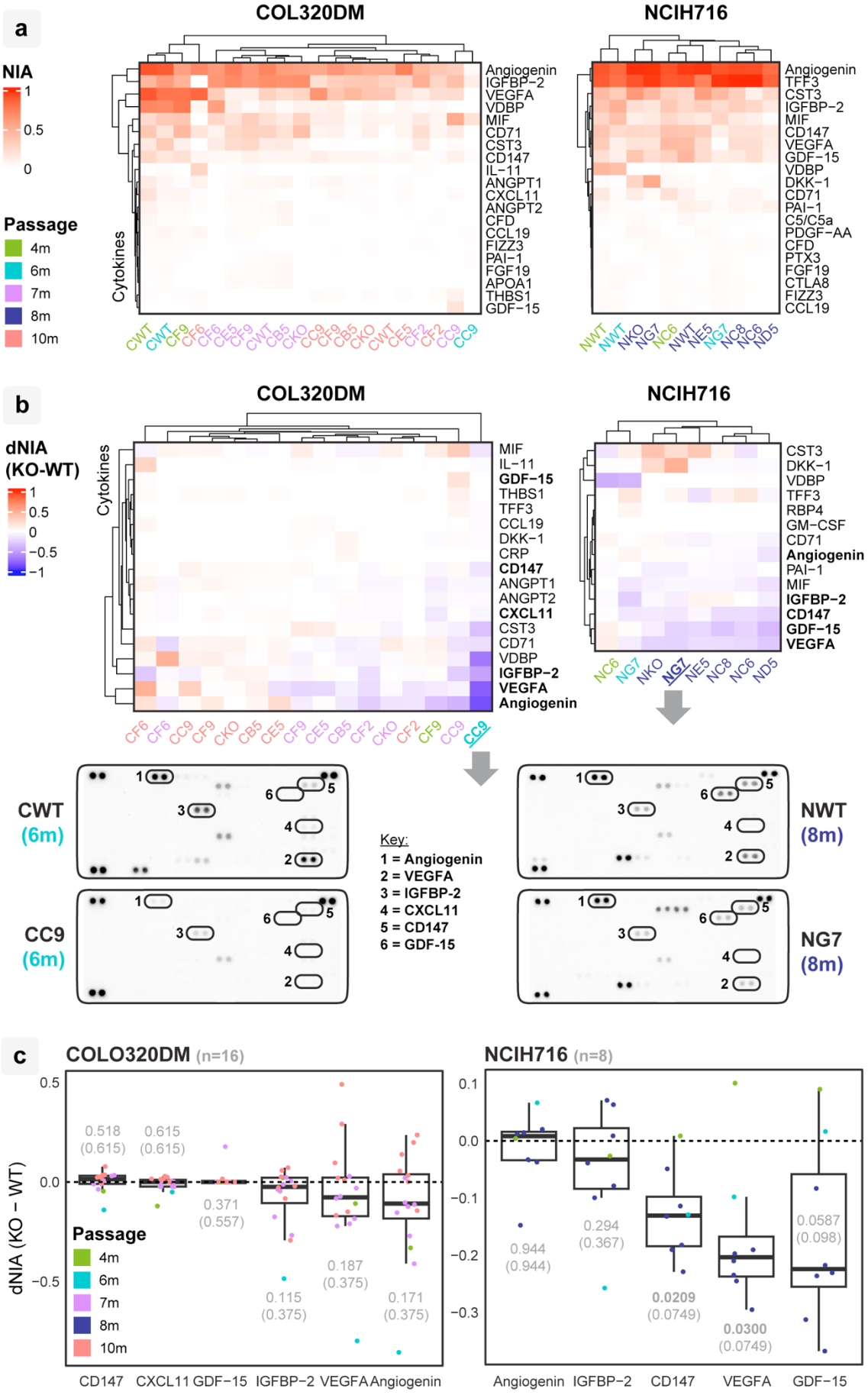

a
COL320DM
NCIH716
NIA
Passage
4m
6m
7m
8m
10m
Cytokines
Angiogenin
IGFBP−2
VEGFA
VDBP
MIF
CD71
CST3
CD147
IL−11
ANGPT1
CXCL11
ANGPT2
CFD
CCL19
FIZZ3
PAI−1
FGF19
APOA1
THBS1
GDF−15
TFF3
PDGF−AA
PTX3
CTLA8
C5/C5a
DKK−1
b
dNIA (KO−WT)
RBP4
GM−CSF
CRP
CWT (6m)
CC9 (6m)
NWT (8m)
NG7 (8m)
Key:
1 = Angiogenin
2 = VEGFA
3 = IGFBP-2
4 = CXCL11
5 = CD147
6 = GDF-15
c
COLO320DM (n=16)
NCIH716 (n=8)
dNIA (KO − WT)
0.518 (0.615)
0.615 (0.615)
0.371 (0.557)
0.115 (0.375)
0.187 (0.375)
0.171 (0.375)
0.944 (0.944)
0.294 (0.367)
0.0209 (0.0749)
0.0300 (0.0749)
0.0587 (0.098)
CD147 CXCL11 GDF−15 IGFBP−2 VEGFA Angiogenin
Angiogenin IGFBP−2 CD147 VEGFA GDF−15

Figure 4.34 | Quantification of cytokine arrays measuring secreted cytokines in *ZFHX4-WT* and *ZFHX4*$^{-/-}$ conditioned medium. (A) Heatmaps of the top 20 secreted cytokines in COLO320DM and NCIH716 samples, with rows of cytokines and columns of samples clustered by hierarchial clustering. Heatmap values represent normalised integrated area (NIA) of dot intensity from respective cytokine membranes. **(B)** Heatmaps of secreted cytokines which present with the largest differences to WT, quantified by delta NIA (KO-WT) with respective to passage-matched WT. Representative cytokine array membranes are shown with annotation of key cytokines. (C) Box plots of dNIA values ofinteresting cytokines, ordered by median. Overlaid dots represent passage and clones as biological ZFHX4^{KO} replicates. N values per box plot are designated by cell line title. Nominal P-values from wilcoxon testing are denoted by its respective cytokine, with adjusted P-values in brackets comparing median dNIA to 0.

To integrate data from five experiments across different passage replicates and clones, median and interquartile range of each protein were calculated, treating all passage replicates and clones as biological replicates of *ZFHX4*$^{-/-}$ (Figure 4.34C). For COLO320DM and NCIH716, the cytokines noted from the heatmaps were plotted with reference to a dNIA of zero (equivalent to WT cytokine levels) and overlaid with individual dNIA values coloured by passage to illustrate variation. Statistical differences from WT were assessed using the Wilcoxon signed-rank test to test whether the median dNIA differed from zero. In COLO320DM *ZFHX4* KOs, Angiogenin was the most downregulated cytokine, displaying a 0.11 median dNIA reduction compared to WT, though this was not significant by Wilcoxon testing. VEGFA (-0.078 median units) and IGFBP-2 (-0.025 median units) also demonstrated the same trends but these were also not significantly different to WT. Interestingly, these trends were also observed in NCIH716 *ZFHX4* KOs. For example, VEGFA showed a nominal significant decrease in secretion from KOs compared to WTs (-0.20 median units, p=0.03) as well as CD147 (-0.13 median units, p=0.02) but these differences didn't remain significant after FDR correction (adj.p=0.0749).

In summary, despite variation between passages, there are clear trends of reduced secretion of pro-angiogenic and ECM remodelling-related factors in ZFHX4 KOs supporting a potential role for ZFHX4 in regulating angiogenesis and ECM remodelling.

4.4.7.4 Investigating angiogenesis

Following a trend of reduced pro-angiogenic factors in *ZFHX4*$^{-/-}$ media, *in vitro* tube formation assays (IVTFAs) were performed using HUVEC endothelial cells (Figure 4.35A) cultured in different control and conditioned mediums following a published protocol (Arnaoutova and Kleinman, 2010). When grown on basement membrane such as Matrigel for up to 12 hours, HUVECs spontaneously organise into interconnected tubular networks and mimic capillary-like structures. IVTFAs therefore serve as a proxy for angiogenic potential, specifically reflecting the later stages of angiogenesis (Figure 4.35B). These stages include migration and movement across the matrix, cell-cell alignment and connection through junction and branch formation and ultimately network stabilisation with the development of meshes and loops.

To quantify these processes, HUVECs were cultured for 12 hours on Matrigel in specialised non-meniscus-forming 96-well plates and images were analysed using the Angiogenesis Analyser macro (Carpentier et al., 2020). This tool quantifies over 20 parameters of angiogenesis, including the gold-standard readouts of total tube length and mesh formation, based on viable-cell GFP signal following Calcein AM staining (Figure 4.35B). Although initial analyses on cytokine arrays didn't reveal statistically significant differences, there was a consistent trend towards reduced secretion of angiogenic factors such as VEGFA in *ZFHX4*$^{-/-}$-cell CM. Furthermore, the cytokine array was limited to 102 cytokines and so other potential angiogenesis-related factors weren't profiled. To test whether cytokine-related differences functionally impact angiogenesis, CM from all WT and KO lines was subsequently applied to IVTFAs under the hypothesis that *ZFHX4*$^{-/-}$ CM would impair HUVEC tube formation compared to WT CM.

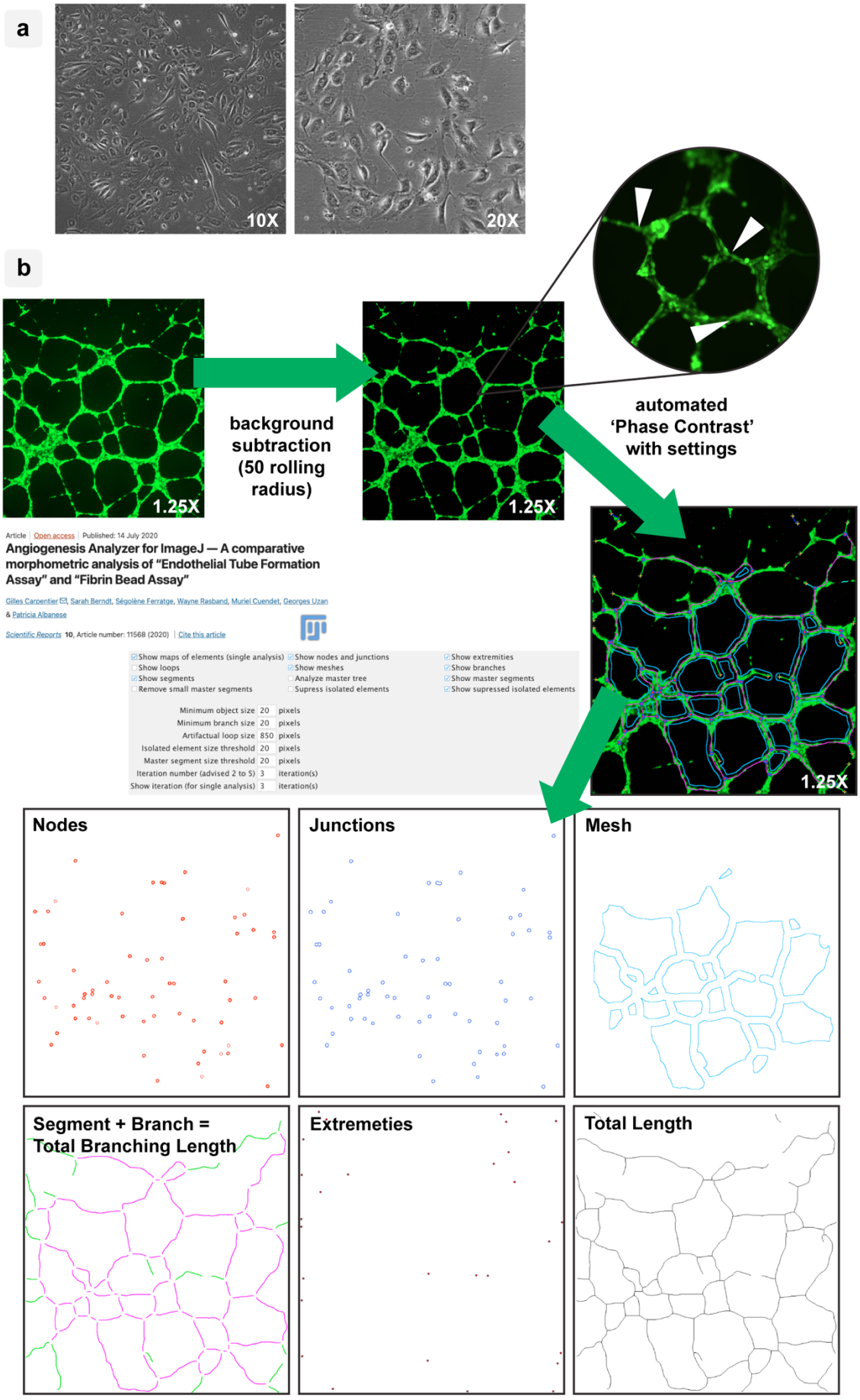

a
10X
20X
b
1.25X
background subtraction (50 rolling radius)
1.25X
automated 'Phase Contrast' with settings
1.25X
Article | Open access | Published: 14 July 2020
Angiogenesis Analyzer for ImageJ — A comparative morphometric analysis of "Endothelial Tube Formation Assay" and "Fibrin Bead Assay"
Gilles Carpentier, Sarah Berndt, Ségolène Ferratge, Wayne Rasband, Muriel Cuendet, Georges Uzan & Patricia Albanese
Scientific Reports 10, Article number: 11568 (2020) | Cite this article
Show maps of elements (single analysis)
Show loops
Show segments
Remove small master segments
Show nodes and junctions
Show meshes
Analyze master tree
Supress isolated elements
Show extremities
Show branches
Show master segments
Show supressed isolated elements
Minimum object size 20 pixels
Minimum branch size 20 pixels
Artifactual loop size 850 pixels
Isolated element size threshold 20 pixels
Master segment size threshold 20 pixels
Iteration number (advised 2 to 5) 3 iteration(s)
Show iteration (for single analysis) 3 iteration(s)
Nodes
Junctions
Mesh
Segment + Branch = Total Branching Length
Extremeties
Total Length

Figure 4.35 | HUVEC cells utilised for in-vitro tube formation assays (IVTFAs) and analysis of angiogenic biology using the Angiogenesis Analyser macro. (A) Morphology of HUVECs grown on plastic at P8. **(B)** Morphology of HUVECs following 12 hours grown on Matrigel during an IVTFA. The computational workflow of IVTFA image analysis using the Angiogenesis Analyser macro on Fiji is depicted with an example image. White arrows indicate formation of tube-like structures.

A total of three independent experiments were performed, utilising CM from 8m, 9m and 10m-cells in different ways with variable n numbers based on sample availability at the time. Across experiments, different CM types were employed including 24-hour and 48-hour conditioning and CM prepared using either a basal or complete base. HUVECs are usually cultured in a supplement-rich media (here referred to as complete EGM) containing proteins such as bFGF and VEGFA. This was consistently used as a positive control throughout all experiments, delineating maximal angiogenesis. In contrast, base media with only 5% FBS as supplement (here referred to as basal EBM) was used as a negative control throughout all experiments, delineating baseline angiogenesis. Other controls included bFGF and VEGFA[165] supplementation in basal EBM (positive controls) as well as the pleiotropic angiogenesis inhibitor Suramin in complete EGM and bFGF/VEGFA targeted removals from complete EGM (negative controls). For every experiment, most of these controls were run additionally to a basal-EBM and complete-EGM non-cell conditioning media that was incubated for the same time. Within the IVTFA, a ‘sample’ is classed as one image taken on one well of HUVECs incubated in a specific CM during the IVTFA, giving many data points.

Given the variability across experimental setup and the large number of parameters generated by the Angiogenesis Analyser, my first step was to identify a reduced set of key metrics to describe angiogenesis. Taking all data (n=227), PCA was performed and plotted to visualise an angiogenic landscape (Figure 4.36A). Variable loadings (i.e. the parameters obtained from the macro) were overlaid, with arrow length representing the strength of each parameter’s contribution to variance and arrow direction reflecting its orientation in PCA space.

PC1 captured a global angiogenic strength signal heavily influenced by measures spanning junctions, nodes, total length, complete segments and mesh annotations. PC2 variation was less pronounced but seemed to be largely driven by branching-related patterns. For simplicity and interpretability, the strongest PC1 contributors were selected as representative metric of angiogenic potential and grouped into three categories: Network Connectivity (number of junctions; reflecting network complexity), Network Extent (total length; reflecting amount of structure formed) and Network Maturation (both number of meshes and mesh area; reflecting network stability and maturity).

Overlaying media controls on the same PCA, it was clear that different media types occupied different PCA space, for example Suramin-treated conditions (grouped irrespective of concentration) occupied the PCA space of least angiogenic potential as expected. Similarly, when overlaying experimental samples, no clear segregation emerged based on ZFHX4 status or cell line, but CM from 48-hour conditioning and those prepared in basal EBM tended to cluster toward reduced angiogenic potential (Figure 4.36B).

To validate this framework, basal EBM and complete EGM controls were directly compared across experiments. As anticipated complete EGM consistently induced a significantly stronger angiogenic phenotype than basal EBM across all metrics (all $p<0.001$, unpaired t-test or Wilcoxon; Figure 4.36C). Notably displaying the highest fold difference, mesh number and network maturity was 3.84-fold more in complete EGM conditions. This suggests that the vast array of supplement in complete EGM is required for optimal HUVEC function and maximal angiogenesis. The images acquired aligned well with these findings, reinforcing the robustness of this analysis pipeline and the suitability of the Angiogenesis Analyser macro for high-throughput quantification of angiogenic behaviour.

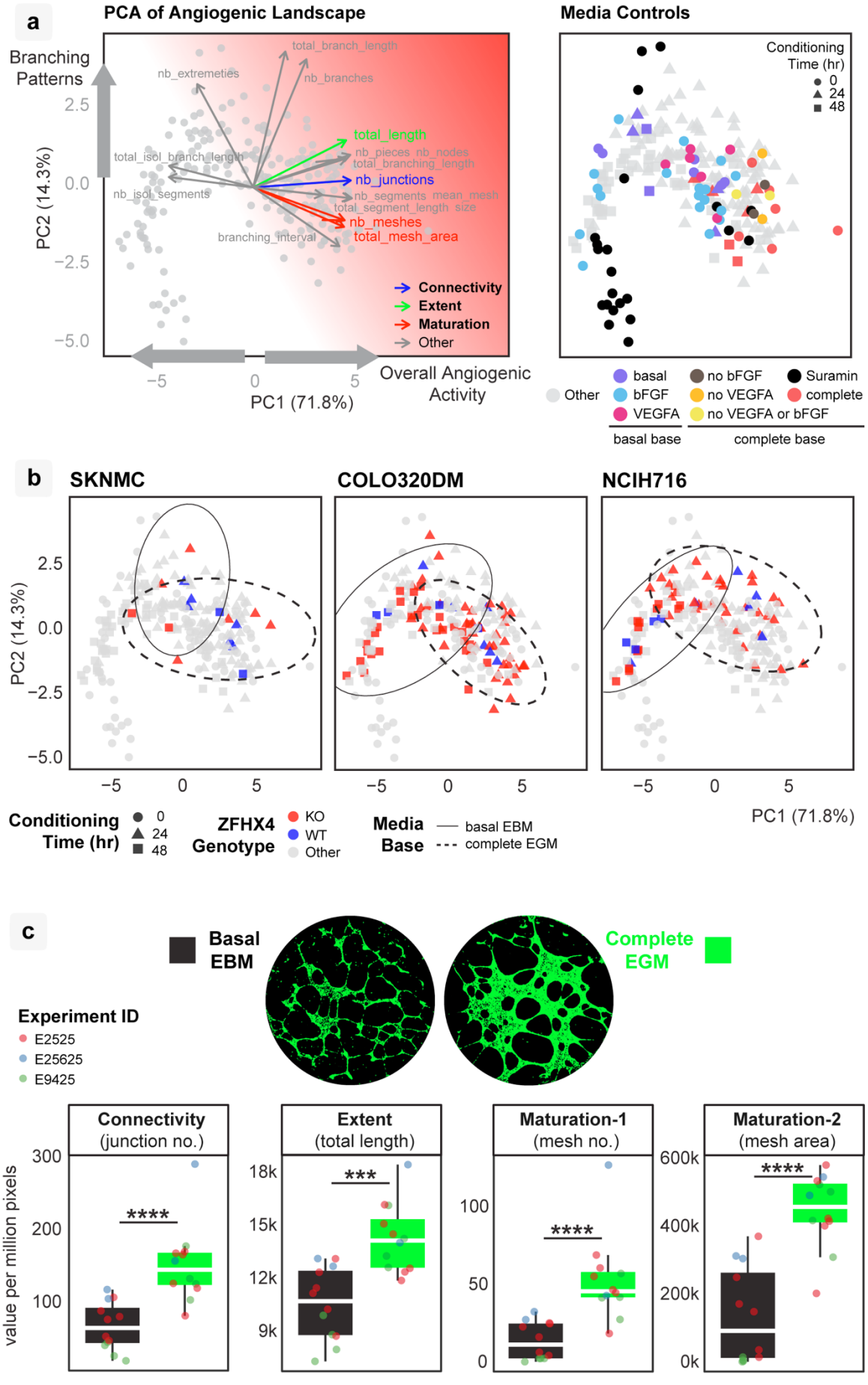

a
PCA of Angiogenic Landscape
Branching Patterns
PC2 (14.3%)
PC1 (71.8%)
Overall Angiogenic Activity
total_branch_length
nb_extremeties
nb_branches
total_length
total_isol_branch_length
nb_pieces nb_nodes
total_branching_length
nb_junctions
nb_isol_segments
nb_segments mean_mesh
total_segment_length size
nb_meshes
branching_interval
total_mesh_area
Connectivity
Extent
Maturation
Other
Media Controls
Conditioning Time (hr)
0
24
48
Other
basal
bFGF
VEGFA
no bFGF
no VEGFA
no VEGFA or bFGF
Suramin
complete
basal base
complete base
b
SKNMC
COLO320DM
NCIH716
PC2 (14.3%)
PC1 (71.8%)
Conditioning Time (hr)
ZFHX4 Genotype
KO
WT
Other
Media Base
basal EBM
complete EGM
c
Basal EBM
Complete EGM
Experiment ID
E2525
E25625
E9425
value per million pixels
Connectivity (junction no.)
Extent (total length)
Maturation-1 (mesh no.)
Maturation-2 (mesh area)
****
***
****
****

Figure 4.36 | The Angiogenic Landscape of HUVECs following multiple IVTFA experiments. **(A)** PCA plot of all IVTFA data points (n=227) irrespective of experiment or sample or replicate. Variable loadings were plotted as vectors, with arrows for parameters mapping to general angiogenesis metrics emphasised to visualise which features contributed the most to sample separation along the principal components. Media control types were overlaid on the same PCA and grouping was irrespective of concentration if applicable. **(B)** The same PCA was overlaid with ZFHX4 genotype of respective samples, and media base was encircled with ellipses. **(C)** Basal EBM and complete EGM controls in all three experiments, comparing angiogenic metrics with t-testing (Extent metric) or wilcoxon testing (all other metrics) based on Shapiro-Wilkes normality. Example images of one replicate at 1.25X are shown. Datapoints are overlaid and coloured by experiment ID *[***=p<0.001, ****=p<0.0001]*

Despite no clear segregation of samples based on *ZFHX4* status, one experiment using CM from 9m cells was further interrogated to compare *ZFHX4*-WT and *ZFHX4*$^{-/-}$ lines. Specifically, we focused on basal EBM-based CM in this experiment, as the predicted reduced VEGFA secretion by *ZFHX4* KOs was expected to have the greatest impact under these conditions. This is comparative to complete EGM-based CM already containing VEGFA which may mask or dilute any biological differences. Furthermore, this experiment had replicates for 24hr and 48hr conditioning.

To process this data, background angiogenesis from non-cell conditioned basal EBM (matched by timepoint) was first subtracted from each replicate. These background-corrected values were then normalised to the corresponding WT mean by division, setting WT to 1. Angiogenic potential (ΔΔ unit) values >1 indicated enhanced angiogenesis relative to WT, while values <1 indicated reduced angiogenesis.

Across the four angiogenic metrics, most ZFHX4 KOs showed a general trend toward reduced angiogenic potential compared with WT, in contrast to some of the COLO320DM monoKOs such as CB5 and CE5. However, none of these differences reached statistical significance (Figure 4.37). This was likely due to considerable heterogeneity observed within replicates of a sample (mostly n=4) as denoted by large SEM error bars. Furthermore, some error bars such as those seen with SWT and SKO demonstrate angiogenesis metrics that were close to baseline EBM controls and thus error was magnified upon normalisation.

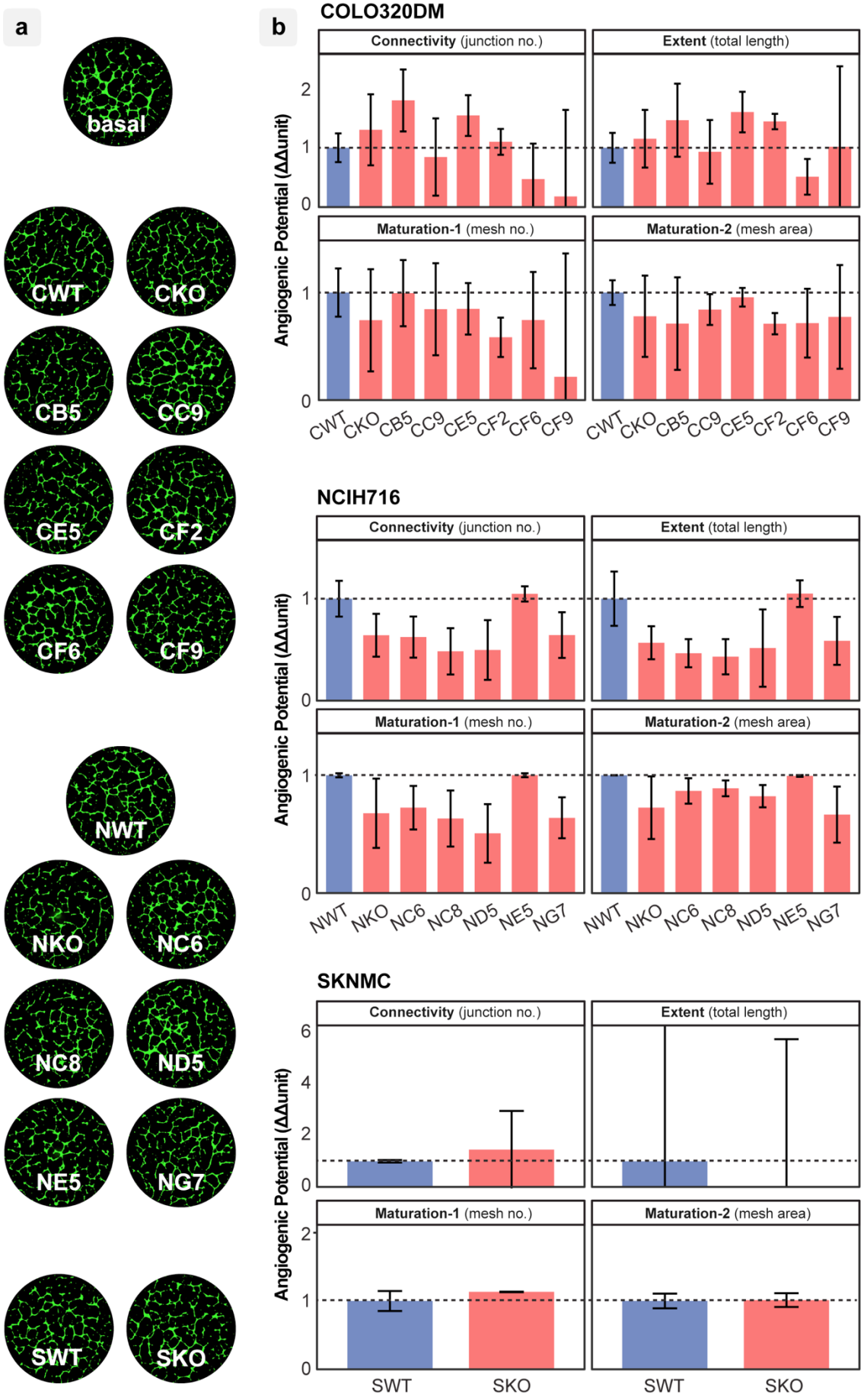

a
basal
CWT
CKO
CB5
CC9
CE5
CF2
CF6
CF9
NWT
NKO
NC6
NC8
ND5
NE5
NG7
SWT
SKO
b
COLO320DM
Connectivity (junction no.)
Extent (total length)
Maturation-1 (mesh no.)
Maturation-2 (mesh area)
Angiogenic Potential (ΔΔunit)
CWT CKO CB5 CC9 CE5 CF2 CF6 CF9
NCIH716
NWT NKO NC6 NC8 ND5 NE5 NG7
SKNMC
SWT SKO

Figure 4.37 | Angiogenic potential of *ZFHX4-WT* and *ZFHX4*$^{-/-}$ cell lines. (A) Example images of HUVECs post IVTFA cultured in each sample are shown with respect to one replicate from the 24hr conditioning group. **(B)** Box plots show mean with error bars denoting SEM. Most samples were n=4 (2 for 24hr and 2 for 48hr). Angiogenesis metric values were background corrected using baseline EBM basal controls and normalised to WT mean by division, presenting y axes as ratios, equivalent to angiogenic potential (ΔΔunit). Statistical testing comparing KOs to WT within each metric employed t-testing or wilcoxon testing based on Shapiro-Wilkes normality.

To assess how sensitive IVTFAs are in detecting biological differences arising from altered protein composition in CM, I included several additional controls across experiments. For quantitation, all values were normalised by division to zero-concentration controls with respect to experiment (angiogenic potential, Δ unit). bFGF was tested as a classical positive control, but across three increasing concentrations no visual or quantitative enhancement of angiogenesis was observed compared with 0ng/ml controls (Figure 4.38A). This was further supported by no significant differences in any angiogenic metric compared to the 0ng/ml control upon statistical testing (Figure 4.38B). Furthermore, a similar outcome was observed with VEGFA that also demonstrated no statistically significant differences. In fact, higher concentrations of either pro-angiogenic factor unexpectedly trended towards reduced angiogenic potential, for example 10ng/ml VEGFA prompted between 1.45-2.32-fold reductions across all angiogenic metrics and the reduction in network connectivity reached significance (p=0.0247, unpaired t-test) but without further correction (p=0.074). Furthermore, preparing complete EGM media without VEGFA, bFGF or both again showed no statistically significant differences compared to standard complete EGM media.

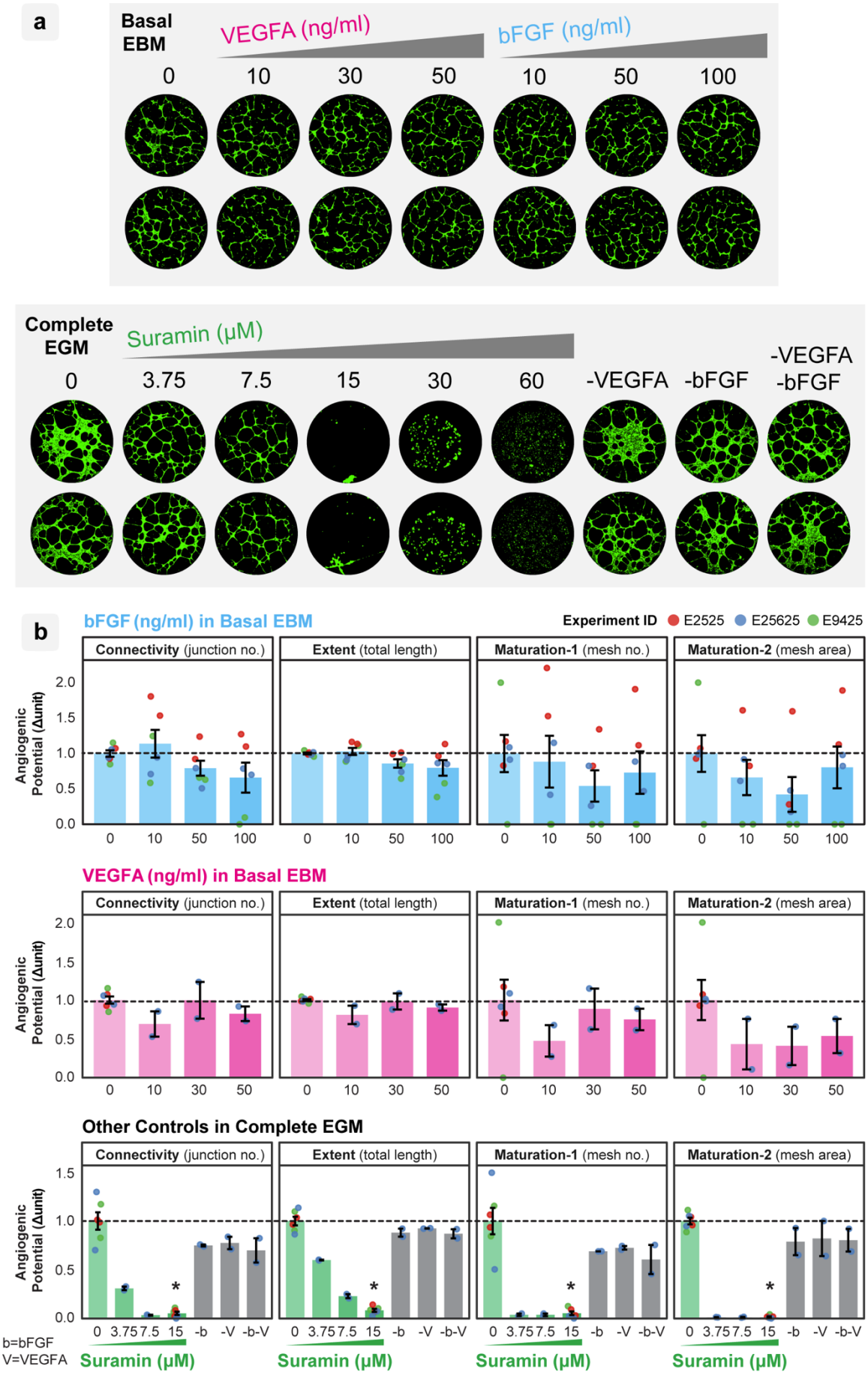

a
Basal EBM
VEGFA (ng/ml)
bFGF (ng/ml)
0
10
30
50
10
50
100
Complete EGM
Suramin (μM)
0
3.75
7.5
15
30
60
-VEGFA
-bFGF
-VEGFA -bFGF
b
bFGF (ng/ml) in Basal EBM
Experiment ID E2525 E25625 E9425
Connectivity (junction no.)
Extent (total length)
Maturation-1 (mesh no.)
Maturation-2 (mesh area)
Angiogenic Potential (Δunit)
VEGFA (ng/ml) in Basal EBM
Other Controls in Complete EGM
b=bFGF
V=VEGFA
Suramin (μM)

Figure 4.38 | Angiogenic potential of pro-angiogenic factors and angiogenesis inhibitors in IVTFAs. (A) Images are with respect to one experiment, and replicates (n=2) are presented vertically to each other. The first set of images refer to conditions in basal EBM-base and the second set refer to conditions in complete EGM-base media. **(B)** Bar plots demonstrating angiogenic potential quantification across four metrics in all conditions from three experiments. Values were normalised by division of the 0-concentration respective controls within each experiment, to generate angiogenic potential on the y-axis. Statistical testing comparing conditions to 0-concentration controls within each metric employed t-testing or wilcoxon testing based on Shapiro-Wilkes normality. Asterisks denote the BH-corrected p-value. Overlaid dots are coloured by experiment. *[*=p<0.05]*

In contrast, the angiogenesis inhibitor Suramin displayed striking and dose-dependent effects that were even observable at microscopy level. As expected, angiogenesis was reduced even at the lowest 3.75μM concentration, demonstrated by reduced network connectivity and extent and ablation in mesh formation and network maturity despite losing significance due to n=2 replicates. Complete ablation in angiogenesis was achieved at 15μM, with significant reductions in all angiogenic metrics (n=6, p=0.013, wilcoxon after correction) compared to 0μM (n=6). Clear differences were noted at the visual level, where cell migration had occurred but was disordered as displayed by cell clumping and disrupted tube formation. At higher concentrations (30-60μM) Suramin abolished cell migration altogether, exhibiting viable GFP-positive cells remaining clustered at their plating sites and mimicking growth on plastic. These samples confounded the macro-based analysis due to detection of artificial networks and thus were not plotted in Figure 4.38B. Overall, these findings suggest that IVTFAs are highly sensitive to angiogenic inhibition but not to subtle differences in the presence or concentration of specific pro-angiogenic proteins. Furthermore, this likely accounts for the minimal and non-significant differences observed between ZFHX4 WT and KO lines as subtleties in VEGFA secretion into CM wouldn't have been sufficient to produce detectable effects in this assay.

4.4.7.5 Investigating cell migration and invasion

Migration and invasion are key functional phenotypes associated with tumourigenicity. Furthermore, *ZFHX4* KD has already been shown to reduce migration and invasion in oesophageal cancer cell lines (Qing et al., 2017). However, through employing Transwell assays and 3D Matrigel Droplet assays, it was observed that my cell line models were

minimally migratory or invasive at baseline and so interpretation relating to ZFHX4 was challenging. This is discussed further in Supplementary Note 4.3.

4.4.8 Integrative analysis of ZFHX4 in clinical cohorts

The difficulty in concluding a definitive ZFHX4 phenotype following *in vitro* investigation across all *ZFHX4*$^{-/-}$ models prompted further *in silico* analyses to integrate experimental results with clinical datasets and improve phenotype prediction. Leveraging four large colon cancer cohorts with matched RNAseq and WES data, along with my RNAseq data of organoid, SKNMC, COLO320DM and NCIH716 *ZFHX4*$^{-/-}$ experiments, a comprehensive integrative analysis was conducted. Initially, to explore the unclear functional consequences of missense mutations in *ZFHX4*, expression levels of *ZFHX4* mRNA were compared across *ZFHX4* mutation subtypes (WT, missense, non-missense/truncating, and >1 mutation) within both nHM and HM contexts across all four cohorts (Figure 4.39). Patients presenting with >1 mutation in *ZFHX4* were segregated separately, as these could be both truncating and non-truncating. Wilcoxon testing revealed no significant differences in *ZFHX4* expression between WT and any mutation subgroup. While non-missense/truncating mutations would be expected to reduce expression, the low frequency of these mutations particularly in nHM samples likely limited statistical power. In contrast, the predominance of missense mutations and their lack of association with reduced expression suggest they do not cause loss of function through transcript depletion. Instead, they may exert functional effects through alternative mechanisms, such as altered protein-protein or protein-DNA interactions. To note, expression was gene-level quantified as transcript-level isoform quantification was not available for all cohorts. As discussed in Section 4.4.4 this should be interpreted with caution and further *in vitro* work is required to demonstrate the functional impact of *ZFHX4* mutations on expression.

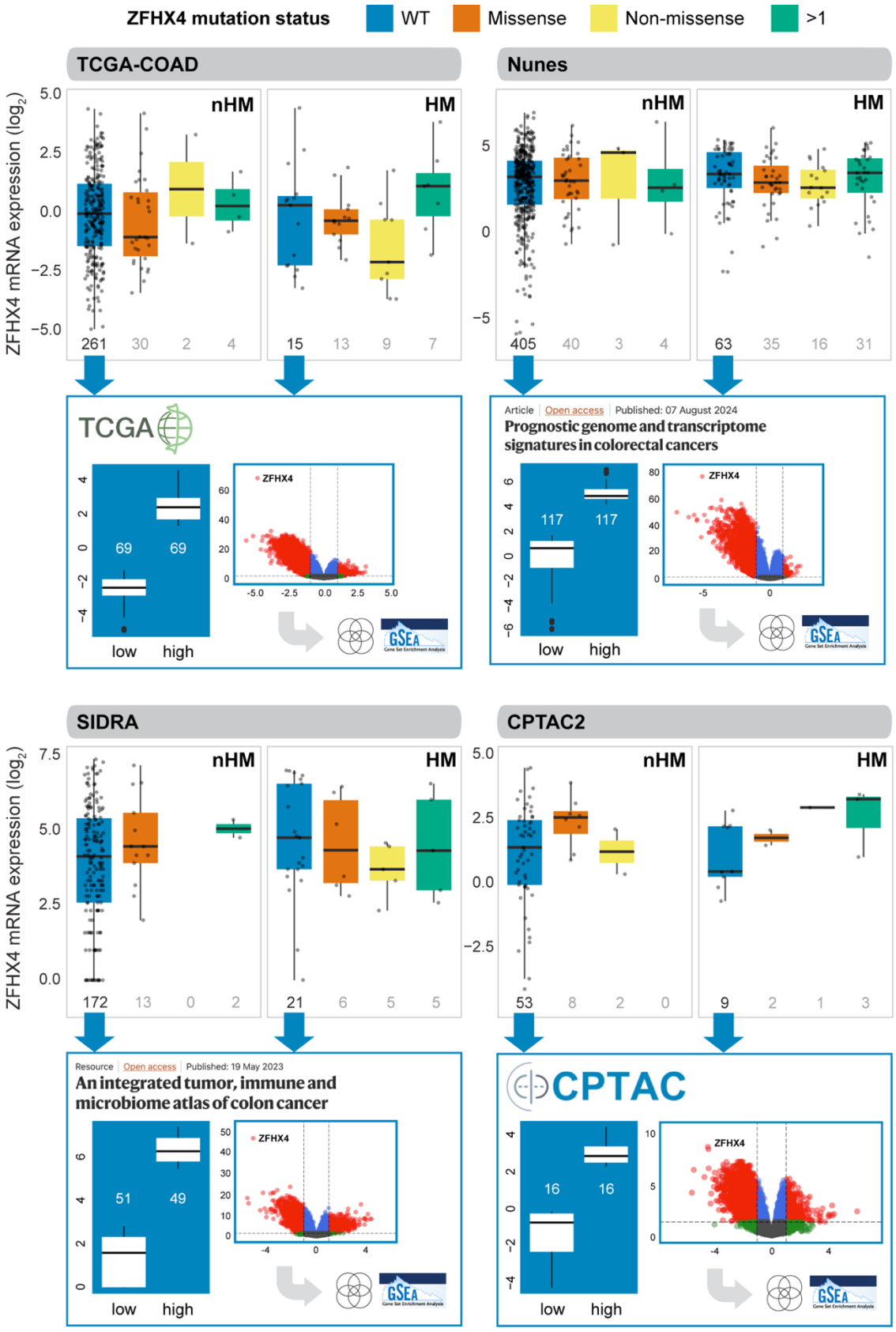


Figure 4.39 | ZFHX4 expression and mutation across four colon cancer cohorts, with schematics demonstrating method of differential analysis. All box plots represent medians and are annotated with N numbers. Mutation subgroup box plots are overlaid with all points which represent individual patients. ZFHX4 expression quartile box plots are represented with the blue background. Volcano plot x axes are log_2FC and y axes are $-log_{10}$(adj.p-value).

Given the potential functional differences between *ZFHX4* mutations (associated with immunogenicity) and expression (associated with tumourigenicity), together with the prognostic significance of *ZFHX4* expression in colon cancer and its links to oncogenic pathways described in this chapter, I focused on *ZFHX4*^{wt} patients only (irrespective of HM status) and stratified them into quartiles based on *ZFHX4* expression (Figure 4.39). The top and bottom quartiles were selected for differential expression analysis followed by pathway enrichment to identify genes and biological programs potentially regulated by non-mutated ZFHX4. This analysis was performed independently within each cohort, followed by cross-cohort comparisons to identify high-confidence ZFHX4-associated genes and pathways with consistent signals across datasets.

Cross-cohort concordance and correlation were subsequently evaluated to assess the consistency of ZFHX4-associated biology across all four cohorts. Differential gene expression (DEG) analyses comparing *ZFHX4*^{high} and *ZFHX4*^{low} samples were performed separately for each cohort and genes were ranked by t-statistic (t-stat) as a more reliable measure that takes both effect size and variance into account comparative to log_2FC. These same t-stat-ranked lists were then used for GSEA with MSigDB-derived gene sets and consequent pathway analysis, achieving pathway ranked lists by NES score. As a side note, ranking of genes and pathways were kept consistent in this chapter, so that experimental *ZFHX4*$^{-/-}$ volcano plots and pathway enrichment bar plots can be compared with clinical *ZFHX4*^{low} expression. Additionally, DEG and pathway enrichment analyses were structured so that directionality was preserved, for example genes with t-stat<0 correspond to pathways with NES<0, simplifying interpretation.

The concordance 'consistency metric' was firstly measured by taking the positive or negative ranked genes (t-stat) and pathways (NES) per cohort, filtering for significance (adj.p<0.05) and then counting overlap of matched genes or pathways across cohorts (Figure 4.40A/B). Genes significantly upregulated in *ZFHX4*^{high} samples (t-stat<0) and *ZFHX4*^{low} samples (t-stat>0) both revealed substantial overlap across cohorts, with over

70% DEGs shared in one cohort with at least one other cohort regardless of directionality. Pathways enriched in *ZFHX4*^{high} groups (NES<0) and *ZFHX4*^{low} groups (NES>0) were also compared, demonstrating even greater concordance with >75% enriched pathways shared across cohorts irrespective of directionality.

The correlation consistency metric was then used to quantify global similarity, utilising the whole set of gene and pathway lists irrespective of significance (Figure 4.40A/C). Only genes with t-stat values (n=14579) or pathways with NES values (n=5502) across all four cohorts were retained, performing Pearson correlation analyses based on the 'pairwise complete observations' function. This revealed strong positive correlations across all cohort pairs, with Pearson correlation coefficients (r)>0.7 and high statistical significance ($p<2.2x10^{-16}$) for all gene-level and pathway-level comparisons. Together, these findings demonstrate a high degree of consistency in ZFHX4-associated biology across four independent colon cancer cohorts, supporting their combined use to explore respective biology while acknowledging downstream analyses remain correlative rather than causal.

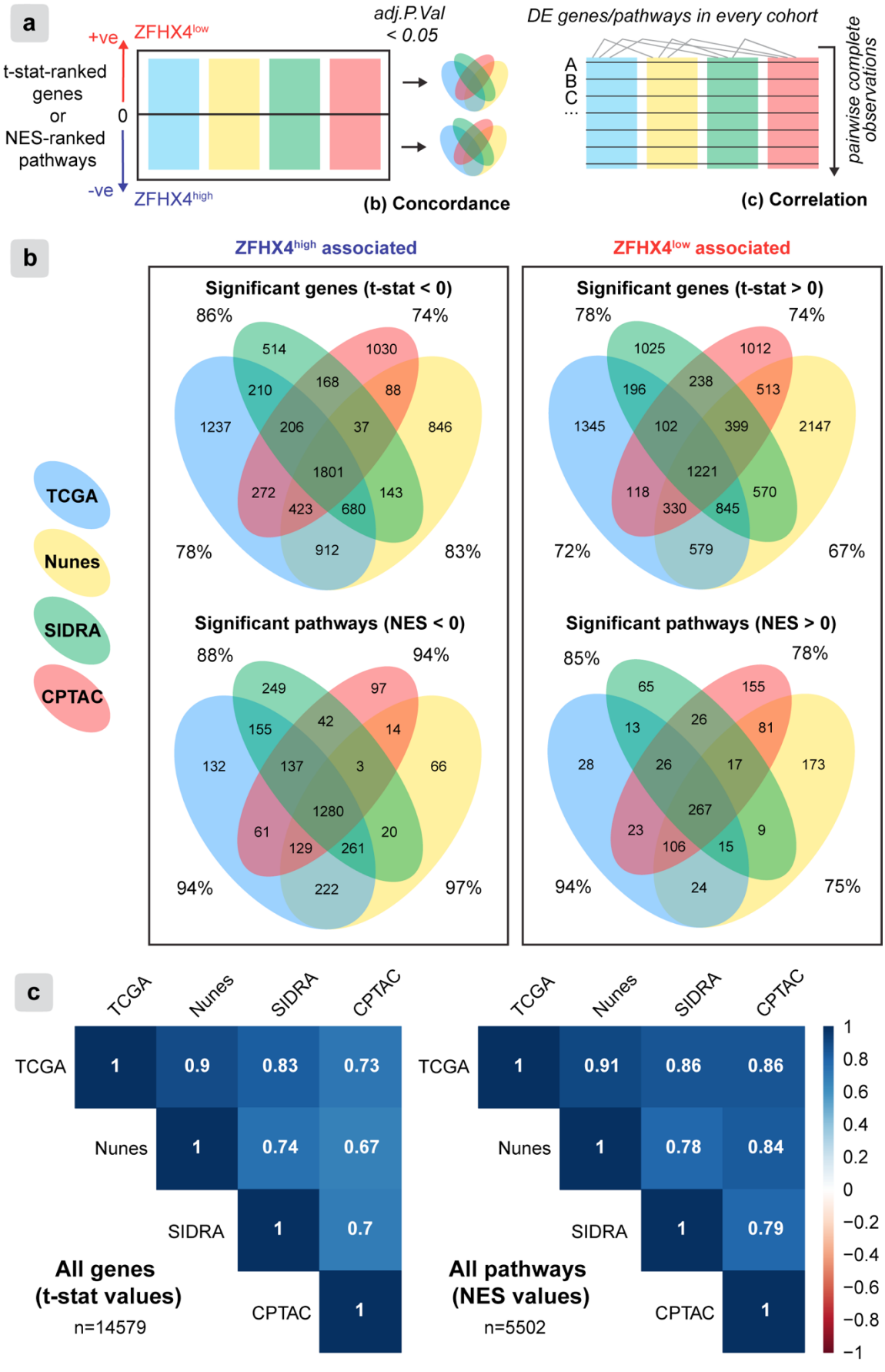

a
+ve
ZFHX4low
t-stat-ranked genes or NES-ranked pathways
0
-ve
ZFHX4high
adj.P.Val < 0.05
(b) Concordance
DE genes/pathways in every cohort
A
B
C
...
pairwise complete observations
(c) Correlation
b
ZFHX4high associated
ZFHX4low associated
TCGA
Nunes
SIDRA
CPTAC
Significant genes (t-stat < 0)
86%
74%
514
1030
210
168
88
1237
206
37
846
1801
272
143
423
680
78%
912
83%
Significant genes (t-stat > 0)
78%
74%
1025
1012
196
238
513
1345
102
399
2147
1221
118
570
330
845
72%
579
67%
Significant pathways (NES < 0)
88%
94%
249
97
155
42
14
132
137
3
66
1280
61
20
129
261
94%
222
97%
Significant pathways (NES > 0)
85%
78%
65
155
13
26
81
28
26
17
173
267
23
9
106
15
94%
24
75%
c
TCGA
Nunes
SIDRA
CPTAC
1
0.9
0.83
0.73
1
0.74
0.67
1
0.7
1
All genes (t-stat values)
n=14579
1
0.91
0.86
0.86
1
0.78
0.84
1
0.79
1
All pathways (NES values)
n=5502
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 4.40 | Concordance and correlation of differential gene and pathway analysis of *ZFHX4^{high}* vs *ZFHX4^{low}* expression across four independent colon cancer cohorts. (A) Schematics demonstrating the concordance and correlation analyses in B and C. **(B)** Venn diagrams of overlapping genes and pathways separated by positive and negative t-statistic or normalised enrichment score (NES) respectively. Percentages of genes or pathways identified in at least one other dataset are annotated by the respective dataset. **(C)** Correlation plots of differentially expressed genes or pathways between all four cohorts, with annotations of Pearson's correlation coefficient. All were significant ($p<2.2x10^{-16}$) by Pearson's pairwise testing for complete observations (n numbers denote observation number).

Therefore, I next investigated the overlapping genes and pathways to give a high-confidence prediction of ZFHX4-related biology. For each cohort, I took the top and bottom 200 significant (adj.p<0.05) genes ranked by t-stat and reported those overlapping in at least 3 cohorts (Table 4.9). 18 genes with expression negatively associated with ZFHX4 expression, and 113 genes with expression positively associated with ZFHX4 expression were identified.

Table 4.9 | Selecting the top and bottom significant (adj.p<0.05) 200 genes by t-statistic per cohort and overlapping across cohorts identifies ZFHX4-associated genes identified in at least 3 cohorts

Negatively associated (n=18)	**Positively associated** (n=113)					
ACY1	ADAM12	CDO1	FGFR1	**MITF**	PALLD	THBS2
CASP5	ADAMTS12	CHRNA3	FLRT2	MMP14	PDGFC	THBS4
CASP6	ADAMTS16	CLIP3	**FNDC1**	**MPDZ**	PDLIM7	**TIMP2**
CCL28	**AEBP1**	COL10A1	GAS1	MRAS	**PKD2**	**TNS1**
COA4	AKT3	COL11A1	GFPT2	MRC2	PLXNA4	TSHZ2
COQ2	**AMOTL1**	COL1A2	GGT5	**MSRB3**	**PRELP**	**TSPYL5**
COX5A	**ANTXR1**	COL5A1	GLI2	MXRA8	**PTGIS**	VCAN
ERI2	AOC3	COL5A2	**GLI3**	MYL9	RAB31	**VGLL3**
HADH	BGN	COL6A3	GPC6	NALCN	RNF150	**WWTR1**
HDHD3	**BNC2**	**COL8A1**	HAND2	NHSL2	**RUNX1T1**	ZEB1
ID1	BOC	**COL8A2**	**HMCN1**	NLGN2	**SERPING1**	**ZFHX4**
KAZALD1	C1R	COMP	HTRA3	NOX4	**SFRP2**	ZNF521
MARVELD3	C20orf194	CPT1C	INHBA	**NPR3**	**SFRP4**	**ZNF532**
NIPAL1	CACNA2D1	CRYAB	**ITGA11**	**NRK**	SLIT2	
PPA2	**CALD1**	**DDR2**	ITGBL1	NRP1	**SPOCK1**	
PXMP2	CCDC8	**DGKI**	KCND2	NRP2	**SSC5D**	
SMAGP	**CCDC80**	EVC	LRRC4C	NXN	**STON1**	
TMEM54	CDH11	**FAM110B**	**MAP1A**	**OLFML2B**	**SULF1**	
	CDH2	FAP	MEIS1	OLR1	TENM3	
	CDK14	FBN1	MEOX2	P4HA3	TENM4	

Genes in bold represent identification in all four cohorts

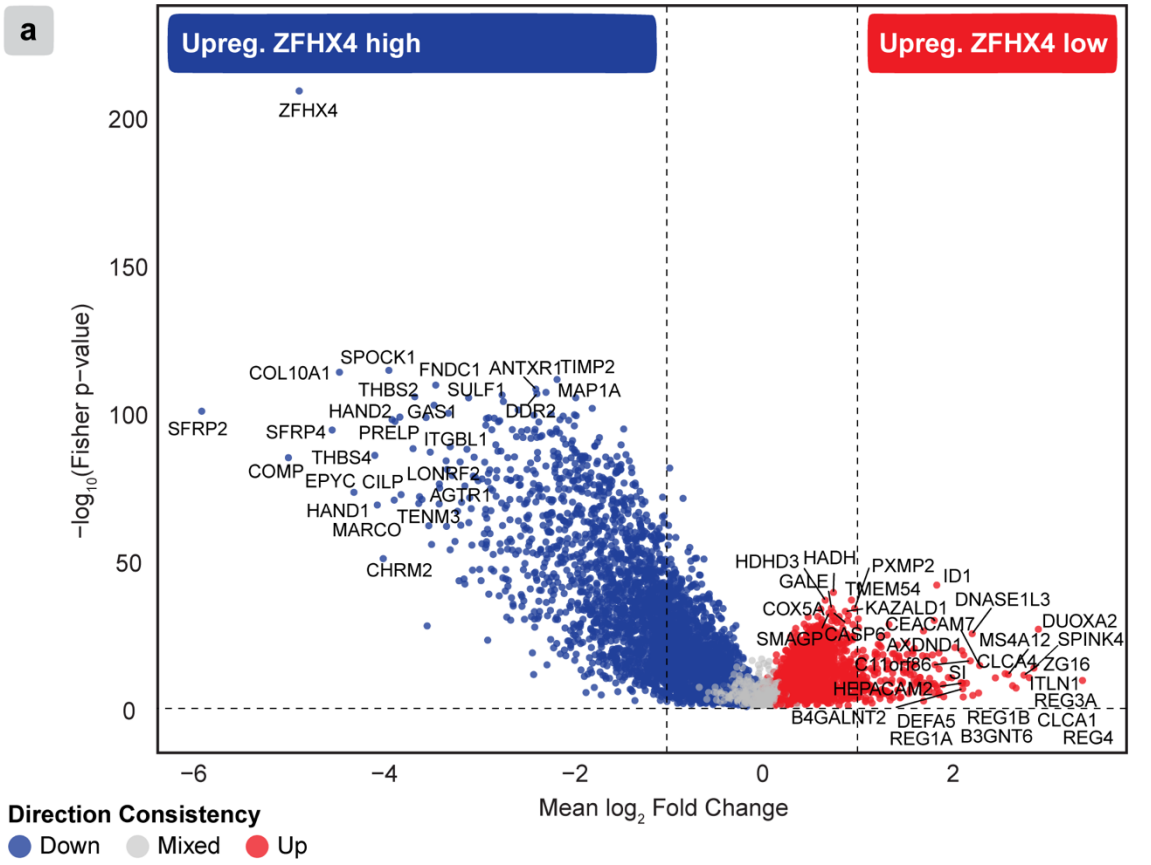
Genes negatively associated with *ZFHX4* expression included epithelial markers, apoptotic/immune regulators and mitochondrial metabolism genes, suggesting ZFHX4 may promote a shift away from oxidative metabolism and apoptotic sensitivity to immune evasion and loss of epithelial identity. Conversely, genes positively associated with *ZFHX4* expression across multiple independent cohorts were enriched for pathways involved in EMT, ECM remodelling, neuronal signalling, non-canonical Wnt signalling, hedgehog signalling and angiogenesis. This suggests ZFHX4 may orchestrate a transition to an invasive, lineage-plastic tumour state through regulation of EMT. However, interpretation here may be confounded by neuronal and stromal compartments (Section 4.4.1) as many of the genes are expressed by fibroblasts or neuronal cells. This suggests DEG analysis may partially reflect tumour microenvironment (TME) infiltration. For example, cancer-associated fibroblasts (CAFs) express mesenchymal-like genes, like those upregulated during EMT in epithelial cells, making it difficult to distinguish between genuine EMT and fibroblast infiltration. Disentangling these signals would require scRNAseq or adjusting for tumour purity as a covariate in DEG analyses, however current scRNAseq datasets in colon cancer still demonstrate a large majority of epithelium express *ZFHX4* (Figure 4.6). Regardless, elevated *ZFHX4* expression is consistently associated with poor prognosis, which may be driven either by its role in promoting a pro-tumourigenic EMT phenotype or by its expression in CAFs. Further investigation is needed to determine whether the adverse prognostic impact of ZFHX4 is primarily due to its activity in epithelial tumour cells or in the stromal compartment, or both.

For a more thorough and alternative investigation into genes and pathways related to *ZFHX4* expression, a meta-analysis was performed combining DEG analysis from all four cohorts. This consisted of converging all four DEG lists with genes, $\log_2$FC and raw p-values by gene names, retaining only those genes present in all four datasets. For each gene, mean $\log_2$FC was calculated to represent the average direction and magnitude of expression differences across datasets. Raw p-values were combined using Fisher's

method, defined by the formula below. This yielded a combined significance value per gene, highlighting those consistently differentially expressed across datasets.

$Fisher\ statistic = -2 \times \sum \log(pvalue)$, which follows a chi-squared distribution with 2 x number of p-values degrees of freedom (i.e. 4).

A meta-volcano plot was generated to visualise genes reproducibly up- or downregulated with respect to *ZFHX4* expression by plotting mean log_2FC and Fisher-combined p-value (Figure 4.41A). To ensure result fidelity, only genes differentially expressed in ≥3 cohorts were retained (why all plotted genes are significant), totalling 7367 genes. If genes achieved the same direction of upregulation ('Up', i.e. associated with *ZFHX4*^{low}) or downregulation ('Down', i.e. associated with *ZFHX4*^{high}) across ≥3 cohorts they were annotated as such, however those where direction was not equal amongst cohorts were labelled as 'Mixed'.


a
Upreg. ZFHX4 high
Upreg. ZFHX4 low
ZFHX4
SPOCK1
COL10A1
FNDC1
ANTXR1
TIMP2
THBS2
SULF1
MAP1A
HAND2
GAS1
DDR2
SFRP2
SFRP4
PRELP
ITGBL1
THBS4
COMP
EPYC
CILP
LONRF2
AGTR1
HAND1
TENM3
MARCO
CHRM2
HDHD3
HADH
PXMP2
ID1
GALE
TMEM54
DNASE1L3
COX5A
KAZALD1
CEACAM7
DUOXA2
SMAGP
CASP6
AXDND1
MS4A12
SPINK4
C11orf86
CLCA4
ZG16
SI
ITLN1
HEPACAM2
REG3A
B4GALNT2
DEFA5
REG1B
CLCA1
REG1A
B3GNT6
REG4
200
150
100
50
0
−log10(Fisher p−value)
−6
−4
−2
0
2
Mean log2 Fold Change
Direction Consistency
Down
Mixed
Up

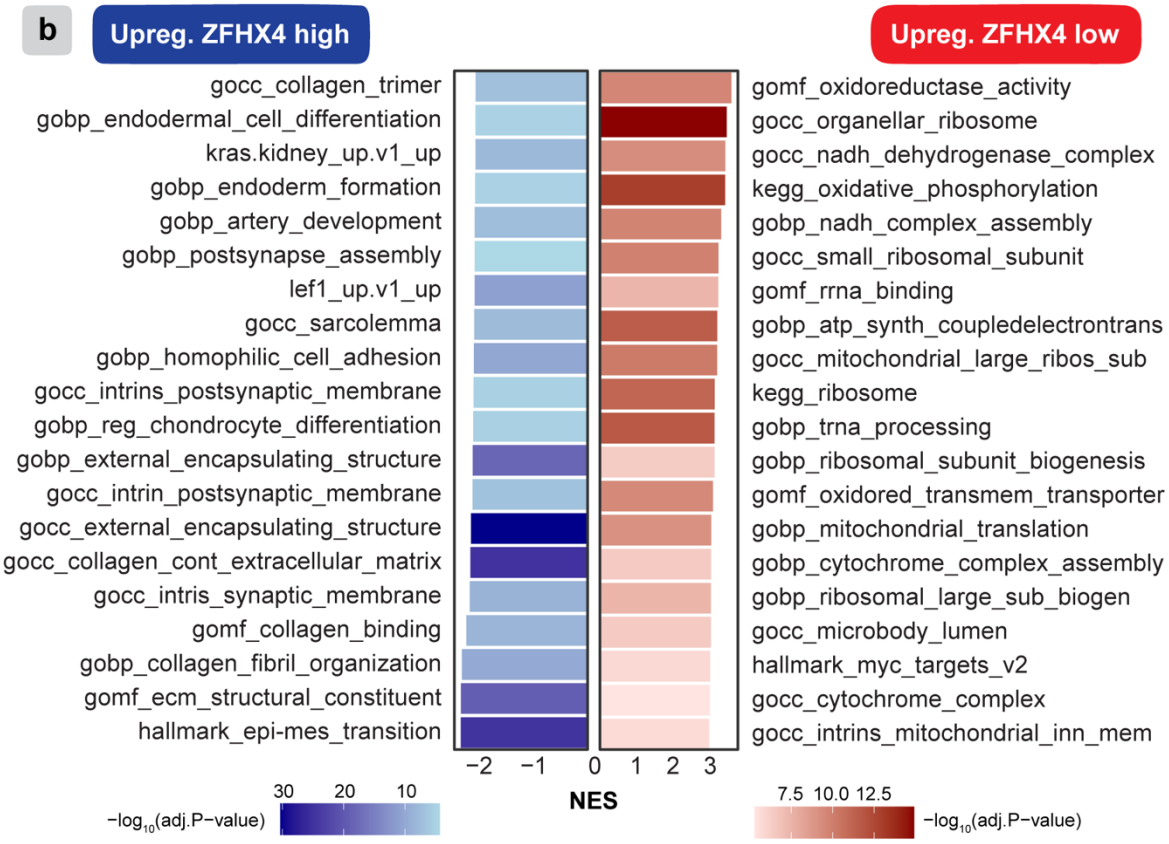

b
Upreg. ZFHX4 high
Upreg. ZFHX4 low
gocc_collagen_trimer
gobp_endodermal_cell_differentiation
kras.kidney_up.v1_up
gobp_endoderm_formation
gobp_artery_development
gobp_postsynapse_assembly
lef1_up.v1_up
gocc_sarcolemma
gobp_homophilic_cell_adhesion
gocc_intrins_postsynaptic_membrane
gobp_reg_chondrocyte_differentiation
gobp_external_encapsulating_structure
gocc_intrin_postsynaptic_membrane
gocc_external_encapsulating_structure
gocc_collagen_cont_extracellular_matrix
gocc_intris_synaptic_membrane
gomf_collagen_binding
gobp_collagen_fibril_organization
gomf_ecm_structural_constituent
hallmark_epi-mes_transition
gomf_oxidoreductase_activity
gocc_organellar_ribosome
gocc_nadh_dehydrogenase_complex
kegg_oxidative_phosphorylation
gobp_nadh_complex_assembly
gocc_small_ribosomal_subunit
gomf_rrna_binding
gobp_atp_synth_coupledelectrontrans
gocc_mitochondrial_large_ribos_sub
kegg_ribosome
gobp_trna_processing
gobp_ribosomal_subunit_biogenesis
gomf_oxidored_transmem_transporter
gobp_mitochondrial_translation
gobp_cytochrome_complex_assembly
gobp_ribosomal_large_sub_biogen
gocc_microbody_lumen
hallmark_myc_targets_v2
gocc_cytochrome_complex
gocc_intrins_mitochondrial_inn_mem
−2
−1
0
1
2
3
NES
30
20
10
−log10(adj.P−value)
7.5
10.0
12.5
−log10(adj.P−value)

Figure 4.41 | Meta-analysis of differential gene expression and pathway enrichment between *ZFHX4*^{high} and *ZFHX4*^{low} patient groups across four independent cohorts. (A) Meta-volcano plot with each dot representing a gene and coloured by consistency of up- or down-regulation across ≥ 3 datasets. Labelled genes represent those with the most positive or negative mean log_2FC and Fisher combined p-value (n=7367 genes). **(B)** Bar plots demonstrating normalised enrichment scores (NESs) of pathway enrichment following GSEA on the differential genes identified in (A), ranking by mean log_2FC. Gene sets were obtained from msigdbr and included hallmark, biocarta, kegg, go-bp, go-mf, go-cc and oncogenic subsets. The top 20 pathways were selected based on NES ranking and coloured by significance. Each bar colour intensity refers to significance of the adj.p-value. Utilised cohorts were TCGA-COAD (Hoadley et al., 2018), Nunes (Nunes et al., 2024), SIDRA (Roelands et al., 2023) and CPTAC2 (Vasaikar et al., 2019).

The top 50 positively and negatively associated genes with *ZFHX4* expression (ranked by mean log_2FC) are summarised in Table 4.10. While many overlap with those in Table 4.9, this meta-analysis approach offers key advantages over simple gene list intersections as it incorporates magnitude, directionality and reproducibility across independent cohorts. For instance, SFRP2 showed the strongest overall association with *ZFHX4* expression, exceeding even ZFHX4 itself by mean log_2FC and this effect could not be fully captured in Table 4.9. Furthermore, the intrinsic skew in *ZFHX4* RNAseq data, observed across nearly all volcano plots in this chapter and likely reflecting its role as a transcriptional activator, confounded the intersection analysis and resulted in fewer negatively associated genes being identified. This underscores the added value of this meta-analysis in identifying consistently regulated genes associated with ZFHX4 biology.

Table 4.10 | Ranked by mean log_2FC, the top 50 positively and negatively associated genes with ZFHX4 expression across all four cohorts.

Positively associated with ZFHX4 expression				Negatively associated with ZFHX4 expression			
Gene	Mean log_2FC	Gene	Mean log_2FC	Gene	Mean log_2FC	Gene	Mean log_2FC
SFRP2	-5.88	LRRC4C	-3.44	CLCA1	3.36	C4BPA	1.91
COMP	-4.97	FNDC1	-3.42	DUOXA2	2.90	DEFA6	1.90
ZFHX4	-4.85	TCEAL2	-3.39	SPINK4	2.85	PIGR	1.90
SFRP4	-4.51	CCBE1	-3.38	ITLN1	2.80	CA1	1.86
COL10A1	-4.43	OMD	-3.38	ZG16	2.74	AKR1B10	1.86
EPYC	-4.28	ADAMTS16	-3.31	REG3A	2.66	LCN2	1.85
THBS4	-4.06	SPP1	-3.31	REG4	2.63	OLFM4	1.85
HAND1	-4.03	PLPP4	-3.30	MS4A12	2.58	VSIG2	1.83
CHRM2	-3.97	COL11A1	-3.29	CLCA4	2.55	ID1	1.83
SPOCK1	-3.91	CHRDL1	-3.27	B3GNT6	2.45	GCG	1.82
HAND2	-3.88	KCND2	-3.27	CEACAM7	2.28	LRRC26	1.81
MARCO	-3.86	MEOX2	-3.25	REG1B	2.27	MUC2	1.80

PRELP	-3.85	OGN	-3.22	REG1A	2.21	CASP5	1.80
GAS1	-3.80	CNTN1	-3.19	DNASE1L3	2.20	BEST2	1.80
CILP	-3.79	SFRP1	-3.18	C11orf86	2.18	BMP5	1.78
ITGBL1	-3.66	NALCN	-3.17	SI	2.14	ATOH1	1.77
THBS2	-3.64	C5orf46	-3.16	AXDND1	2.11	RETNLB	1.76
TENM3	-3.60	PDZRN4	-3.15	DEFA5	2.11	LCN15	1.75
LONRF2	-3.59	FABP4	-3.13	B4GALNT2	2.10	SPINK1	1.73
AGTR1	-3.57	GDF6	-3.12	HEPACAM2	2.10	DMBT1	1.71
NPR3	-3.52	DES	-3.12	C10orf99	2.09	ADH1C	1.71
ADIPOQ	-3.51	CDH2	-3.10	CWH43	2.09	FABP2	1.70
CASQ2	-3.50	ADAM12	-3.08	DUOX2	2.02	NXPE4	1.69
NRK	-3.48	PENK	-3.08	CA4	1.99	CCL28	1.69
GPM6A	-3.47	EBF2	-3.07	FCGBP	1.95	ADH6	1.68

All were reported significant by Fisher combined p-value ($p<0.05$)

To characterise these genes with respect to pathway enrichment and use the entire ranked gene lists, GSEA was performed using the same gene sets as previous, ranking genes by mean log_2FC (Figure 4.41B). EMT, ECM-remodelling, developmental plasticity and neuronal-like signalling pathways were associated with high *ZFHX4* expression comparative to high oxidative phosphorylation and mitochondrial function, ribosomal biogenesis, protein synthesis and tRNA/rRNA processing pathways associated with low *ZFHX4* expression. Together, these findings suggest that *ZFHX4* may act as a master regulator of tumour cell plasticity, promoting invasive and mesenchymal phenotypes while repressing metabolic and epithelial programs. However, this work is purely association-based and further *in vitro* work would need to establish these mechanisms.

To assess the similarity of clinical $ZFHX4^{high}$ vs $ZFHX4^{low}$ comparisons and experimental $ZFHX4^{-/-}$ vs $ZFHX4^{wt}$ comparisons, using $ZFHX4^{-/-}$ as a surrogate for low *ZFHX4* expression, all datasets were combined, maintaining the same direction for DEG analysis. For example, negative fold change will be respect to both clinical $ZFHX4^{high}$ and experimental $ZFHX4^{wt}$ and positive fold change will be respect to both clinical $ZFHX4^{low}$ and experimental $ZFHX4^{-/-}$. Additionally, representing the lowest *ZFHX4* expressing clones, CC9 and NG7 DEGs lists were created by performing a fold change calculation using respective WT triplicate values and singular KO values from the RNAseq experiment. Whilst no adjusted P-value significance can be obtained, these lists still hold valuable information

regarding expressed genes. Furthermore, log_2FC was used from this point to allow inclusion of these extra DEG lists where t-statistic was not available. To evaluate how closely experimental datasets of *ZFHX4*$^{-/-}$ recapitulated the differences in transcriptional programs between *ZFHX4*^{high} and *ZFHX4*^{low} tumours observed in clinical colon cancer datasets, I performed similar correlation analysis as earlier but between all differentially expressed genes (log_2FC) and pathways (NES) across all clinical and experimental datasets (Figure 4.42). Again, only features present in every dataset were retained (n=12947 genes, n=5187 pathways) thereby minimising confounding from TME-specific transcripts absent in cell lines and organoids.

At the gene level, the clinical datasets were highly and positively correlated with each other (Pearson's r>0.7) as reported in Figure 4.40 by t-stat, confirming the robustness of the ZFHX4-associated transcriptional signature across independent patient cohorts. Moreover, the strong positive correlation of CC9 and NG7 towards their respective monoKO comparisons is a good quality control to show the reliability of this correlative analysis. Interestingly, the experimental datasets correlated weakly with the clinical profiles. NCIH716 polyKOs and monoKOs displayed a weak but positive correlation with the clinical data ($r \approx 0.0–0.2$), whereas COLO320DM polyKOs and monoKOs showed a weak negative correlation ($r \approx 0.0$ to $–0.2$). SKNMC polyKOs showed virtually no correlation with either clinical or experimental datasets, consistent with its distinct tissue origin and potentially inefficient knockout. Organoid *ZFHX4*$^{-/-}$ displayed similar but attenuated trends, also likely reflecting incomplete depletion of ZFHX4.

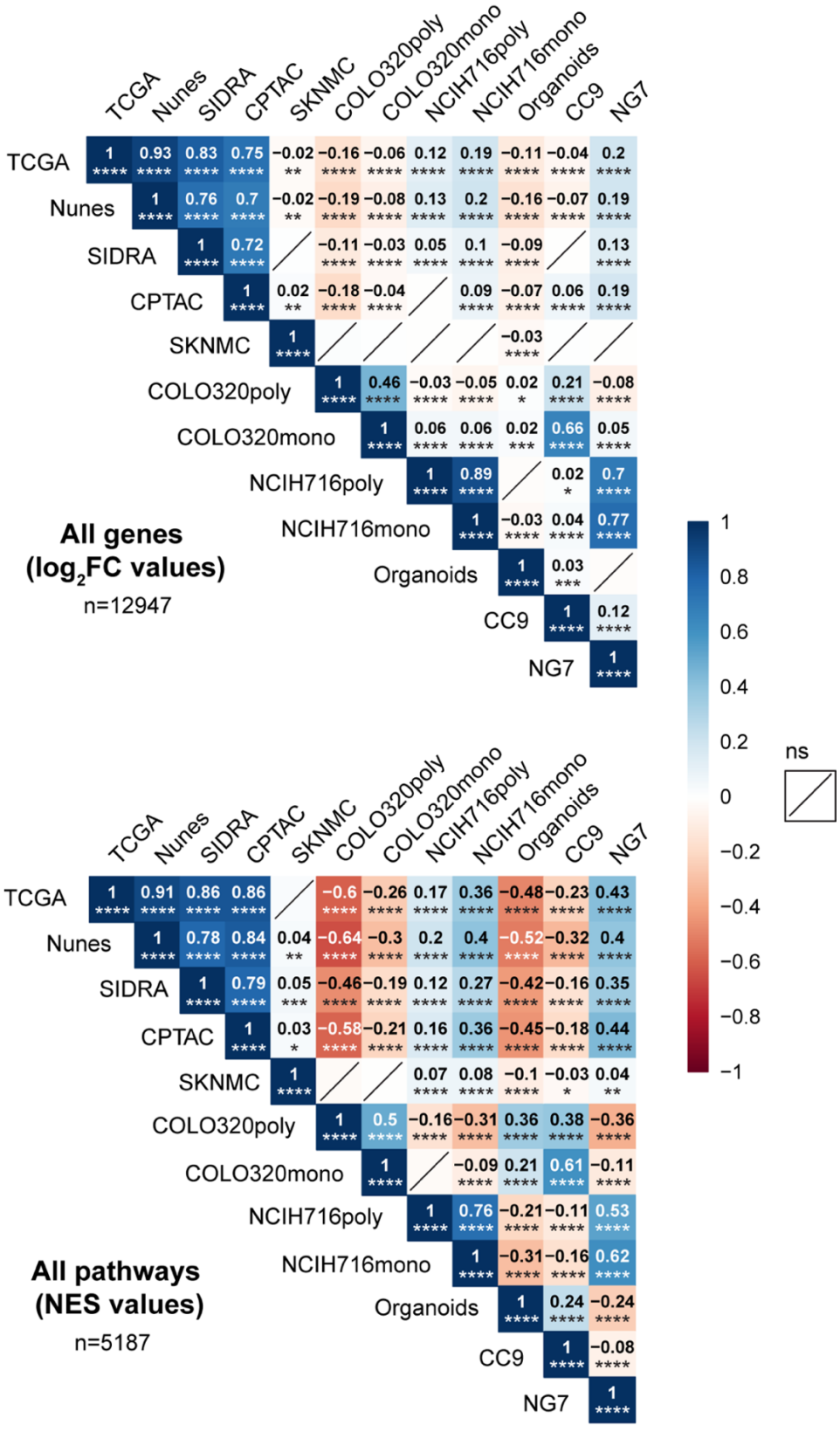


Figure 4.42 | Correlation plots of differentially expressed genes and pathways between all four clinical cohorts and 8 experimental datasets. Comparisons are annotated with Pearson's correlation coefficient. Asterisks denote significance by Pearson's pairwise testing for complete observations (n numbers denote observation number). Clinical *ZFHX4*^{high} vs *ZFHX4*^{low} comparisons are TCGA, Nunes, SIDRA and CPTAC, and experimental *ZFHX4*^{wt} vs *ZFHX4*$^{-/-}$ comparisons represent the rest. *[*p<0.05, **p<0.01, ***p<0.001, ****p<0.0001, ns=not significant]*

At the pathway level, correlations were more pronounced and biologically informative. Again, SKNMC *ZFHX4*$^{-/-}$ failed to correlate with any dataset at either the gene or pathway level. *ZFHX4*$^{-/-}$ in NCIH716 was positively correlated with the clinical *ZFHX4*^{low} pathway signature ($r \approx 0.12–0.44$), indicating this model most faithfully recapitulates the *in vivo* LoF phenotype. In contrast, *ZFHX4*$^{-/-}$ COLO320DM lines produced negative correlations with the clinical data ($r \approx –0.5$ to $–0.6$) suggesting that *ZFHX4* ablation in these contexts induced transcriptional programs opposite to those observed in *ZFHX4*^{low} tumours. Organoid models showed the same trend, which is notable as these models are generally expected to most closely reflect clinical data owing to their recent derivation from metastatic tissue. However, the inefficient CRISPR knockout likely contributed to this outcome. Additionally, NCIH716 and COLO320DM *ZFHX4*$^{-/-}$ cell lines were inversely correlated with each other ($r \approx –0.1$ to $–0.3$).

In summary, this apparent divergence likely reflects context-dependent regulation of *ZFHX4*, differences in genetic background (such as MYC amplification in COLO320DM) and the absence of tissue-level signalling cues *in vitro*. Thus, among the *in vitro* models utilised in this thesis, NCIH716 may exhibit the most suitable *in vitro* system for further mechanistic studies of *ZFHX4*, while COLO320DM and organoid models appear to recapitulate the clinical biology less faithfully. However, considering the clonal-dependent effects observed *in vitro* thus far, and the need for further validation in loss of *ZFHX4* mRNA expression in *ZFHX4*$^{-/-}$ samples, further work would be necessitated to refine model selection going forward.

Despite the apparent differences and similarities between experimental and clinical cohorts, specific pathways of interest were further investigated. Bubble plots were produced to visualise fold change differences and significance of specific genes between either *ZFHX4*^{high} vs *ZFHX4*^{low} clinical samples, or *ZFHX4*^{wt} vs *ZFHX4*$^{-/-}$ experimental models (Figure 4.43). For clarity, colour coding in this figure is unique compared to others in this chapter, with red representing a ZFHX4-high context and blue representing a ZFHX4-low context, as observed with ZFHX4 vs common housekeeper controls. Looking at NuRD complex members specifically due to their reported association with ZFHX4 in GBM such as CHD4 (Chudnovsky et al., 2014), CHD3/4/5 expression appears associated with the *ZFHX4*^{high} setting clinically and CHD3/4 were also upregulated in ZFHX4^{wt} comparative to ZFHX4$^{-/-}$ (CC9 and NG7) indicating they are not only expressed in these models but may also be transcriptionally regulated by *ZFHX4*. The large fold upregulation of CHD5 with CC9 *ZFHX4*$^{-/-}$ suggests potential negative feedback loops may also be present. However, as NuRD exerts its function through protein-protein interactions of its members, it may be better to overlay a NuRD-related phenotype rather than NuRD member genes.

Because EMT-related pathways were consistently altered in both clinical and experimental datasets, we further evaluated a curated set of epithelial and mesenchymal markers sourced from Cell Signalling Technology to align with potential future protein validation (CST #9782) and the general literature. Epithelial genes such as CHD1 and CLDN1 were downregulated in ZFHX4-high conditions in both settings. More notably, mesenchymal markers enriched in *ZFHX4*^{high} tumours including CDH2, FN1 and ZEB2 were strongly downregulated upon *ZFHX4* knockout in cell line models. These findings provide compelling evidence that *ZFHX4* actively regulates an EMT program within epithelial cells, suggesting that EMT signal observed in clinical samples is not solely derived from stromal components such as CAFs or fibroblasts, but reflects a genuine tumour cell-intrinsic EMT phenotype. It would be of great interest to validate these changes at the protein level and establish a true EMT phenotype associated with *ZFHX4* expression.

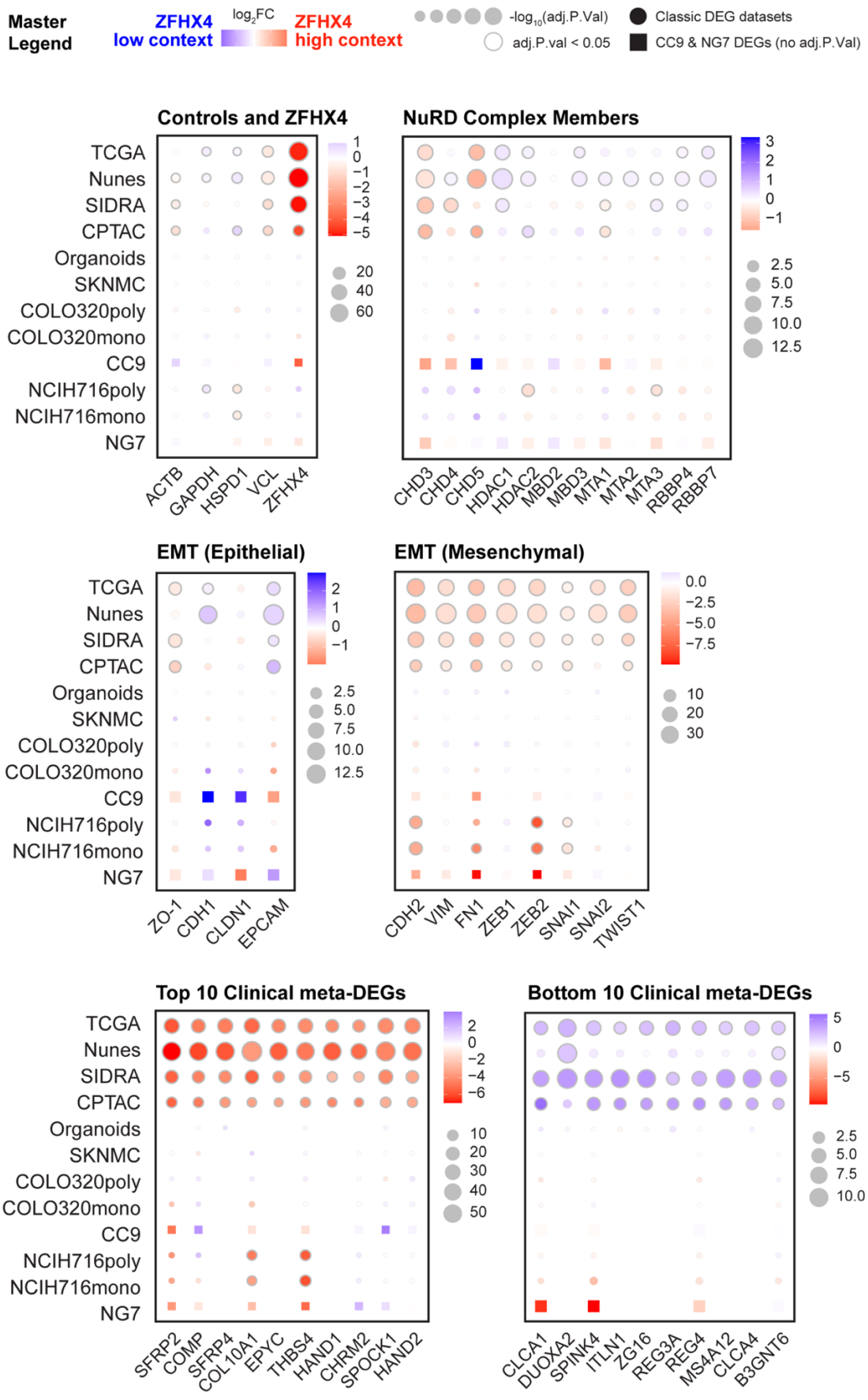


Figure 4.43 | Bubble plots of genes up- or down-regulated in differential gene expression comparisons amongst clinical and experimental datasets.

To further support this integration between clinical and experimental models, I profiled the top and bottom 10 clinical meta-DEGs from Table 4.10. Genes such as *SFRP2*, *COL10A1* and *THBS4* associated with the clinical *ZFHX4*^{high} phenotype were also significantly downregulated upon *ZFHX4* knockout in cell line models, reinforcing their regulation by *ZFHX4* and highlighting the value of combining patient-derived data with experimental validation to uncover core *ZFHX4*-driven biology.

More work is warranted here for a truly exhaustive analysis of genes associated with a clinical ZFHX4^{high} phenotype and downregulated through ZFHX4 knockout in experimental models and vice versa, similar to the meta-analysis performed in Figure 4.41. Furthermore, despite only n=1, individual clones could be analysed in the same way as CC9 and NG7, as this may highlight which clones were most representative of the clinical phenotype.

4.5 Discussion

ZFHX4's potential role in tumourigenicity was explored in this chapter because even though its mutations were associated with an immunogenicity-related phenotype (the original focus of this thesis), its biology in colon cancer had not been previously examined. Moreover, given that tumourigenicity and immunogenicity are interlinked, understanding ZFHX4's basic cancer biology is a prerequisite to probing its role in immunogenicity. Only a few studies in other cancer types had implicated ZFHX4 with tumourigenic phenotypes such as epigenetic regulation through NuRD complex interactions, EMT and migration (Table 4.1). Additionally, the prognostic association of *ZFHX4* mRNA expression (Figure 4.4) and its correlation with tumour stage (Figure 4.5) in colon cancer supports its investigation as a potential regulator of both immunogenicity and tumourigenicity.

However, investigating ZFHX4 *in vitro* proved challenging and came with unexpected complexities. These difficulties arose not only from limitations inherent to the employed cell line and organoid models, of which ZFHX4 had not been validated before, but also from technical issues such as those relating to the availability and reliability of anti-ZFHX4 antibodies. Such issues confounded many key experimental approaches such as WB and CUT&RUN, both of which depend heavily on reliable and efficient antibody binding. Furthermore, *ZFHX4* itself is a large, complex gene with poorly characterised biology making it difficult to interpret results or isolate a definitive phenotype. Although the precise functional role of ZFHX4 remains unresolved, the data presented in this chapter consistently implicate EMT-related processes as a principal axis of its activity. These findings provide a strong rationale for further mechanistic investigation.

In this section I will reflect on the models and techniques used in this chapter, discuss potential mechanistic insights and outline future experiments that could more precisely dissect the role of ZFHX4 in colon cancer biology.

4.5.1 Organoid and cell line models

The original aim of my work was to employ primary patient-derived colon organoids as a model system. These organoids originally pioneered by Clevers and Sato in the early 2010s (Jung et al., 2011, Sato et al., 2011) are widely considered state-of-the-art models that more faithfully recapitulate normal and neoplastic colon biology compared to conventional models such as cell lines. Since their introduction they have inspired a wide range of research opportunities including functional modelling of somatic mutations using genome-editing technologies such as CRISPR/Cas9 (Drost et al., 2015, Matano et al., 2015) and dissection of TME interactions through co-culture with immune (Cattaneo et al., 2020, Dijkstra et al., 2018) and stromal (Qin et al., 2023) compartments. These advances have also led to the establishment of biobanks and subsequent high-throughput drug screening (van de Wetering et al., 2015). They have further enabled studies across multiple clinical stages, supported by the successful derivation of organoids from normal tissue, primary tumours and metastatic lesions (Fujii et al., 2016). Despite holding great potential to study colon cancer development, these models remain technically demanding to culture. For example, normal human colonic organoids are typically challenging to culture beyond passage 8, likely because growth conditions are not yet fully optimised, consistent with varied reported growth durations and protocols in the literature referenced above and thus lack of standardisation across research groups. While tumour-derived colonic organoids generally grow more rapidly and for longer periods than normal organoids, they are inherently limited by their mutated genetic background which must be fully characterised with tools such as WES before functional studies can be confidently interpreted.

Specifically, organoids derived from adjacent normal tissue of colon cancer patients undergoing resection are a central model system used in the Buczacki lab. These “genetically clean” organoids enable the study of specific gene perturbations without the confounding influence of pre-existing mutations commonly found in cancer-derived organoids. This approach has proven valuable in previous work where mutations in genes

like FBXW7 led to striking phenotypes (Chan et al., 2023a). One goal of my work was to phenotypically validate at least three *ZFHX4*$^{-/-}$ patient-derived normal colon organoid lines to study a ZFHX4-related phenotype in the primary setting. This would've been ideal in combination with or without CRISPR/Cas9-mediated backgrounds such as APC$^{-/-}$ and TP53$^{-/-}$ which have been shown to replicate tumourigenic phenotype (Matano et al., 2015, Drost et al., 2015, Chan et al., 2023a). Despite extensive efforts, generating normal-derived *ZFHX4*$^{-/-}$ models proved exceptionally challenging because of both technical and biological issues. These included inefficient CRISPR editing, poor organoid growth, frequent bacterial and fungal contamination, limited gDNA/mRNA yield and most importantly an unexpected cell death or senescence phenotype that was repeatedly observed. Even when switching to a metastatic-derived organoid line previously characterised for optimal ZFHX4 mRNA expression and WT status and confirmation of an efficient CRISPR-mediated KO, the loss of this KO identified following RNAseq exemplified inherent limitations and unpredictability of labour-intensive organoid work, especially within time-constrained research projects. Whether the senescent phenotype seen was related to ZFHX4 loss or simply technical is moot and remains uncertain.

As a consequence, the focus of my work shifted toward cell line models which still remain cornerstone of cancer biology and continue to provide valuable platforms for functional genomics, traditionally supplementing similar work in organoids (Chan et al., 2023c). These models offered long-term study of ZFHX4-related phenotype using many *in vitro* techniques on genetically distinct clonal populations, and whilst this was advantageous, this also provided a double-edged sword by introducing confounding of:

(1) variation of the same *ZFHX4*$^{-/-}$ line **between passages.** While long-term culture allowed for repeated experimentation, I observed considerable variation in phenotype between different passages of the same *ZFHX4*$^{-/-}$ clone. This was particularly evident in the proliferation assays using CellTiter-Glo where experiments conducted at different time points, using distinct passage replicates for reliability, yielded inconsistent results. For

example, the COLO320DM-CC9 clone, which displayed the most pronounced ZFHX4 mRNA loss, exhibited markedly reduced proliferation in 2m cells. However, by month 11 the same clone showed a reversion to normal proliferation rates, complicating interpretation of whether this effect was truly ZFHX4-dependent or an artifact of prolonged culture. Additionally, despite maintaining consistent levels of %indel and %KO across three months, the distribution of specific indel variants shifted over time (Table 4.6). This suggests either poor sensitivity of Sanger sequencing for characterising a specific mutation, inefficiencies in the single-cell cloning process or most likely genetic drift or adaptation during long-term culture. These factors highlight the need to revalidate clones at later passages, especially as most post-3m cells were used for functional assays. It also underscores the difficulty in attributing specific phenotypes to particular indels. Furthermore, while RNAseq was performed to capture the transcriptional consequences of ZFHX4 knockout, it was limited to a single replicate per clone as a compromise of using different clones to represent biological replicates. Ideally, RNAseq across multiple passage replicates would provide a more comprehensive view of ZFHX4-related transcriptional dynamics.

(2) variation **between different *ZFHX4*$^{-/-}$ lines** from the same parent. A key advantage of employing cell lines is the ability to isolate single-cell clones, enabling precise generation of independent, genetically defined ZFHX4 knockouts. Six of these were utilised per cell line as by the original premise of eventually selecting the most representative to study ZFHX4. However, irrespective of cell line these clones often displayed divergent phenotypes across multiple assays, despite confirmation as complete knockouts at the gDNA level. For example, proliferation assays indicated that phenotypic differences in proliferation arose from clonality rather than ZFHX4 genomic status, as similar variability was observed across wild-type single-cell clones. This finding is consistent with prior reports that phenotypic and even transcriptional heterogeneity exists between clonal wild-types derived from the same parent, and that this contributes to the variability of phenotype observed between genome-edited clones (Westermann et al., 2022). To account for this

phenotypic variability, all KO lines were consistently profiled in every experiment, amplifying *in vitro* work substantially, considering biological replicates were n=17 and that technical passage replicates also needed incorporating. To alleviate these issues, a more robust and reproducible approach for future studies would be to first derive monoclonal wild-type parental lines and then perform CRISPR/Cas9 editing to generate ZFHX4 knockouts from these defined backgrounds. This matched isogenic system would help minimise phenotypic noise from clonal variability and has been previously proposed as a best practice in genome editing workflows with cell lines (Westermann et al., 2022). However, given the divergent clonal phenotypes associated with the loss of ZFHX4, even this approach may not yield the 'true' *ZFHX4*$^{-/-}$ phenotype and importantly highlights the limitation of cell line models to infer phenotype more generally.

Present with all models utilised in this thesis was limited correlation to the ZFHX4 phenotype observed clinically (Figure 4.43), unsurprising considering clinical comparisons of *ZFHX4*^{high} vs *ZFHX4*^{low} capture a graded, context-dependent expression landscape, whereas CRISPR knockout imposes a binary loss that can invoke nonlinear dose effects, feedback, and compensatory rewiring. Consequently, complete ablation may diverge from, or even invert, the transcriptional programs associated with clinical low expression depending on context. My data is consistent with this principle: pathway concordance is strongest in NCIH716 cells but the reverse in COLO320DM cells which negatively correlated with ZFHX4^{high} clinically. Intrinsic idiosyncrasies between COLO320DM and NCIH176 lines likely explain this, for example COLO320DM is MYC-amplified, neuroendocrine-like, adherent and colon-derived (Bianchi et al., 1991, Alitalo et al., 1983, Berg et al., 2017) whereas NCIH716 is FGFR2-amplified, suspension and derived from ascites of a metastatic patient (Mathur et al., 2014). Even though both demonstrated loss of a pro-growth program upon ZFHX4 knockout, the specificity of MYC/ribosome pathway downregulation in COLO320DM is likely related to COLO320DM's inherent MYC-amplified biology, highlighting ZFHX4 may regulate distinct transcriptional networks depending on

cellular context. Despite this inter-model heterogeneity, the selection of both lines as core models of biological replicates for this thesis was well justified. Both expressed moderate to high levels of ZFHX4 and share not only ZFHX4^{wt} status but MSS/TMBL status too. No other available colon-related cell lines satisfied all these criteria making them the most suitable models for functional investigation of ZFHX4 within the constraints of available systems.

The limitations of cell line models in fully capturing ZFHX4-associated biology underscore the potential of primary tissue systems as more informative platforms for phenotype discovery. This is supported by prior work in GBM where ZFHX4 knockdown led to a dramatic reduction in proliferation in primary patient-derived, stem-like TICs and healthy neural progenitor cells (NPCs), but not in more differentiated cancer cell lines such as U251 (Chudnovsky et al., 2014). Similarly, ZFHX4 knockdown in two oesophageal cancer cell lines showed no effect on proliferation, further illustrating model-dependent variability (Qing et al., 2017). Another study performing CUT&RUN on ZFHX4 in stem-like NPCs demonstrated different transcriptional targets compared to CUT&RUN on differentiated neurons (Perez Baca et al., 2025). Taken together, this is strong evidence that ZFHX4 function is highly context-dependent and may be governed by chromatin state.

Further work in the primary tissue setting, namely scRNAseq and IHC have given particularly interesting insight. I demonstrated consistent ZFHX4 protein expression in a distinct epithelial subset localised to the crypt base and in villi regions across all matched tissues from two patients. This represents the first colon-specific ZFHX4 IHC characterisation beyond what is available through the Human Protein Atlas, offering a novel direction for further validation. Techniques such as immunofluorescence (IF) could be utilised to co-stain for canonical epithelial subset markers and even extended to organoid models. For example, some studies have driven organoids towards certain lineages such as enteroendocrine cells and could offer a novel model system to investigate ZFHX4 biology within specific epithelial subsets using both IF and scRNAseq (Zeve et al., 2022, Yang et

al., 2025). One study even used human intestinal organoids to investigate NuRD and NuRD-targeting in different LGR5/stemness-related contexts (Yang et al., 2025). Preliminary scRNAseq data suggests ZFHX4 can be expressed by multiple epithelial cell types in both normal and cancer settings (Section 4.4.1) and further work here is required to explore this in depth. For example, integrating all key scRNAseq datasets and performing differential gene expression and pathway enrichment analyses between ZFHX4^{low} and ZFHX4^{high} cells within specific cell subsets would provide the most refined *in silico* assessment of ZFHX4 function. It would also help overcome the limitations of bulk-RNAseq which cannot disentangle expression across epithelial, stromal or neuronal compartments, each of which may exhibit distinct ZFHX4 regulatory programs. As an initial exploration, I used the cellxgene portal for this purpose and identified preliminary associations between ZFHX4 and NuRD complex members as discussed later, reinforcing the value of this approach. This is particularly exciting in light of recent advances in understanding epithelial subset function in colon cancer development, especially since the discovery of LGR5+ intestinal stem cells (Barker et al., 2007) and positions this work at the intersection of developmental biology, immunogenicity and tumourigenicity. The combination of prognostic relevance, mutational impact on immune phenotypes and subset-specific biology of ZFHX4 could therefore have significant translational implications.

4.5.2 gDNA, mRNA and protein profiling of ZFHX4

A central challenge in my work was reconciling gDNA, mRNA and protein level changes after CRISPR targeting of *ZFHX4* in cell lines. Although genomic *ZFHX4* editing was confirmed in all clones (with some expected WT carryover in polyKO pools), RNAseq didn't uniformly show loss of *ZFHX4* expression, contrary to the original premise of engineering a knockout. Notably, transcript-level quantification proved more informative than gene-level counts, which can be confounded by degraded or partial transcripts. For example, MANE

transcripts were clearly reduced in NC8, ND5, and CF9 by transcript-level analysis, whereas gene-level summaries suggested little change or even apparent increases compared to WT. This illustrates the added resolution of transcript-aware quantification particularly for large, multi-isoform genes like *ZFHX4*.

Amongst all clones, only CC9 showed near-ablated *ZFHX4* mRNA levels that was concordant with the loss of protein by HPA antibody. CC9 also exhibited a striking slow-growth phenotype in early passages and required extended culture to yield sufficient material for experiments. This was also consistent with cell cycle analysis demonstrating less cells in S-phase which were replicating DNA much more slowly comparative to WT and other clones. Though by 11 months, it had adapted and recovered proliferative capacity, consistent with adaptive cellular rewiring or compensatory mechanisms that may be attributable to upregulation of paralogs like ZFHX3. This also highlights the potential pitfalls of prolonged cell line culture following genome editing without consistent validation checks as previously discussed.

Investigating ZFHX4 at all the central dogma levels prompted a plausible hypothesis: complete loss of ZFHX4 is detrimental to cancer cells, given its predicted role in promoting tumourigenesis and predominance of non-truncating vs truncating mutations pan-cancer (Chapter 5). In initial CRISPR experiments, cells with complete ZFHX4 knockout may have undergone negative selection, leaving behind a population of hypomorphic clones (cells with in-frame indels or partial disruptions that preserved minimal function). No clones described in this chapter used for all ZFHX4-related work consistently displayed complete loss of ZFHX4 at the protein level by the HPA antibody. This is in line with a hypomorphic phenotype. Additionally, this idea is supported by the observation that many wells in single-cell cloning plates failed to grow (data not shown), hinting early lethality in cells with complete ZFHX4 loss. The survival of CC9 may reflect a rare case where adaptive survival mechanisms allowed persistence and eventual normalisation of proliferation. More experimentation here is required to definitively prove a reduced cloning efficiency as a result

of lethality by performing single-cell cloning in parallel to the wild-type parent immediately after CRISPR. Another large piece of evidence satisfying this hypothesis in my work is the loss of *ZFHX4*$^{-/-}$ in the metastatic colon organoid line. Despite an initial 100%-indel and 81%-KO score confirmation following CRISPR, two weeks of culture resulted in an unexpected >75% wild-type population expansion across all RNAseq replicates. This negative selection of *ZFHX4*$^{-/-}$ does suggest lethality but experiments exploring this such as viability and Sanger validation checks at multiple time points post-CRISPR are necessitated. Another experiment could describe GFP labelling through lentivirus, performing CRISPR, mixing 1:1 with isogenic mCherry-labelled control organoids and then observing populations overtime with fluorescence microscopy. Furthermore, it also suggests that timing of CRISPR and RNAseq as well as sufficient organoid material pre-CRISPR would be important considerations for future CRISPR knockout if not knockdown of *ZFHX4* in organoids.

There is precedent for this dosage sensitivity in the literature. An important ZFHX4 study reported throughout this thesis, the work in GBM linking ZFHX4 to NuRD and its proliferation phenotype was performed entirely using shRNA knockdown, not genetic knockouts highlighting the significance of partial suppression over complete loss. Similar patterns were observed in zebrafish, where knockdown of ZFHX4 prompted a more pronounced phenotype than knockout despite lower protein levels in the latter (Ishorst et al., 2025). This difference between knockdown and knockout in zebrafish has been reported before and highlights deleterious mutations through genomic editing can introduce unexpected phenotypic discrepancies as a result of compensatory mechanisms (Rossi et al., 2015). My western blots also revealed different banding patterns between knockdown and knockout samples suggesting complex post-transcriptional or post-translational regulation following different *ZFHX4*-peturbing mechanisms. Additionally, *ZFHX4*$^{-/-}$ mice generated by large *ZFHX4* deletions with PiggyBac and Cre-loxP technologies demonstrated post-natal lethality, consistent with ZFHX4 function essential for development

(Zhang et al., 2021a, Nakamura et al., 2021). Some of the strongest evidence supporting this hypothesis comes from the human germline data from neurodevelopmental disorder studies, concluding *ZFHX4* is highly LoF-intolerant, strongly selected against, dosage-sensitive and consistent with haploinsufficiency.

Together, these findings suggest that modulating ZFHX4 dosage rather than ablation may be a more reliable strategy for modelling its biology. This would also be a better comparison to the graded expression observed clinically between *ZFHX4*^{high} and *ZFHX4*^{low} tumours. Further work employing techniques such as CRISPR interference (CRISPRi), RNAi or CRISPR activation (CRISPRa)/overexpression may better reflect physiological regulation and phenotypes observed in clinical datasets. These observations also underscore the importance of validating genetic knockouts at multiple central dogma levels, especially for large and complex genes like *ZFHX4*. In my work I've demonstrated %-indel and %-KO scores from gDNA alone were poor predictors of transcript or protein loss, emphasising the need for cautious interpretation.

Although protein loss remains the gold-standard for confirming functional knockout, ZFHX4's large MW made western blotting technically challenging. Only a few human anti-ZFHX4 blots exist in the literature but these are either barely interpretable (Chudnovsky et al., 2014), don't report the antibody used (Xia et al., 2019), or don't provide enough detail on their parameters used for high MW blotting. The most convincing example (Perez Baca et al., 2025) used the HPA antibody and reported clean, discrete bands – hence I adopted their protocol through personal communication. Despite this, I was unable to reproducibly detect a single band at the expected 393kDa size. This may have been due to biological factors including model context, PTMs or protein stability/turnover within cells. It may also have been technical, relating to anomalous migration of large proteins, nucleic acid tethering or protein instability and/or degradation during sample preparation. Further experiments are required here to investigate these possibilities.

4.5.3 Other methods and techniques

Attempts to map ZFHX4-DNA binding by CUT&RUN were unsuccessful, which was disappointing given that integration with matched RNAseq could have provided definitive target identification. Importantly, both CUT&RUN experiments utilising two independent kits were technically successful as evidenced by robust H3K4me3 signal. Across all models, H3K4me3 peak-bearing genes correlated positively with high expression, confirming the assay captured real chromatin biology. An antibody-related failure was deemed unlikely. This is because I used the same HPA antibody that yielded >100 ZFHX4 peaks with the same CUT&RUN platform and enriched ZFHX4 by IP-MS in another study, albeit on healthy NPCs and neurons (Perez Baca et al., 2025). The more plausible explanation is context-dependent ZFHX4 biology which likely differs depending on chromatin state. In my differentiated cell line models, ZFHX4 may lack access to its targets due to closed chromatin or required cofactors. This is consistent with the broader pattern that ZFHX4-linked phenotypes are stronger in primary stem-like models than in differentiated ones, as discussed earlier. Supporting this, Baca et al. reported ~50% overlap of ZFHX4 peaks between NPCs and subsequently differentiated neurons, implying binding is highly state-dependent (Perez Baca et al., 2025). Given ZFHX4 expression in stem/progenitor epithelial subsets across normal and cancer settings, and its known involvement in embryonic and foetal development (Section 4.4.1), a reasonable hypothesis is that ZFHX4's functional DNA binding and downstream program require a stem-like chromatin context and specific cofactors. A clear next step is to repeat CUT&RUN with NPCs as a positive control and with primary-derived normal colon and cancer epithelium (such as organoid suspensions enriched for intestinal stem cell subsets) to test whether chromatin state and cell identity enable productive ZFHX4-DNA engagement.

Functional assays in this chapter spanning cell secretome analysis, IVTFAs and cell line migration and invasion offered insights into potential ZFHX4-regulated phenotypes but also came with their limitations. Secretome profiling revealed ZFHX4 loss reduced secretion of

pro-angiogenic factors such as VEGFA and Angiogenin, pointing to a potential angiogenesis phenotype, considered a key cancer hallmark (Hanahan and Weinberg, 2000, Hanahan, 2022). Consistent with this, IVTFAs generally trended towards reduced angiogenesis with *ZFHX4*$^{-/-}$ CM. However, across experiments these effects did not reach significance despite robust normalisation and pooling of replicates. The most likely reasons for this pertain to variability in CM technique (conditioning duration, medium type, cell number normalisation) and from the passage differences of cells from which it was obtained. It was also demonstrated that IVTFAs lacked sensitivity to subtle changes in factors such as VEGFA highlighting the difficulties in method to explore the angiogenic phenotype functionally. Other angiogenesis-related assays could've looked at HUVEC viability, proliferation and migration upon CM incubation. Despite this *in vitro* ambiguity, the Hallmark-Angiogenesis signature was significantly enriched in ZFHX4^{high} patients (NES -1.88, adjP=0.0008) across all four clinical cohorts in the meta-DEG analysis. Furthermore, ZFHX3 is an established regulator of angiogenesis in liver cancer (Fu et al., 2020). By analogy, ZFHX4 may exert a similar function, making angiogenesis a biologically plausible axis of its activity.

Migration and invasion assays remain inconclusive, largely because COLO320DM and NCIH716 lines were inherently non-migratory/invasive at baseline. SKNMC served as a robust mesenchymal-like positive control, confirming assay performance. While some intriguing shifts were seen (for example in CC9), these single-experiment trends are difficult to interpret given the passage variability described earlier. Future work should employ more motile colon cancer models that better align with the EMT-linked biology of ZFHX4.

4.5.4 Mechanistic insights of ZFHX4

ZFHX4 was shown to interact with CHD4 and the nucleosome remodelling and histone deacetylase (NuRD) complex in GBM (Chudnovsky et al., 2014), and this axis is plausibly relevant in colon cancer. Additionally, considering no NuRD members were detected in a

healthy NPC-based ZFHX4 IP-MS, ZFHX4-NuRD associations may be context- and cancer-specific (Perez Baca et al., 2025). This fits with extensive evidence for NuRD activity in colon cancer and with my data linking ZFHX4 to EMT – a process NuRD is known to modulate. Together, these observations support a model in which ZFHX4 engages NuRD to drive EMT and colon tumourigenesis.

ZFHX4 and EMT

EMT describes a functional transition of polarised epithelium into motile, ECM-secreting mesenchymal cells (Figure 4.44) (Kalluri and Weinberg, 2009). It is triggered in response to pleiotropic signals from the microenvironment which induce the expression of master EMT-TFs (TWIST1/2, ZEB1/2, SLUG and SNAIL) which further interact with post-transcriptional, post-translational and epigenetic regulators (Nieto et al., 2016). Ultimately this leads to changes in gene expression and consequent downregulation of epithelial features such as cell-cell junctions, apical-basal polarity and basement membrane interactions (Yang et al., 2020). Eventually, mesenchymal-like cells display fibroblastic morphology and enhanced migratory capacity with often invasive properties. Sometimes these can also gain enhanced migratory speed and become single migrating cells in an ameboid state (Graziani et al., 2022). A variety of markers are known to roughly characterise each EMT state which are regarded as diverse and plastic, as cells can exist in a ‘partial’ or ‘hybrid’ continuum displaying collective epithelial migration with mesenchymal invasiveness (Nieto et al., 2016, Nie et al., 2025). Both EMT and its reverse mesenchymal to epithelial transition (MET) are essential for normal embryonic development and adult tissue repair but regularly aberrant in cancer (Thiery et al., 2009). This is because it provides invasive phenotypes and metastatic potential, permitting dissemination to distant sites and through its plasticity can halt migration for malignant colonisation. Particularly in colon cancer, EMT signatures are poorly prognostic, confer stem-like properties and therapeutic resistance, regulating both disease progression and metastasis (Nie et al., 2025).

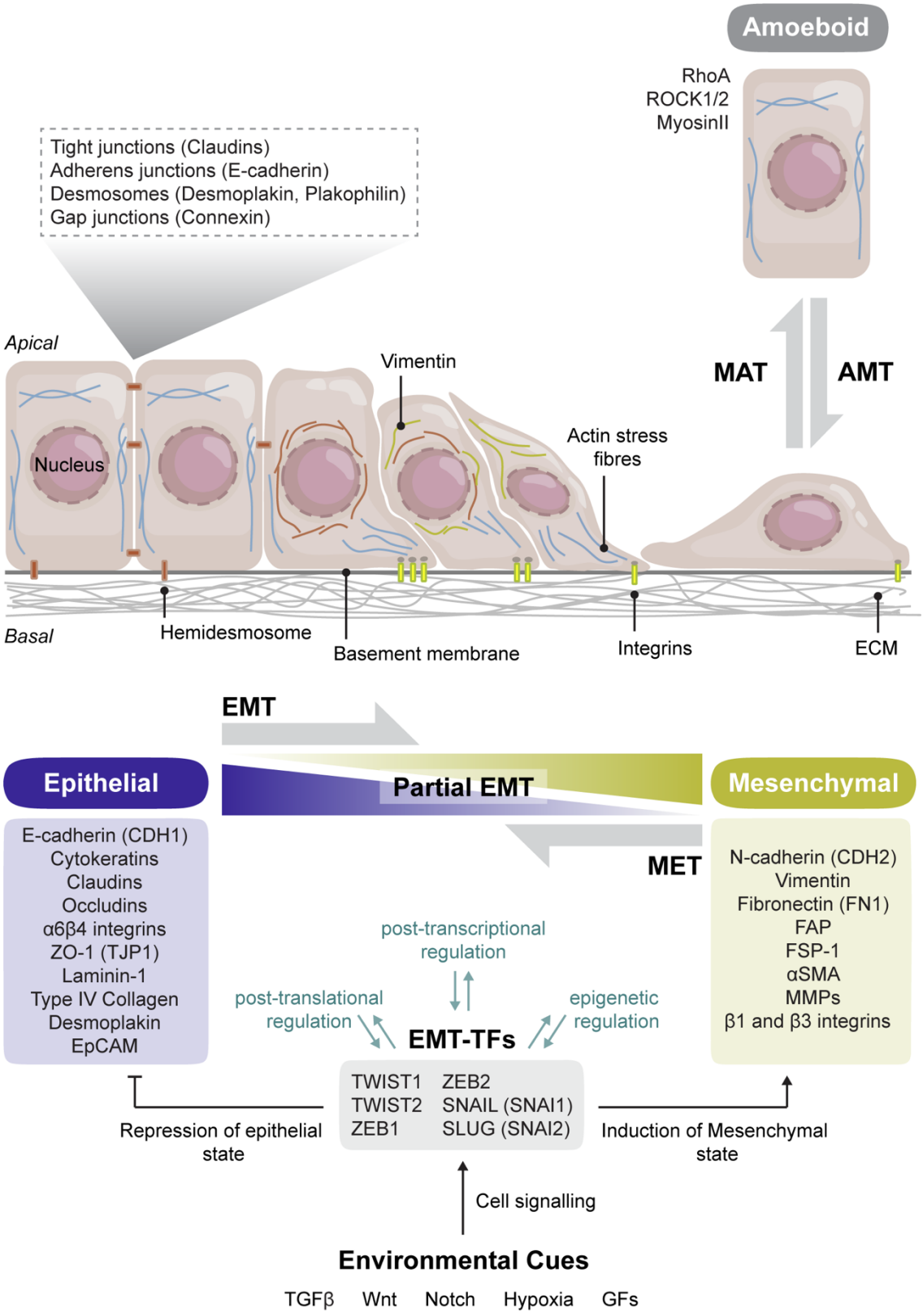


Figure 4.44 | The translation from epithelial to mesenchymal cell states through a multi-layer regulatory network, known as EMT. The figure was adapted from multiple seminal reviews on EMT (Nie et al., 2025, Du et al., 2017, Yang et al., 2020, Kalluri and Weinberg, 2009, Graziani et al., 2022, Dongre and Weinberg, 2019, Zhang et al., 2021b, Lu and Kang, 2019) *[EMT=epithelial-mesenchymal transition, TF=transcription factor, MAT=mesenchymal to ameboid transition, AMT=ameboid to mesenchymal transition, MET=mesenchymal to epithelial transition]*

Supporting evidence linking ZFHX4 with EMT is the consistent association of EMT with *ZFHX4*^{high} clinical phenotypes, and downregulation of CDH2, FN1 and ZEB2 genes in NCIH716 *ZFHX4*$^{-/-}$ models. Moreover, a known epigenetic regulator of EMT, Runx2 was identified to interact with Zfhx4 through protein-protein interactions in mice chondrocytes (Nakamura et al., 2021) and both appear to be similarly epigenetically regulated by the super-enhancer complex Zfp260 in mouse skeletal stem cells (Weng et al., 2024). Furthermore, not only reemphasising the transcriptional and epigenetic heterogeneity within cell lines, one study using SW480 cells highlighted two subpopulations with distinct Wnt and EMT activity (Yi et al., 2020). Investigators found this was regulated by RUNX2 which altered chromatin landscape and bound EMT-associated genes, causing transcriptional activation as well as enhanced metastasis in mice. Taken together this could highlight similar functions of RUNX2 and ZFXH4 particular in driving EMT.

Other supporting evidence of ZFHX4 with a role in EMT comes from its other ZF-TF family members ZFHX1A and ZFHX1B, otherwise known as ZEB1 and ZEB2, both known canonical drivers of EMT (Dongre and Weinberg, 2019). ZFHX4 was also shown to bind ZEB1 in IP-MS (Perez Baca et al., 2025). In lung adenocarcinoma, ZFHX4 was found to positively regulate the TF TWIST1 mRNA expression through a competing endogenous RNA (ceRNA) effect rather than a TF-related effect (Xia et al., 2019). Here, in ZFHX4 mRNA-high contexts, miR-514a-3p is bound and sequestered to greater levels by ZFHX4 mRNA molecules, prompting reduced repression on TWIST1, enhancing TWIST1 mRNA and protein levels and thus EMT phenotype (Xia et al., 2019). Collectively, both literature and work in this chapter presents a likely link between ZFHX4 and EMT.

The NuRD complex in colon cancer

The NuRD complex is a multifaceted chromatin remodelling complex that utilises ATP hydrolysis to alter chromatin state by destabilising, moving, ejecting or restructuring the repeating units of chromatin structure, known as nucleosomes (Clapier and Cairns, 2009).

Chromatin remodellers such as NuRD are essential to the cell and provide dynamic regulation of DNA repair, replication, transcription and chromosome segregation primarily through changing access to and composition of packaged DNA (Clapier and Cairns, 2009). This is mediated through recruitment to specific loci by TFs and methylated DNA (Lai and Wade, 2011). Highly conserved across animals and plants with broad expression in most tissues, the stoichiometric assembly of six core protein subunit modules (CHD, MBD, RBBP, HDAC, MTA, GATAD) are paramount for conferring its functional specificity and both co-repressor and co-activator functions (Ramirez and Hagman, 2009, Low et al., 2020). This is predominantly mediated through paralog switching where subunit paralogs can exist in different combinations, for example CHD3, CHD4 or CHD5 for the CHD module, and this can drive specific functional outcomes in different cellular contexts (Reid et al., 2023). Similar to other chromatin remodelling complexes, NuRD is essential for normal development and homeostasis such as the maintenance of pluripotency (Kaji et al., 2006) and differentiation programmes (Zhu et al., 2009). Therefore, it has unsurprisingly been linked to tumourigenesis, specifically through the association of subunit expression to oncogenic or tumour-suppressive phenotypes and suggest complex roles for both promoting and suppressing tumour growth (Reid et al., 2023, Bornelov et al., 2018, Ramirez and Hagman, 2009, Lai and Wade, 2011).

In colon cancer, the NuRD member MBD2 has been associated with intestinal tumourigenesis, where deficiency in Apc^{min} mice suppressed tumourigenesis indicating a pro-tumourigenic role (Sansom et al., 2003, Sansom et al., 2007). Interestingly, the authors suggested tumourigenesis in MBD2-functional settings may be due to MBD2-mediated transcriptional repression of important negative Wnt regulators such as secreted frizzled-related protein (SFRP) genes previously reported as epigenetically inactivated in colon cancer (Suzuki et al., 2004, Cai et al., 2014). I identified *SFRP1*, *SFRP2* and *SFRP4* as some of the top genes positively associated with *ZFHX4* expression in the four clinical datasets, however as negative Wnt regulators this seems counterintuitive considering the

prediction that ZFHX4 is positively associated with tumourigenesis. Despite this, further investigation into SFRP genes here may be of interest as some investigators also suggest a Wnt-independent role involved in BMP signalling and collagen processing which may relate to the EMT phenotype (Nathan and Tzahor, 2009). Most importantly, SFRP2 has been shown to contribute to tumour aggressiveness through its expression in stromal CAF populations rather than epithelium, prompting immune suppression by neutrophil recruitment (Kasashima et al., 2021), Treg differentiation (Wang et al., 2025) and inhibiting $CD8^+$ T-cell infiltration (Zhang et al., 2025). In another study, SFRP1 secreted by colon tumour epithelium promoted expansion of pro-invasive CAF populations which restricted tumour growth but encouraged invasiveness (Mosa et al., 2020). Investigators also revealed SFRP1 induced EMT in tumour cells without disrupting intrinsic Wnt signalling, consistent with a $ZFHX4^{high}$ phenotype and underscoring further work in utilising scRNAseq for disentangling ZFHX4 roles in epithelial and stromal compartments where crosstalk is likely complex.

Conversely, the NuRD member MBD3 has been shown to interact with unphosphorylated c-Jun to repress AP-1 target genes such as the intestinal stem cell marker Lgr5 (Aguilera et al., 2011). Gut-specific conditional Mbd3 deletion in mice led to stimulated c-Jun activity, progenitor cell proliferation and susceptibility to colitis-induced tumourigenesis (Aguilera et al., 2011). Taken together, this reemphasises opposing roles of NuRD in facilitating tumourigenesis through its composition of subunit members. Furthermore, it also suggests NuRD is implicated in colon tumourigenesis, and it's plausible that ZFHX4 could mediate its pro-tumourigenic effects through NuRD recruitment considering it does this in GBM through CHD4 interactions (Chudnovsky et al., 2014).

NuRD and EMT

There are several associations of NuRD linked to EMT in the literature. In HEK cells TWIST1 has been shown to interact with many NuRD subunits and regulate transcription

of EMT-related genes such as repression of E-cadherin (Fu et al., 2011). Repression of epithelial target genes by non-acetylated TWIST1-NuRD complexes have also been reported in breast MCF7 cells, where diacetylated TWIST1 can also associate with other epigenetic complexes such as TIP60 and activate mesenchymal genes, enhancing metastatic capability in mouse xenografts (Yu et al., 2023). Some investigators demonstrated ZEB1 interacts with NuRD via CHD4 through acetylation of its K811 residue in a co-repressor manner, suppressing a GTPase activating protein (TBC1D2b), promoting E-cadherin internalisation and enhancing *in vivo* metastasis in lung adenocarcinoma (Manshouri et al., 2019, Perez-Oquendo et al., 2023). Another study in cervical cancer showed SNAIL interacted with PRMT5 and NuRD to repress E-cadherin expression, and the subsequent EMT phenotype was reversed through PRMT5 inhibition (Gao et al., 2021). This highlights not only protein-protein binding but also PTMs are critical for EMT-TFs to exert distinct functions mediated through NuRD interactions and suggests ZFHX4 could similarly demonstrate the same effect.

Collectively, it's therefore plausible to suggest ZFHX4 could regulate EMT through interactions with the NuRD complex, enhancing colon tumourigenesis in both epithelial and stromal compartments. Key experiments to explore this in colon cancer using a variety of models would involve:

(1) Determining if ZFHX4 physically interacts with DNA and regulates gene transcription (ZFHX4 CUT&RUN);

(2) Proving ZFHX4 physically interacts with NuRD (IP-MS);

(3) Showing ZFHX4 and NuRD co-occupy EMT loci (NuRD subunit CUT&RUN);

(4) Demonstrating (A) ZFHX4 drives an EMT phenotype and (B) that this depends on NuRD (KD and KO techniques, and WB/IF);

(5) Determining compartment specificity with regards to epithelium and stroma (co-culture models)

4.5.5 Conclusion

In summary, my work argues ZFHX4 is likely a dose-sensitive, context-dependent regulator of colon tumour biology. Clinically, *ZFHX4*^{high} tumours associate with EMT and stromal-related programs, angiogenesis and poor prognosis, whereas *ZFHX4*^{low} contexts align with epithelial identity and oxidative metabolism. *In vitro*, seemingly straightforward assays proved technically challenging, especially with ZFHX4 protein detection which offered more questions than answers. Organoid and cell line models provided their own strengths and weaknesses but did permit large-scale genotyping, phenotyping and functional assays to answer plausible hypotheses. Despite the lack of a concrete phenotype with this work, converging evidence from primary tissues through IHC and scRNAseq and a plausible NuRD-mediated EMT mechanism likely involving microenvironmental crosstalk argues an interesting role for ZFHX4 in colon tumourigenesis. Given this thesis' original focus on immunogenicity, these findings broaden ZFHX4's translational relevance and provides many avenues for further work. They also underscore the novelty of this work, as ZFHX4 remains poorly characterised in general and its role in colon cancer hasn't been previously studied.

Chapter Five: ZFHX4 as a potential regulator of immunogenicity

5.1 Introduction

Considering the potential of single-dose immunotherapy to effectuate complete durable remission particularly in late-stage cancer patients (Kurian and Markabawi, 2024), understanding the biology of this response is crucial to improve clinical response in more patients. In immuno-oncology, progress has largely been driven by biomarker research, where tumour-associated characteristics are linked to treatment response and guide clinical decisions, much like tumourigenic biomarkers inform targeted therapy. Whilst more general characteristics such as TMB have propelled immunotherapy research in the past decade (Samstein et al., 2019), more specific genomic and transcriptomic biomarkers are emerging to associate with immunotherapy response as a function of gene-altering processes rather than indirect perturbations from mutational processes (Ishino et al., 2023, Wang et al., 2023).

Most biomarkers have been mined from genomic and transcriptomic sequencing of patients that later undergo immunotherapy treatment and thus can be associated with response. Because immunotherapy was first applied in melanoma and lung adenocarcinoma, most datasets represent these cancers treated with anti-PD1 or anti-CTLA4, although datasets from other malignancies are now emerging. These indications include head and neck, bladder, renal and colorectal cancer and further comprehensive analyses identifying biomarkers across large or harmonised cohorts are emerging too (Litchfield et al., 2021, Thakkar et al., 2024, Mao et al., 2023, Usset et al., 2024).

Interestingly, one of these meta-analyses identified *ZFHX4* mutations correlating with high inferred immunotherapy objective response rates and significantly conferring better overall survival in melanoma (HR=0.69[95%CI:0.50-0.98], p=0.0358) and progression-free survival in non-small cell lung carcinoma (NSCLC; HR=0.30[95%CI:0.15-0.61], p=0.000472) (Mao et al., 2023). Despite the potential confounding of TMB in these results, an independent

cohort demonstrated significant enrichment of *ZFHX4* mutations in 14% more patients with clinical benefit compared to no clinical benefit following immunotherapy even in the TMBL setting (Fisher's p=0.0261, Table 3.3, Chapter Three) (Miao et al., 2018). Furthermore, analysing three independent NSCLC public cohorts, *ZFHX4* mutations remained an independent predictor of immunotherapy response despite TMB and PD-L1 correction (Shen et al., 2022). Interestingly, stratifying into four patient groups based on TMB and *ZFHX4* status with reference to TMBL-ZFHX4-WT, *ZFHX4* mutations provided a significant 25% survival advantage alongside TMBH (HR=0.30[95%CI:0.14-0.64], p=0.002) compared to TBMH-ZFHX4-WT (HR=0.55[95%CI:0.26-1.17], p=0.12) which was reflected in another meta-analysis on 10 cohorts (Shen et al., 2022, Fu et al., 2024). Irrespective of a significant association between *ZFHX4* mutations and TMB as found in both studies and as expected with large genes such as *ZFHX4*, these data highlight the predictive value of *ZFHX4* mutational status in immunotherapy response regardless of TMB. Specifically in COAD, datasets such as these are limited especially in the nHM setting due to the relatively recent implementation of immunotherapy for this indication, however newly emerging datasets with matching genomic and transcriptomic alongside immunotherapy response data require further investigation (Thakkar et al., 2024, Taraborrelli et al., 2023).

Following the identification of enriched *ZFHX4* mutations in a nHM TCGA-COAD subgroup with heightened immunogenicity (ICS3; Chapter Three), and with aforementioned published strong evidence of *ZFHX4's* association with immunotherapy response, ZFHX4 was further explored *in silico* and *in vitro* to ascertain a mechanism of immunogenicity.

5.2 Hypothesis

ZFHX4 mutations are associated with immunogenicity in many clinical COAD datasets, and using cell line-based *ZFHX4*$^{-/-}$ models as a surrogate for mutation, *ZFHX4*$^{-/-}$ conditioned medium (CM) will enhance T-cell activation and migration phenotypes

Aims

1. Investigate mutational spectra of *ZFHX4* across four independent COAD cohorts including TCGA
2. Investigate potential of *ZFHX4*$^{-/-}$ CM generated from cell line models to enhance an activated phenotype in healthy donor T-cells as a proxy for enhanced immunogenicity
3. Investigate potential of CM in aim-2 to enhance the migratory phenotype of healthy donor T-cells as a proxy for enhanced immunogenicity

5.3 Experimental Design

Firstly, the mutational spectra of ZFHX4 was investigated across four independent COAD cohorts including TCGA to identify clinically relevant *ZFHX4* mutations affecting common amino acid (aa) residues. Subsequently, the cell line models employed in Chapter Four were investigated to establish if *ZFHX4*$^{-/-}$ was associated with a more immunogenic state. This was predominantly through cell line conditioned medium approaches and healthy donor pan-T-cell populations to investigate potential changes to T-cell migration via Transwell assays or T-cell activation and exhaustion via flow cytometry.

5.4 Results

5.4.1 ZFHX4 mutations in cancer and COAD

To observe *ZFHX4* mutational spectra pan-cancer irrespective of TMB, the TCGA Pan-Cancer Atlas spanning 10953 patients with 70 different cancers was investigated using cBioPortal. *ZFHX4* presented the top 12th mutated gene of all cancers, displaying a somatic small-variant (<49bp) mutation frequency of 9.4% and 87% predominance for missense SNVs across the entire gene (Figure 5.1A). All missense SNVs are variants of unknown significance (VUS) and whilst some missense mutations affecting the same codon occur in up to 8 patients, they don't appear to confine to specific hotspots. Additionally, collating four independent COAD datasets including TCGA, *ZFHX4* is mutated (*ZFHX4*^{mt}) in ~19% COAD patients on average and 12% in TMBL tumours, with missense mutations also predominating (Figure 5.1B).

Following Fisher's exact testing on TMBL vs TMBH and *ZFHX4*^{wt} vs *ZFHX4*^{mt} proportions in COAD cohorts, *ZFHX4* mutations were found significantly enriched in TMBH tumours across nearly all cohorts (TCGA $p=1.37x10^{-12}$, Nunes $p<2.2x10^{-16}$, SIDRA $p=7.44x10^{-7}$, CPTAC2 p=0.069). This enrichment was expected considering its large size and thus higher probability of mutation in hypermutated contexts such as MSI. However, more than 10% of TMBL-COAD tumours harboured ZFHX4 mutations across all four cohorts, indicating these alterations are not solely artefacts of hypermutation and may contribute functionally to tumourigenesis and/or immunogenicity. *ZFHX4* is not currently recognised as a cancer driver by COSMIC (Sondka et al., 2024) or bioinformatic tools such as IntOGen (Martinez-Jimenez et al., 2020), likely reflecting the paucity of functional evidence for its predominantly missense variants catalogued as VUS. Though sub-setting for TMBL tumours, *ZFHX4* ranked the 21st most frequently mutated gene out of 19122 mutated genes, underscoring the need for further investigation given the research emphasis already placed on other top-ranking cancer-associated genes.


a
Pan-Cancer (subtype and TMB agnostic)
Missense (1412)
Truncating (190)
In-frame (4)
Splice (4)
patients
R361 L411 G754 R783 S1068 R1422 K1442 P1723 R1983 E2641 R2607 R2647 R2933 R3066
Q86UP3-5
Sumoylation
Phosphorylation
O-linked Glycosylation
Acetylation
b
All Samples
nHM/TMBL
TCGA 19.1% [341] 12.2% [294]
Nunes 21.6% [597] 10.4% [450]
SIDRA 13.8% [224] 8.02% [187]
CPTAC2 20.5% [78] 15.9% [63]
Combined mean 18.75% [1240] mean 11.63% [994]
Missense
Frameshift DEL
Nonsense
Frameshift INS
In-frame DEL

Figure 5.1 | Mutational spectra of *ZFHX4* pan-cancer and in four independent COAD datasets additionally segregated into TMBL groups. (A) Lollipop plot of *ZFHX4* mutations aligned to the Q86UP3-5 isoform and inferred PTMs obtained from Figure 4.1. Common mutations affecting the same aa residue in >5 patients are annotated. Data is obtained directly from cBioPortal TCGA PanCancer Atlas. Truncating mutations refer to nonsense, nonstop, FS ins and FS del. **(B)** Lollipop plots of *ZFHX4* mutations aligned to the Q86UP3-5 isoform in different COAD datasets. Y axis depicts number of patients. Values in grey depict small variant frequency and total n number of the cohort in square brackets. All cohorts are aligned to allow direct cross-cohort comparison, and all mutations were combined into a single lollipop plot. *[FS=frameshift]*

In the combined COAD lollipop plot, *ZFHX4* mutations clustered around N-terminal/exon-2 and homeodomain/exon-10 regions, with L411 residues mutated in 8 COAD patients and identified as a top pan-cancer altered residue. To assess positional enrichment, the most frequently altered residues across the four independent cohorts were compiled (Table 5.1). Considering the probability a mutation hits a specific aa residue in *ZFHX4* ($P(m) = \frac{1}{3616} = 0.000276$) and a null hypothesis that mutations are randomly distributed across residues, the expected number of patients to have an affected residue would be ~0.34 ($1240 \times P(m)$) and ~0.28 ($994 \times P(m)$) in the all-sample and nHM-cohorts respectively. Comparing to observed patient numbers all listed residues exceeded these expectations, however binomial testing with multiple-testing correction (α=0.05) required ≥6 (all-sample) and ≥5 (nHM) patients with the same residue for significance; only L411 in the all-sample cohort met this threshold. Interestingly, most recurrently mutated residues were leucine (L), arginine (R), alanine (A) or glycine (G) and over half were located within functional domains.

Table 5.1 | Top 10 commonly altered residues in *ZFHX4* by mutations in all four COAD cohorts

All COAD samples (n=1240)				**nHM/TMBL COAD samples** (n=994)			
Residue	**Domain**	**N patients**	**N cohorts**	**Residue**	**Domain**	**N patients**	**N cohorts**
L411	na	12	3	L2458	ZF-19	3	3
L2865	HD-4	4	3	R2607	HD-3	3	2
L2458	ZF-19	3	3	A3533	na	3	2
R2647	ZF-20	5	2	G355	na	2	2
T2934	HD-4	4	2	A386	na	2	2
K340	na	3	2	G671	ZF-4	2	2
A386	na	3	2	G2914	HD-4	2	2
R2607	HD-3	3	2	L27	na	2	1
G2914	HD-4	3	2	R768	ZF-6	2	1
A3533	na	3	2	R2131	HD-1	2	1

Domain annotation was provided by InterPro (Blum et al., 2025). [ZF=zinc-finger, HD=homeodomain]

When examining mutation types across all samples versus TMBL samples in the four COAD cohorts, significant differences were observed (two-sided Fisher's test, p=0.000175). For example, missense mutations were significantly enriched in the TMBL setting, occurring at 15% higher frequency. Conversely, frameshift deletions, insertions and nonsense mutations were proportionally depleted in TMBL by 8.71%, 4.67% and 1.96% respectively. This suggests a potential negative selection for frameshift-inducing/truncating or likewise positive selection for non-frameshift-inducing/truncating mutations in the TMBL setting.

Taken together, these data suggest a prevalence of *ZFHX4* mutations particularly in the TMBL setting across multiple independent COAD cohorts. They are predominantly missense in nature and weakly cluster around specific residues notably in the N-terminal region and within defined functional domains. Although the precise functional consequences remain unclear, these mutational spectra collectively point toward a plausible role in altering ZFHX4 function.

5.4.2 Obtaining a pan-T population and testing artificial activation technique

To study immunogenicity, healthy donor T-cells were used as an *in vitro* model and flow cytometry was largely used as the investigative method of choice. These were obtained regularly from the Seymour group (Department of Oncology, University of Oxford) in a population of lymphocytes which underwent negative selection to obtain pan-T-cell populations. Acknowledging that T-cell binding MHC/antigen on antigen presenting cells (APCs) by T-cell receptor (TCR) and downstream activation and effector functions are a primary readout of immunogenicity *in vivo*, human anti-CD3/CD28 Dynabeads were employed to induce artificial T-cell activation *in vitro* by mimicking MHC/TCR and B7/CD28 interactions required for T-cell activation and effector function (Figure 5.2A). Standard artificial activation consisted of 1:1 T:bead ratios with 3-day incubations alongside recombinant IL-2 for sustaining proliferation, and these cells were termed ‘activated’ (A). This consistently prompted observable changes in culture comparative to ‘non-activated’ (NA) cells used as parallel controls, demonstrating significant clumping, roughly 2-fold increases in cell size and a prominent irregular shape (Figure 5.2B). Furthermore, obtaining pan-T populations by negative selection was consistently highly successful, achieving >95% CD3/TCR+ cells (Figure 5.2C). In every experiment, forward (FSC) and side (SSC) scatter flow cytometry plots demonstrated distinct populations of NA and A cells, with A cells displaying higher FSC and SSC in the plot, consistent with changes observable by eye (Figure 5.2C).

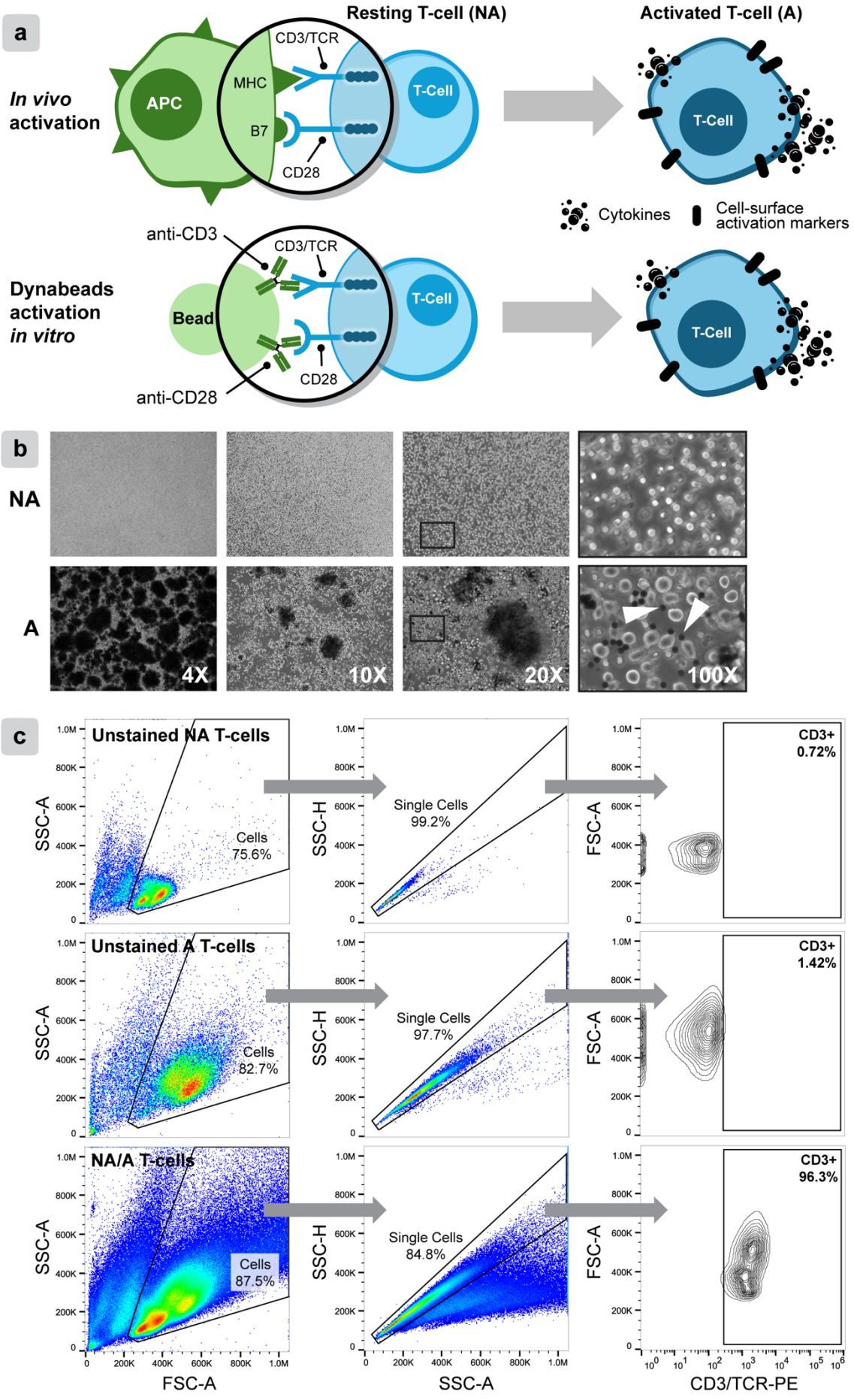

a
Resting T-cell (NA)
Activated T-cell (A)
In vivo activation
APC
MHC
B7
CD3/TCR
CD28
T-Cell
Cytokines
Cell-surface activation markers
Dynabeads activation in vitro
anti-CD3
anti-CD28
Bead
b
NA
A
4X
10X
20X
100X
c
Unstained NA T-cells
Cells 75.6%
Single Cells 99.2%
CD3+ 0.72%
Unstained A T-cells
Cells 82.7%
Single Cells 97.7%
CD3+ 1.42%
NA/A T-cells
Cells 87.5%
Single Cells 84.8%
CD3+ 96.3%
SSC-A
SSC-H
FSC-A
CD3/TCR-PE

Figure 5.2 | Healthy donor pan-T-cells and artificial activation by Dynabeads *in vitro.* (A) Schematic demonstrating the mechanism of anti-CD3/CD28 Dynabeads mimicking *in vivo* APC/T-cell interactions. **(B)** Microscope images at increasing magnifications taken after a 3-day incubation of 1:1 T:bead cultures. White arrows indicate observable black-coloured Dynabeads. 20X images with were further zoomed to 100X as by black boxes to inspect cell size and shape. **(C)** Gating of unstained NA, unstained A and stained NA+A samples following flow cytometry, obtaining cells, single cells and CD3+ cells with identical gates. The CD3+ gate was selected based on the unstained controls (first two rows). *[NA=non-activated, A=activated, TCR=T-cell receptor, APC=antigen presenting cell]*

Accompanying visual confirmation of artificially activated T-cells, key activation markers CD69, CD25 and PD-1 were tested with flow cytometry. In most experiments, a live-dead (LD) control was created by heat-killing 50% sample and mixing it back into the original sample so that identifiable live and dead peaks could define a live-cell gate (Figure 5.3A). In an initial experiment, 79-fold, 45-fold and 18-fold A cells expressed CD25, CD69 and PD-1 respectively comparative to NA cells, suggesting the success of Dynabeads in creating an artificially activated phenotype (Figure 5.3B). Due to sample consisting of a NA/A mix here, populations were gated based on distinct size in FSC/SSC plots.

The activation markers presented a spread of expression rather than binary, thus antibody titration was performed using 1:50 to 1:800 dilutions. Across CD25, CD69 and PD-1 antibodies, 1:100 provided the most optimal dilution as a compromise between the highest separation index (Methods Section 2.9) and lower concentration for antibody conservation.

Taken together, optimised methods for isolating pan-T-cell populations and artificial activation of T-cells were robust and provided the basis for all further experiments to investigate the effects of *ZFHX4*$^{-/-}$ on immunogenicity.

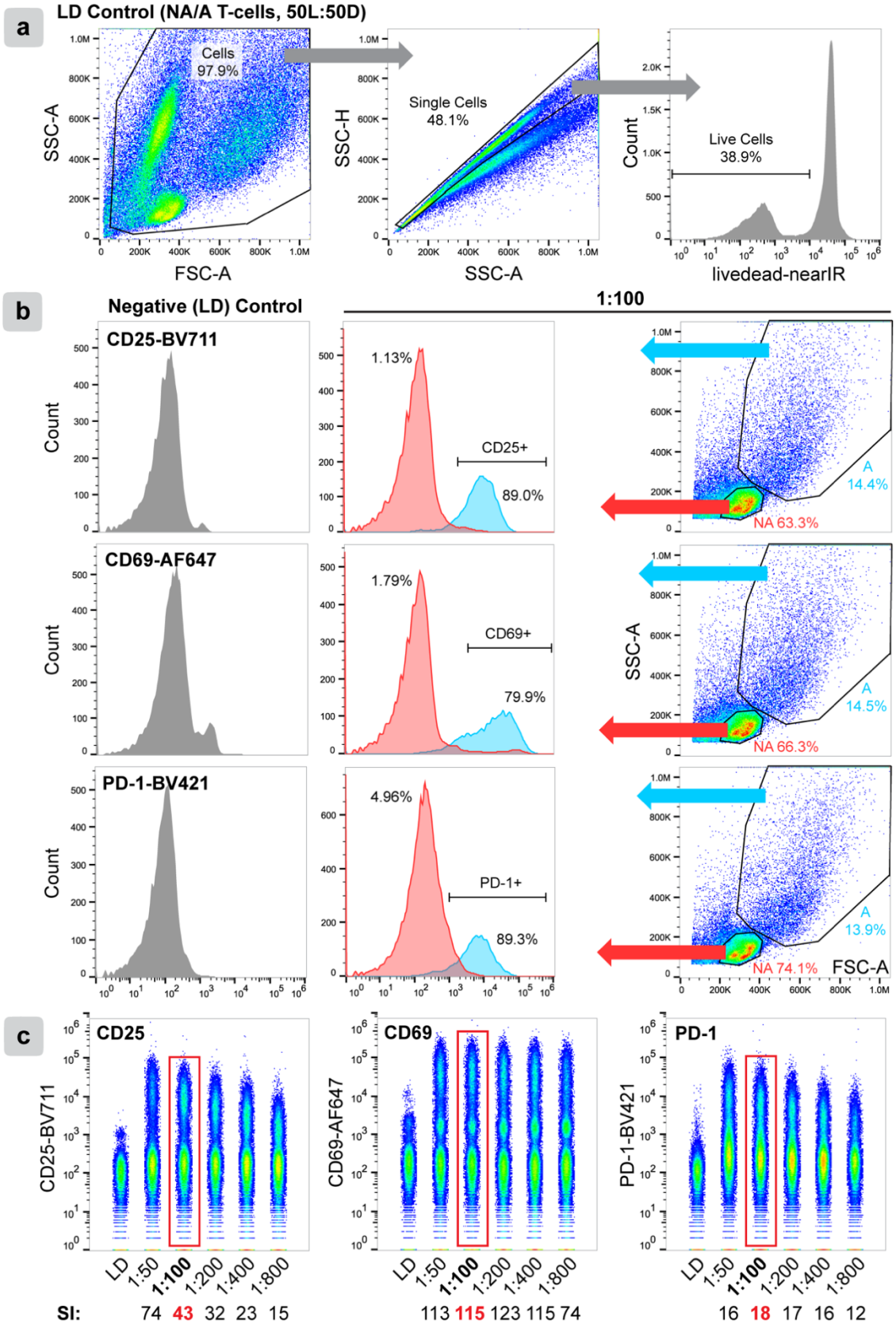


Figure 5.3 | Investigating activation marker expression following Dynabead incubation. (A) Flow cytometry gating strategy to obtain live single T-cells from one donor. **(B)** Gating of CD25, CD69 and PD-1 based on unstained LD controls. NA (red) and A (blue) populations were annotated based on expected FSC and SSC distribution. **(C)** Antibody titration with separation index calculations, highlighting the chosen concentration in red. *[LD=live-dead]*

5.4.3 Effects of *ZFHX4*$^{-/-}$ conditioned medium on T-cell activation and exhaustion

Dynabead-activated healthy T-cells were employed in this chapter because my cell line models lack autologous T-cell models and my project was largely focused on antigen-independent immunogenicity. Instead, I asked whether *ZFHX4*$^{-/-}$ epithelium alters T-cells activation, exhaustion or migratory state, hypothesising an overall enhancement in immunogenicity. Due to utilisation of CM in other experiments such as those in Chapter Four, CM was also employed here to examine anything unique to *ZFHX4*$^{-/-}$ CM that can alter T-cell activation and exhaustion, despite not identifying immunogenic factors in the cytokine arrays which were limited to around 100 cytokines.

To test this, NA cells were incubated with CM from two passages (P1; 10m cells, P2; current-3m cells) to assess activation, while A cells were incubated in parallel to evaluate reduced activation, exhaustion or further activation beyond Dynabeads. This was performed at time 0, describing freshly isolated pan-T-cells incubated with (A) or without (NA) Dynabeads 1:1 + IL-2 within specific CM or control medias. Following a 3-day incubation, flow cytometry was performed on a much larger panel with further optimised concentrations as Figure 5.3 (data not shown), additionally utilising IFNγ as an intracellular marker of activation and CD4/CD8 markers to identify specific T-cell populations. Furthermore, three independent T-cell donors (D1/D2/D3) and different media passage replicates (P1/P2) were used to enhance reliability of this single experiment.

As a consequence of large panel size and expected spillover across channels (Supplementary Figure S5.1), compensation was performed within Attune NxT software and quality-control (QC) was performed on FlowJo to examine if this was successful. Using A cells from D1 as an example, compensation appeared successful as indicated by vertical and horizontal populations aligned to biexponential axes with no diagonal spill in different channels (Figure 5.4).

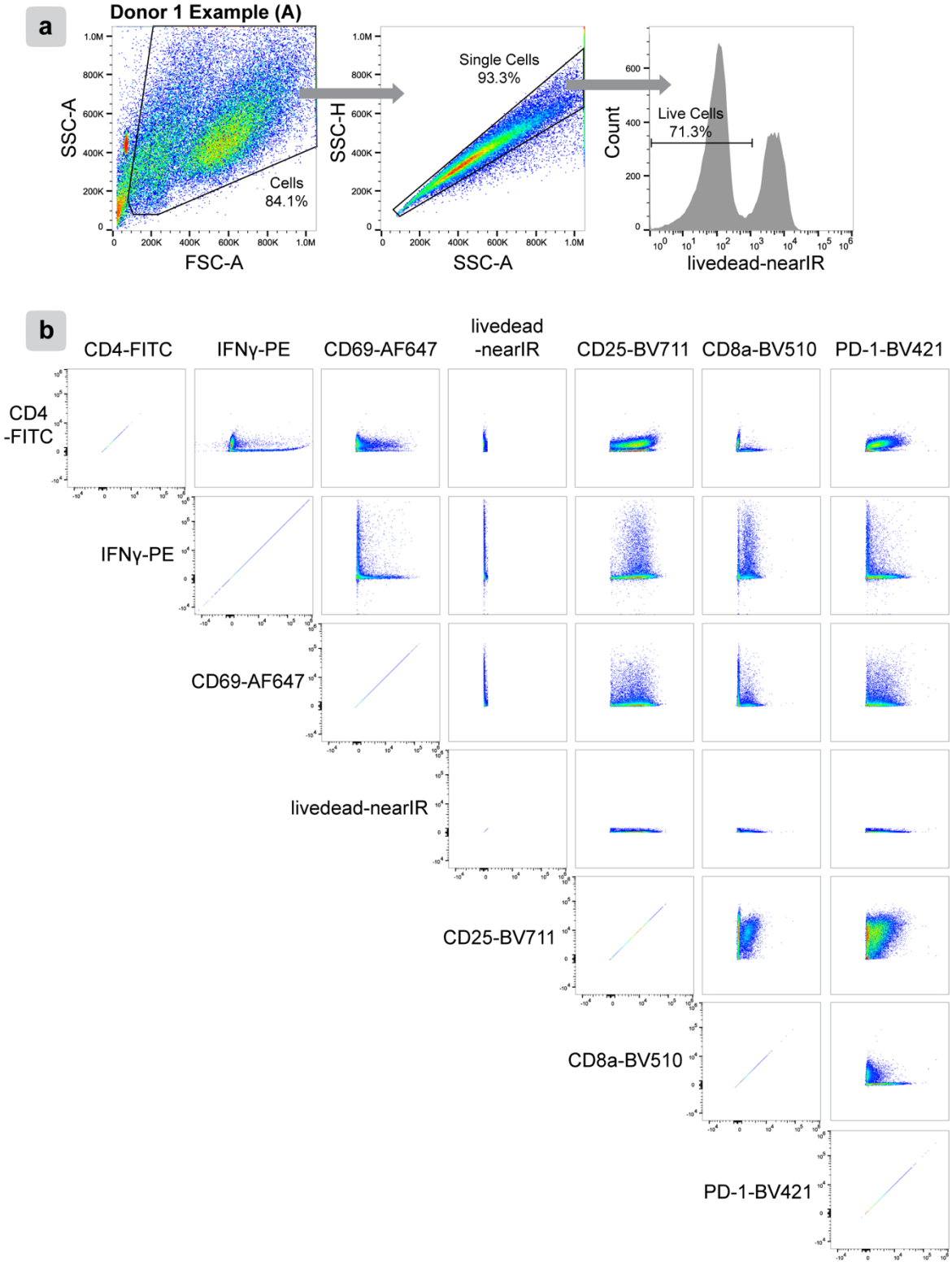
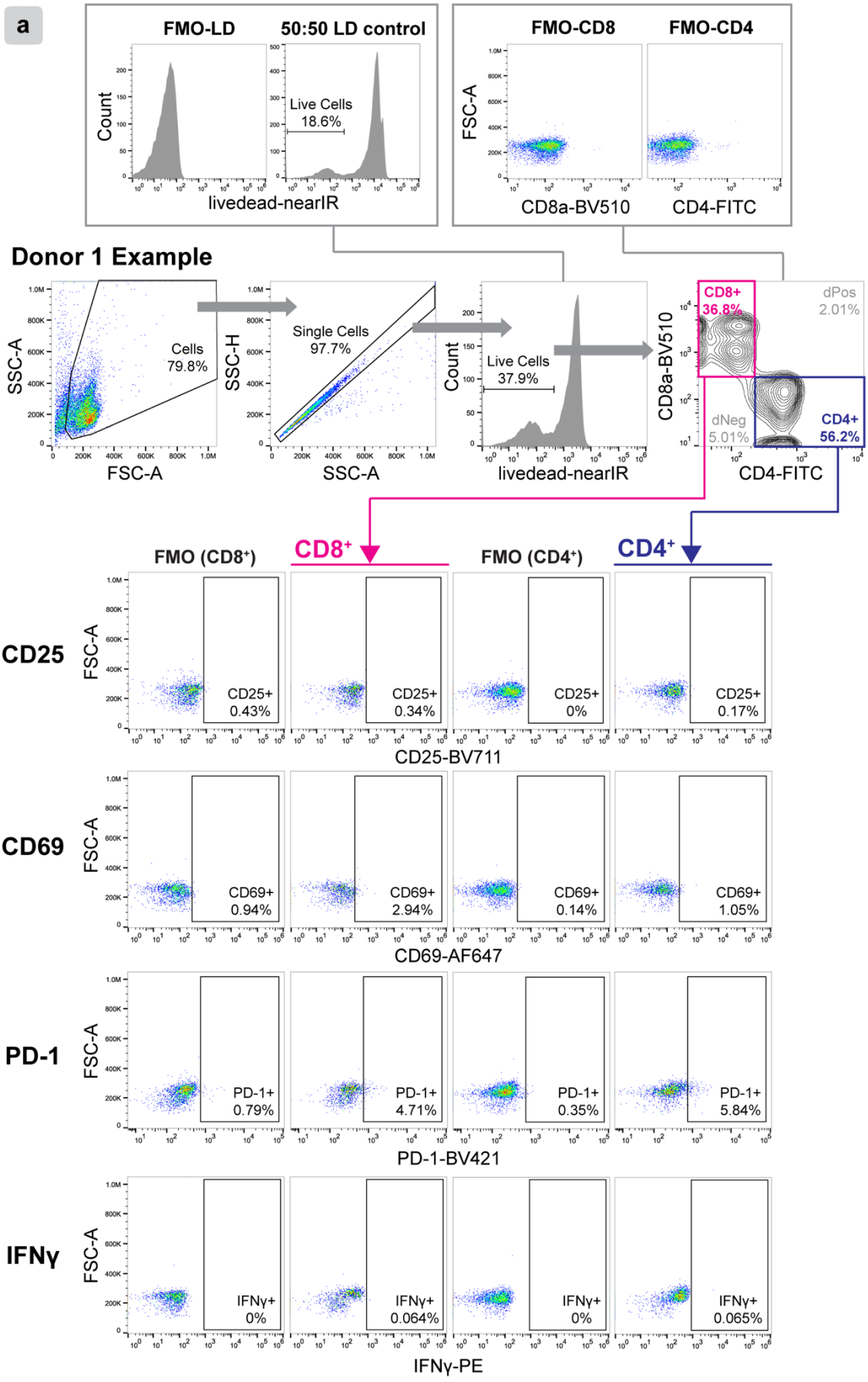


Figure 5.4 | QC of instrument-based compensation using A cells from D1 as an example. (A) Flow cytometry gating strategy to select live single cells so that dead cells don't confound autofluorescence. **(B)** Comparison matrix of bivariate plots of all markers used in the flow panel (n=7), using biexponential axes to observe potential bending or diagonals as a result of inadequate compensation. *[A=activated, D1=donor 1]*

As NA and A cells display different sizes and thus shifts in fluorescence even with identical antibody staining, NA (Figure 5.5A) and A (Figure 5.5B) gating strategies followed the same principles but with slightly different gate thresholds to select positive live populations. Firstly, single live cells were selected based on FSC/SSC-A/H plots and live-dead controls. $CD4^+CD8^-$ (referred to as $CD4^+$ T-cells) and $CD4^-CD8^+$ (referred to as $CD8^+$ T-cells) gates were drawn based on respective FMOs. Further gating was performed within these populations applying the same gates for CD25, CD69, PD-1 and IFNγ based on respective FMOs subset to either $CD8^+$ or $CD4^+$ T-cells. To note, IFNγ staining represented a slight median fluorescence intensity (MFI) shift and a high proportion of cells with fluorescence greater than FMO but this was likely not biologically real as no NA cells and only a small subset of A cells should express IFNγ. Therefore, gating for IFNγ was more stringently set to define true positivity. This gating was performed on one replicate of D1 control T-cells incubated in their normal media and applied across all NA or A samples respectively. All FMOs were generated using D1 control T-cells.


a
FMO-LD
50:50 LD control
Count
Live Cells
18.6%
livedead-nearIR
FMO-CD8
FMO-CD4
FSC-A
CD8a-BV510
CD4-FITC
Donor 1 Example
SSC-A
Cells
79.8%
FSC-A
SSC-H
Single Cells
97.7%
SSC-A
Count
Live Cells
37.9%
livedead-nearIR
CD8a-BV510
CD8+
36.8%
dPos
2.01%
dNeg
5.01%
CD4+
56.2%
CD4-FITC
FMO (CD8+)
CD8+
FMO (CD4+)
CD4+
CD25
FSC-A
CD25+
0.43%
CD25+
0.34%
CD25+
0%
CD25+
0.17%
CD25-BV711
CD69
FSC-A
CD69+
0.94%
CD69+
2.94%
CD69+
0.14%
CD69+
1.05%
CD69-AF647
PD-1
FSC-A
PD-1+
0.79%
PD-1+
4.71%
PD-1+
0.35%
PD-1+
5.84%
PD-1-BV421
IFNγ
FSC-A
IFNγ+
0%
IFNγ+
0.064%
IFNγ+
0%
IFNγ+
0.065%
IFNγ-PE

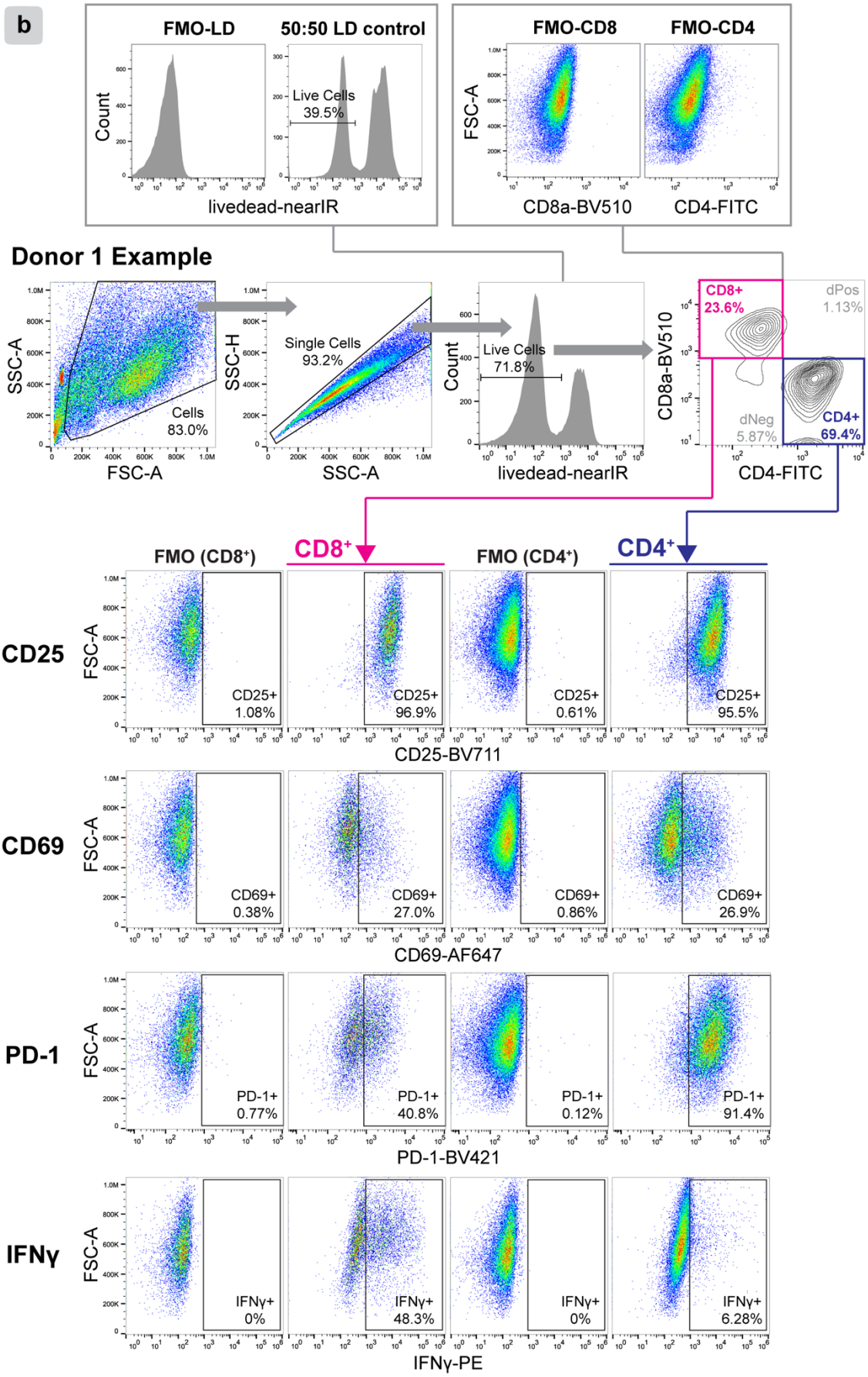

b
FMO-LD
50:50 LD control
Count
Live Cells
39.5%
livedead-nearIR
FMO-CD8
FMO-CD4
FSC-A
CD8a-BV510
CD4-FITC
Donor 1 Example
SSC-A
Cells
83.0%
FSC-A
SSC-H
Single Cells
93.2%
SSC-A
Count
Live Cells
71.8%
livedead-nearIR
CD8a-BV510
CD8+
23.6%
dPos
1.13%
dNeg
5.87%
CD4+
69.4%
CD4-FITC
FMO (CD8+)
CD8+
FMO (CD4+)
CD4+
CD25
FSC-A
CD25+
1.08%
CD25+
96.9%
CD25+
0.61%
CD25+
95.5%
CD25-BV711
CD69
FSC-A
CD69+
0.38%
CD69+
27.0%
CD69+
0.86%
CD69+
26.9%
CD69-AF647
PD-1
FSC-A
PD-1+
0.77%
PD-1+
40.8%
PD-1+
0.12%
PD-1+
91.4%
PD-1-BV421
IFNγ
FSC-A
IFNγ+
0%
IFNγ+
48.3%
IFNγ+
0%
IFNγ+
6.28%
IFNγ-PE

Figure 5.5 | Flow cytometry gating strategies to identify populations of live CD4+ and CD8+ T-cells that express a variety of activation markers. Gating for both NA **(A)** and A **(B)** cells followed the same structure but with slightly different thresholds due to size and fluorescence differences. This figure depicts Donor 1 (D1) control T-cells incubated in their normal media as an example and with respect to a single experiment. This gating was applied identically to all samples with respect to NA or A cells. FMOs were conducted on control D1 T-cells. *[NA=non-activated, A=activated, LD=live-dead]*

Both %-positive (p-) populations and MFI (m-) within positive populations were used as metrics to quantify not only the proportion of cells with expression but the intensity of expression per cell respectively. These were produced for each activation marker within $CD4^+$ and $CD8^+$ subsets, generating 20 metrics alongside %-viability and CD4:CD8 ratio. Some MFI values were highly spurious and unreliable due to few events within positive gates so filtering was performed to set MFI to 'missing' when its respective %-positive population metric was <1%.

Due to this dataset's high dimensionality, PCA was performed on all samples to observe immunogenic landscape in PCA space of the first two PCs capturing the largest sources of variation in this dataset (Figure 5.6A). Samples clearly clustered by cell state (shape) as expected with strong phenotypes observed previously upon robust Dynabead activation (Figure 5.2). Interestingly, samples also tightly clustered by donor, indicating substantial batch differences. Furthermore, contributions of each metric to the PCs (loadings) were extracted and the top 5 for each PC were plotted as arrows, with direction and length indicating which metrics most strongly drove separation along each component. Separation across PC1 was largely driven by most %-positive activation metrics (CD25, PD-1, CD69) within either $CD4^+$ or $CD8^+$ subsets, an expected factor explaining the separation of NA and A samples and the largest obvious source of variation in this dataset. Intriguingly, PC2 separation was mostly driven by $CD4^+$ and $CD8^+$ enrichment in opposite directions, suggesting baseline CD4:CD8 ratios differed between donors and potentially across different samples of varied activation state. Whilst not demonstrated in the PCA plot, PC3 separation was mostly driven by %-IFNγ parameters suggesting the added utility of IFNγ in this flow panel for providing a more informative activation phenotype.

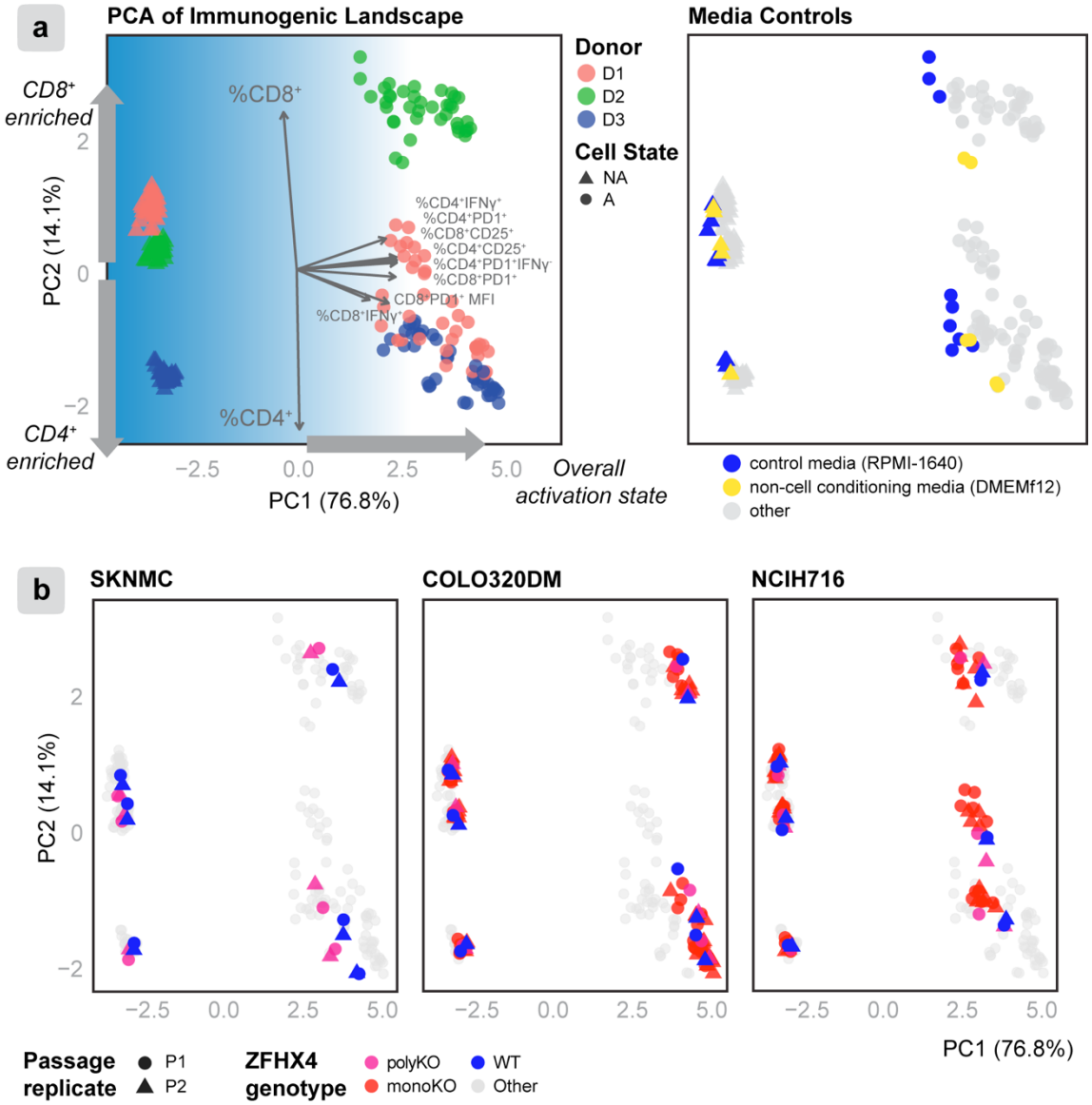


Figure 5.6 | PCA of the flow cytometry dataset spanning samples from different donors, cell activation states and conditioned media conditions. (A) PCA of all samples, with arrow length and direction indicating the top 5 contributing metrics to PC1 and PC2. 6 metrics were removed from this analysis due to presence of missing values as a result of unreliable MFI. The two types of media controls were also overlaid, representing technical and donor replicates (RPMI-1640 n=9, DMEMf12 n=6) for each cell state. **(B)** Experimental conditioned media samples were overlaid on the same PCA plot, spanning replicates of different passages and also donors (n=6 per genotype) for each cell state. *[P1=passage 1, 10m cells, P2=passage 2, current-3m cells]*

Overlaying the two types of controls in this experiment, donor T-cells incubated in their own media (RPMI-1640) and non-cell CM (DMEMf12) were investigated to identify potential donor-to-donor differences and baseline NA and A states. PCA overlay again demonstrated separation by donor but general clustering of the two control types within the same donor, suggesting differences attributed to RPMI-1640 and DMEMf12 medium were negligible.

In experimental samples, donor differences outweighed *ZFHX4* genotype effects suggesting that downstream WT-KO analyses should control for donor variability (Figure 5.6B). Within the same donor, samples irrespective of derived cell line appeared to generally cluster by *ZFHX4* genotype (colour) rather than CM passage replicate (shape) suggesting the differences between 10m and current-3m passage replicates were negligible and thus reduce the confounding of cell age with prolonged culture in this analysis.

As a further quality control check, all metrics were compared in NA and A cells incubated in both control media types to observe the degree of activation across all three donors. As both control media types clustered together on the PCA plot in Figure 5.6, they were grouped together as a single control. Activated T-cells showed 1.5-1.8-fold higher viability than resting NA cells from the same donor (D1 p.adj=$4.4x10^{-8}$, D2 p.adj=$1.97x10^{-4}$, D3 p.adj=$1.97x10^{-4}$) (Figure 5.7A). Furthermore, Dynabead activation appeared to adjust CD4:CD8 compartment proportions which was different across donors. For example, CD4:CD8 ratios in A cells exhibited significant skewing towards $CD4^+$ populations in D1 (p.adj=$8.76x10^{-5}$) but significant skewing towards $CD8^+$ populations in D2 (p.adj=$9.43x10^{-4}$), with no differences in ratio observed in D3 controls (p.adj=0.088). Across donors, A cells skewed toward $CD4^+$ in D1 (mean 2.81) and D3 (mean 4.68) but toward $CD8^+$ in D2 (mean 0.94); all differences were significant (adj.p<0.0001; testing not presented in Figure 5.7). This highlights donor-specific heterogeneity in T-cell responses to Dynabeads and the confounding impact of CD4/CD8 differences when comparing activation across donors. It also further supports the separation of metrics into both CD4 and CD8 compartments.

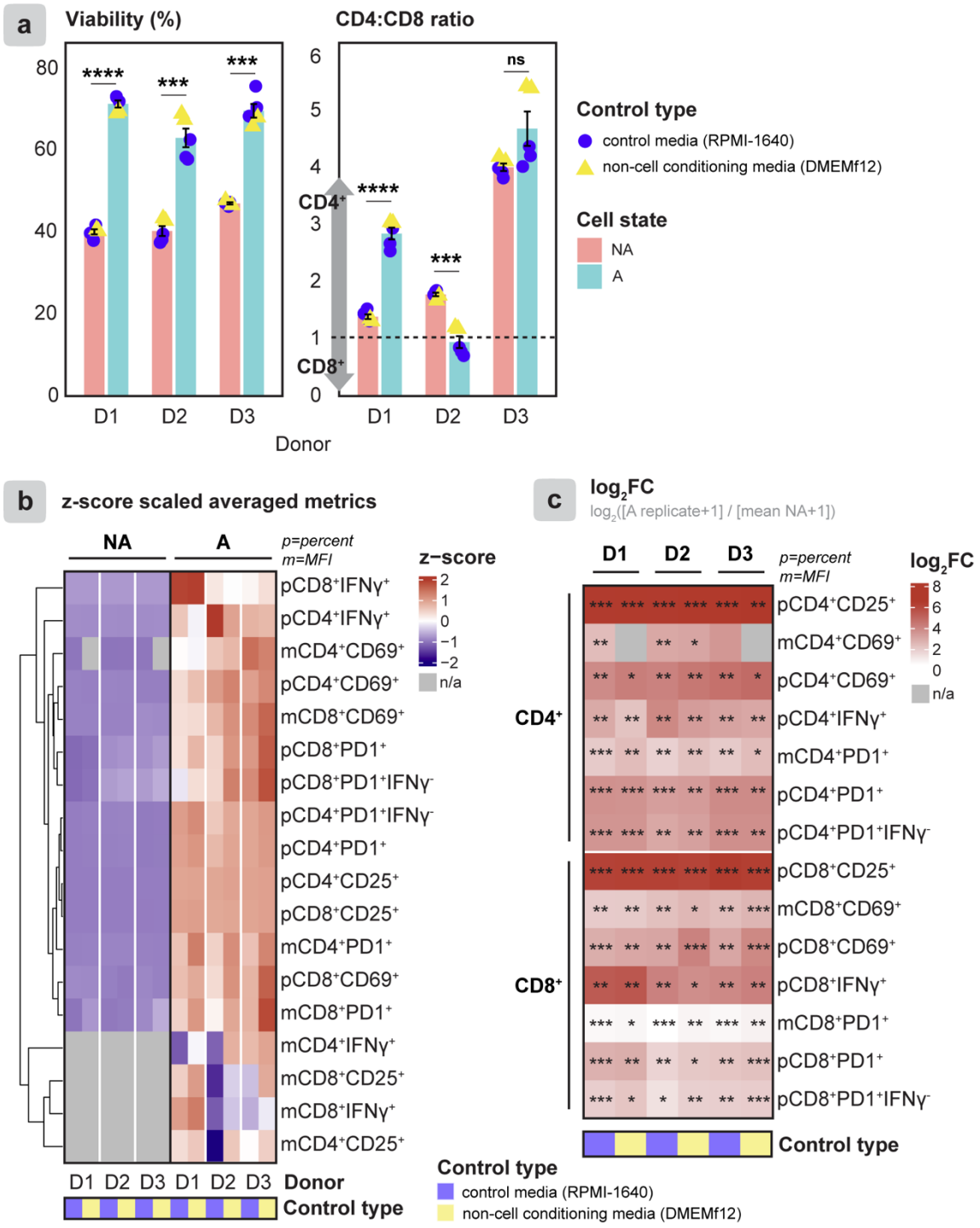


Figure 5.7 | Viability, CD4:CD8 compartment and activation metric differences between NA and A cells cultured in control media across all donors. (A) Bar plots of %-viability and CD4:CD8 ratio averages of technical replicates across both media control types (n=5), with error bars displaying SEM. Following normality testing, unpaired t-tests generated p-values annotated on respective bars. **(B)** Heatmap of z-scored of technical averages (RPMI-1640 n=3, DMEMf12 n=2). Some activation metrics were removed due to too many missing values (i.e. some MFI metrics). Rows were clustered with hierarchical clustering. **(C)** Log_2FC heatmap demonstrating relative activation metric changes in A cells vs respective NA cells from matched donor and control type. Asterisks indicate significance following unpaired t-tests between raw NA and A technical replicates per control type per donor. Some activation metrics were removed due to too many missing values (i.e. some MFI metrics). *[ns=not significant, *=p<0.05, **=p<0.01, ***p<0.001, ****p<0.0001]*

Activation metrics were averaged across replicates, donors and control types, then z-scored to assess changes in percentage or MFI (Figure 5.7B). Across all donors and control types, the majority of activation metrics demonstrated z-scores>0 and thus values higher than the mean, suggesting an overall upregulation of all activation marker expression and intensity in A cells compared to NA cells. Furthermore, notable heterogeneity between donors of A cells is consistent with donor-donor differences as described previously, for example in both control types D1 A cells appear to have the highest proportion of $CD8^+IFN\gamma^+$ populations with more intense IFNγ-MFI suggesting potentially enhanced effector function in this donor.

Z-score analysis was further supported by log_2FC analysis (Figure 5.7C), where values in A cells divided values in respective donor- and control-type paired NA cells to give a relative readout. All activation markers of both percent and MFI metrics exhibited $log_2FC>0$ and thus were significantly upregulated in A cells compared to NA cells as expected, with largest fold changes observed with $\%CD25^+$ cells in both CD4 and CD8 compartments. $\%PD1^+IFN\gamma^-$ populations were used to represent an exhaustion phenotype, since PD-1 alone can also indicate activation and combining it with IFNγ helps distinguish functional states. These populations were also significantly enriched upon Dynabead activation but more exhaustion markers such as TIM3 and LAG3 would be required to completely define an exhausted phenotype.

In summary, despite donor variability, Dynabeads robustly activated T-cells, increasing both the proportion and intensity of CD25, CD69, PD-1 and IFNγ expression across $CD4^+$ and $CD8^+$ subsets.

Given donor variability in media controls, this factor was accounted for in analyses of experimental CM samples. Instead of normalising by subtraction of media control baselines which inevitably could generate negative values that would confound downstream analyses when comparing KOs to WTs, relative fold changes were calculated using respective WTs, donors and passage replicates (Figure 5.8A). These fold changes were log_2-transformed and provided n=6 replicates which were then averaged and used for statistical testing with a one-sample t-test to identify significant differences from $log_2FC=0$ (i.e. respective WT).

Performed for both NA/A samples and across all metrics, heterogeneity amongst clones was evident in their ability to alter T-cell activation through CM (Figure 5.8B). Overall, *ZFHX4*$^{-/-}$ CM slightly reduced viability but in A cells (already maximally activated by Dynabeads) it significantly increased both the frequency and intensity of IFNγ expression in $CD4^+$ and $CD8^+$ T-cells. Furthermore, particularly in COLO320DM-KOs there appeared broad upregulation of activation metrics such as $CD25^+$ MFI, PD-1^+ MFI and %CD69 in both $CD4^+$ and $CD8^+$ T-cells, however most of these appeared significantly downregulated in NCIH716-KOs.

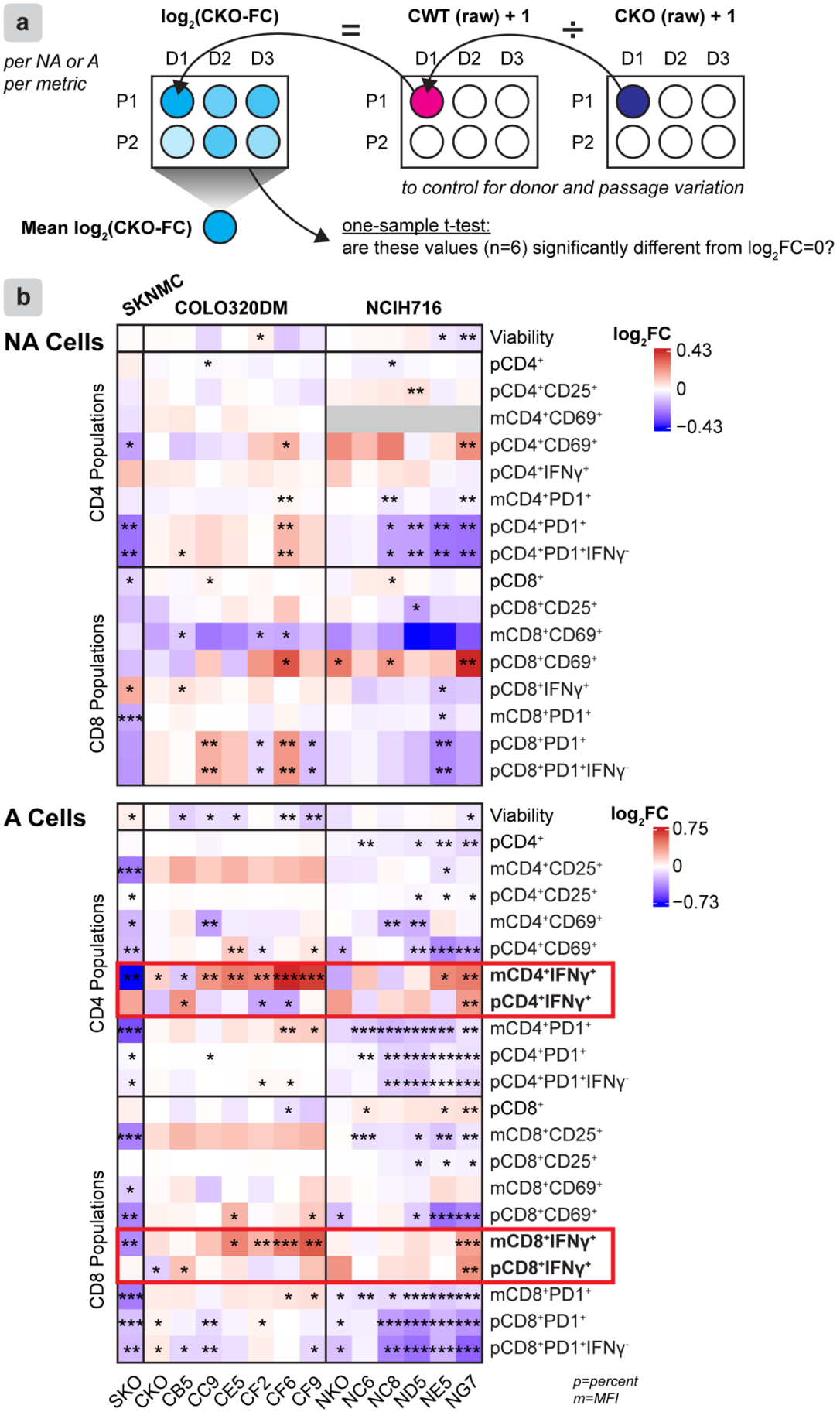

a
per NA or A
per metric
log2(CKO-FC)
=
CWT (raw) + 1
÷
CKO (raw) + 1
D1 D2 D3
P1
P2
to control for donor and passage variation
Mean log2(CKO-FC)
one-sample t-test:
are these values (n=6) significantly different from log2FC=0?
b
SKNMC
COLO320DM
NCIH716
NA Cells
A Cells
CD4 Populations
CD8 Populations
Viability
pCD4+
pCD4+CD25+
mCD4+CD69+
pCD4+CD69+
pCD4+IFNγ+
mCD4+PD1+
pCD4+PD1+
pCD4+PD1+IFNγ-
pCD8+
pCD8+CD25+
mCD8+CD69+
pCD8+CD69+
pCD8+IFNγ+
mCD8+PD1+
pCD8+PD1+
pCD8+PD1+IFNγ-
mCD4+IFNγ+
pCD4+IFNγ+
mCD8+IFNγ+
pCD8+IFNγ+
log2FC
0.43
0
−0.43
0.75
−0.73
SKO CKO CB5 CC9 CE5 CF2 CF6 CF9 NKO NC6 NC8 ND5 NE5 NG7
p=percent
m=MFI

Figure 5.8 | T-cell activation changes in *ZFHX4*$^{-/-}$ CM relative to *ZFHX4-WT* CM. (A) Schematic demonstrating the calculation of log_2FC to account for donor and passage variation and allow for statistical testing using CKO as an example. Each circle represents a single .fcs file spanning 20 activation metrics. **(B)** Mean log_2FC heatmaps demonstrating relative activation metric changes in *ZFHX4*$^{-/-}$ CM vs respective *ZFHX4-WT* CM. Asterisks indicate significance following one-sample t-tests between log_2FC values and $log_2FC=0$. Some activation metrics were removed due to too many missing values (i.e. some MFI metrics). *[ns=not significant, *=p<0.05, **=p<0.01, ***p<0.001, ****p<0.0001, FC=fold-change, P1=passage 1, D1=donor 1, NA=non-activated, A=activated]*

To highlight the differences observed with IFNγ, mean log_2FC was also plotted as bars for better visualisation of differences in percent expression and MFI (Figure 5.9A). Replicating the heatmap, >50% KO-CM induced IFNγ expression and to a higher degree in $CD4^+$ and $CD8^+$ cells relative to WT-CM. Selecting the most different KOs from COLO320DM and NCIH716 lines, CF9 and NG7 density plots alongside their respective WT of A cells from D1 in P1 media (Figure 5.9B) demonstrated subtle differences by eye that were emphasised with quantification provided by flow gates (Figure 5.9C). This highlights the value of flow cytometry in detecting subtle differences and of using analytical approaches such as mean log_2FC to interpret high-dimensional datasets. Additionally, CF9 and NG7 also demonstrated significant depletion of $PD1^+IFNγ^-$ populations particularly within $CD8^+$ compartments, indicating a reduction in exhaustive phenotype comparative to WT and a more functional state (Figure 5.8B).

These results suggest that the CM may contain an immunogenic factor or reduced levels of an anti-immunogenic factor not captured in the cytokine arrays. With robust replicates across three independent donors, these findings highlight a potential immunogenic role for *ZFHX4*$^{-/-}$ CM.

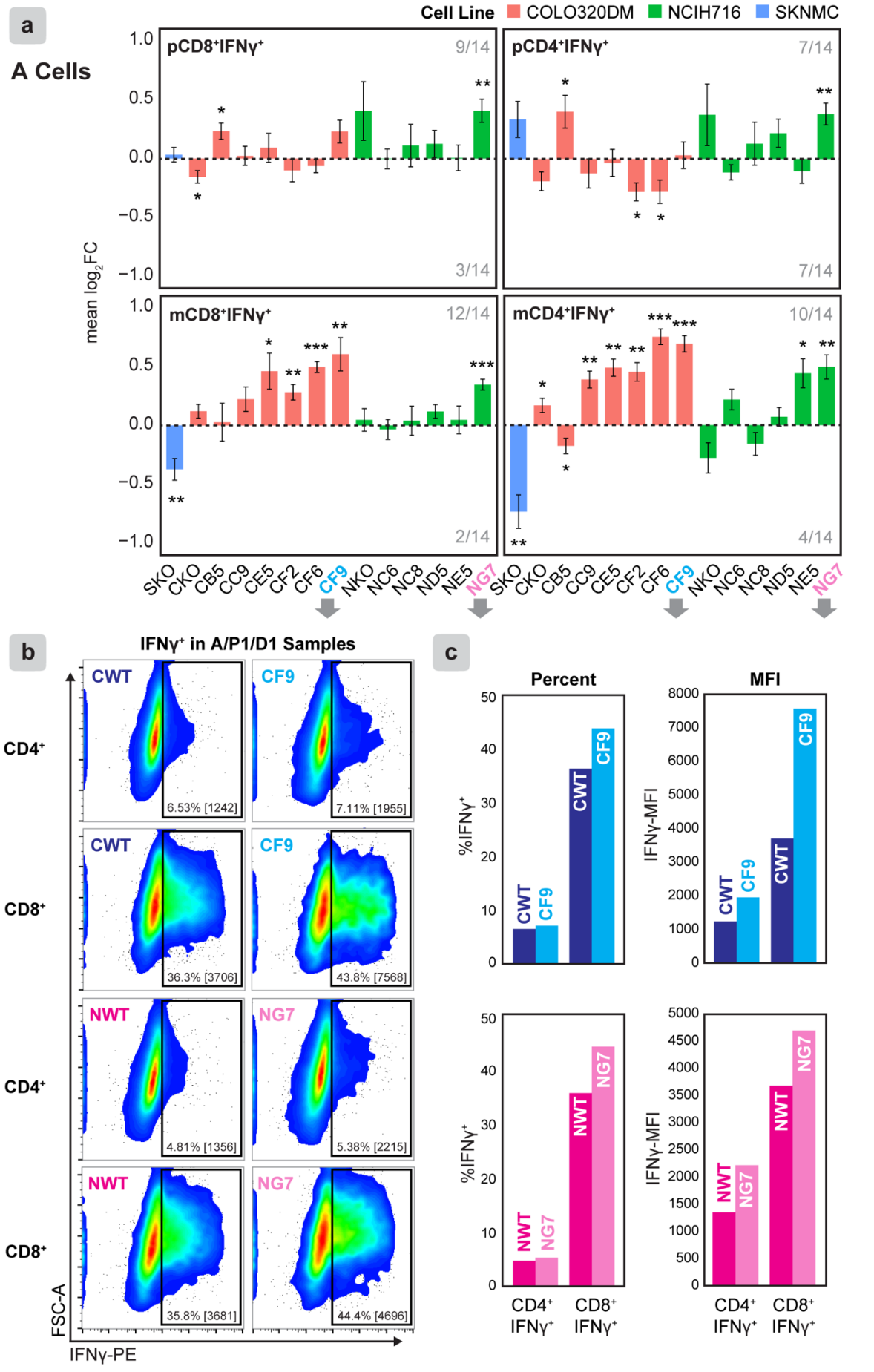
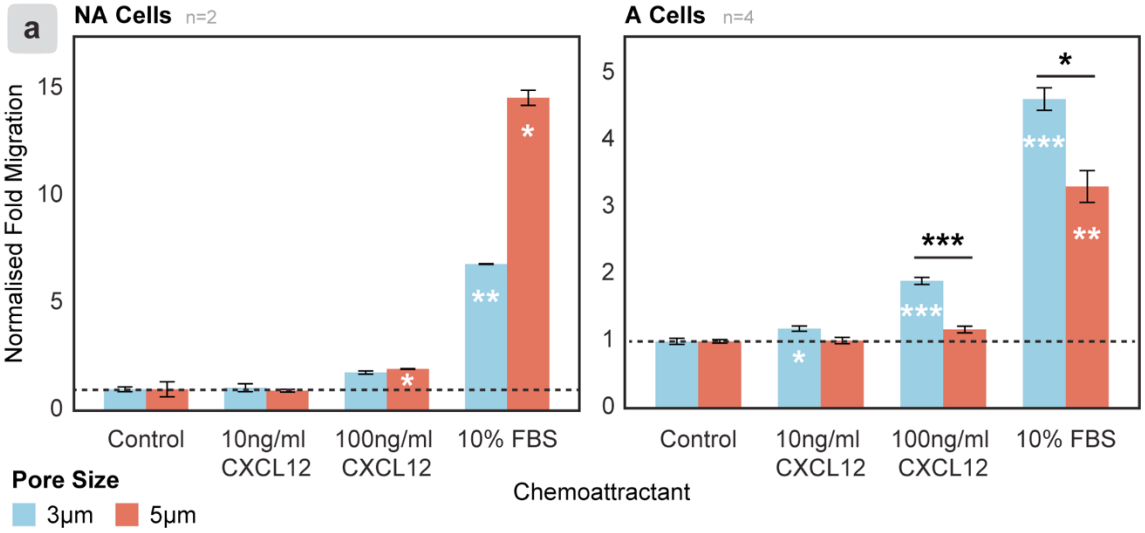
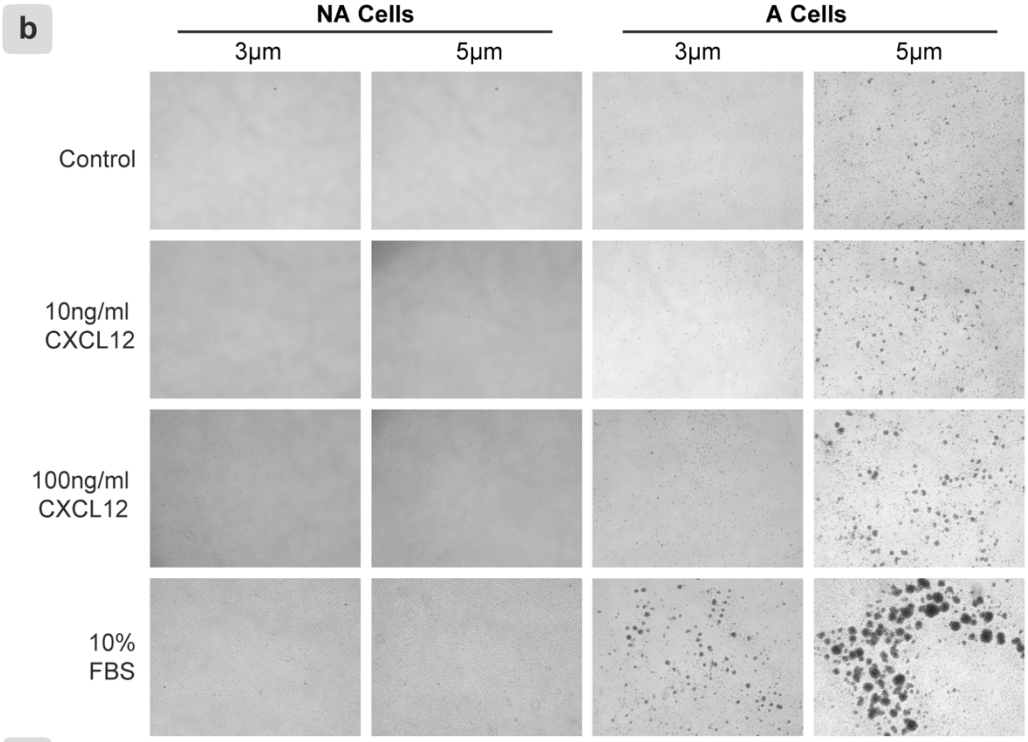

a
A Cells
Cell Line COLO320DM NCIH716 SKNMC
pCD8⁺IFNγ⁺ 9/14
pCD4⁺IFNγ⁺ 7/14
mCD8⁺IFNγ⁺ 12/14
mCD4⁺IFNγ⁺ 10/14
3/14
7/14
2/14
4/14
mean log2FC
SKO CKO CB5 CC9 CE5 CF2 CF6 CF9 NKO NC6 NC8 ND5 NE5 NG7
b
IFNγ⁺ in A/P1/D1 Samples
CWT
CF9
NWT
NG7
CD4⁺
CD8⁺
6.53% [1242]
7.11% [1955]
36.3% [3706]
43.8% [7568]
4.81% [1356]
5.38% [2215]
35.8% [3681]
44.4% [4696]
FSC-A
IFNγ-PE
c
Percent
MFI
%IFNγ⁺
IFNγ-MFI
CD4⁺ IFNγ⁺
CD8⁺ IFNγ⁺

Figure 5.9 | T-cell IFNγ changes in *ZFHX4*$^{-/-}$ CM relative to *ZFHX4-WT* CM. (A) Bar plots of mean log_2FC demonstrating relative IFNγ changes in *ZFHX4*$^{-/-}$ CM vs respective *ZFHX4-WT* CM (n=7). Error bars represent SEM and asterisks indicate significance (not multiple-testing corrected) following one-sample t-testing to log_2FC=0. Fractions in grey represent the number of KOs inducing IFNγ upregulation in T-cells (log_2FC>0). **(B)** Smoothed density plots of CWT, CF9, NWT and NG7 samples pertaining to A cells from D1 and P2 media as an example. Percent positive and [MFI] values are annotated. **(C)** %IFNγ$^+$ and IFNγ-MFI metrics of samples in (B). *[D=donor, P=passage, E=experiment]*

5.4.4 Optimisation of T-cell migration Transwell assays

To investigate the ability of *ZFHX4*$^{-/-}$ CM to attract T-cells and cause migration, Transwell migration assays were used similarly to Chapter 4 but required further optimisation. CellTiter-Glo-luminescence was used to quantify the number of T-cells which migrated through the membrane and was normalised to control media not containing chemoattractant (i.e. random migration) to calculate normalised fold migration. First, pore size and chemoattractant were compared with NA or A cells from one donor following a 20hr assay using 500k T-cells in 24-well transwells (Figure 5.10A/B). Across both pore sizes and cell activation types, there was a dose-dependent migration trend observed with potent T-cell chemoattractant CXCL12, prompting up to 1.95-fold increased migration at 100ng/ml compared to control media. However interestingly, FBS provided the most potent chemoattractant for both pore sizes and cell activation types, prompting significant increased migration across conditions and up to 14.6-fold increased migration relative to control media with NA cells and 5μm pores. On average across conditions, FBS provided 4.14-fold (±2.31s.d.) further migration than 100ng/ml CXCL12 suggesting this was an optimal chemoattractant. In general, NA cells migrated more towards FBS than A cells relative to control media (3μm pore; 1.48-fold more, 5μm pore; 4.41-fold more), however considering the biological relevance of using A cells and the confirmed significant migration towards FBS in this cell type with 3μm pores (adj.p=0.0008) prompted choosing 3μm pores, 10% FBS and A cells for all future assays.


a
NA Cells n=2
A Cells n=4
Normalised Fold Migration
Control
10ng/ml CXCL12
100ng/ml CXCL12
10% FBS
Chemoattractant
Pore Size
3μm
5μm


b
NA Cells
A Cells
3μm
5μm
Control
10ng/ml CXCL12
100ng/ml CXCL12
10% FBS

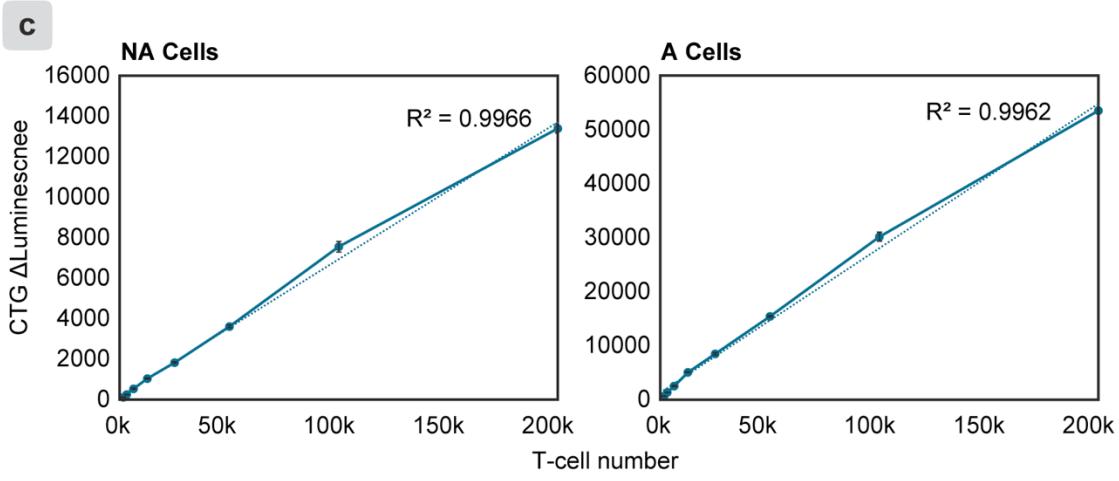

c
NA Cells
A Cells
R² = 0.9966
R² = 0.9962
CTG ΔLuminescnee
T-cell number

Figure 5.10 | T-cell migration assay optimisation using 24-well transwells. (A) Normalised fold migration calculated from background-subtracted cell-titre-glo luminescence values and division by control media containing no chemoattractant. Each bar represents mean from technical replicates (NA n=2, A n=4) and error bars indicate SEM. Following normality testing, unpaired t-tests were performed within grouped bars (black asterisks) and one-sample t-tests were performed comparing each mean to control (i.e. 1), designated by white asterisks. Non-significant comparisons are not shown. **(B)** Microscope images of replicates used in (A) following Transwell insert removal before extracting for CellTiter-Glo. All are at 4X magnification. **(C)** From a different donor, scatter plots indicate the linear relationship between T-cell number and CellTiter-Glo luminescence. Error bars are SEM and all dots are mean (n=3). A linear trendline is presented with the dashed line, also displaying the R^2 value.

As a quality-control check, microscope images were taken to observe potential differences following insert removal before performing CellTiter-Glo (Figure 5.10B). Generally, images on the 24-well plate were not representative as T-cells tended to cluster in specific areas, supporting the utilisation of using CellTiter-Glo as a more sensitive and reliable quantification of cell number. Furthermore, following serial dilution using T-cells from another donor, an expected strong linear trend was observed between known T-cell number and CellTiter-Glo luminescence values (R^2 values > 0.9960 for both NA and A linear trendlines) (Figure 5.10C). Interestingly, on average luminescence per cell was 4.65-fold higher for A cells (mean 0.36U/cell) compared to NA cells (mean 0.078U/cell) likely due to enhanced cell size and metabolism. This suggests luminescence values between NA and A cells shouldn't be compared without normalisation to respective controls.

5.4.5 Effects of *ZFHX4*$^{-/-}$ conditioned medium on T-cell migration

To use this assay for testing *ZFHX4*$^{-/-}$ CM specifically, two independent experiments were performed and summarised in Table 5.2. Both employed A cells from different donors and comprised chemoattractant-free and chemoattractant controls. The readout of these assays followed the same protocol for suspension Transwell assays, obtaining technical replicates of CellTitre-Glo luminescence values to quantify migrated cell number.

Table 5.2 | Experimental setup of two T-cell migration Transwell-based experiments

ID	Plate Format	D (n)	Cell line passage	Conditioning (i.e. BC)	TC	Assay duration
23325	24w	1	7m	• 24hr • DMEMf12 10%FBS/1%PS, • Diluted 2:1 with fresh DMEMf12 10%FBS/1%PS	RPMI-1640 0.5%FBS 1%PS (150k cells)	42hr
24625	96w	3	10m and current-3m	• 72hr • DMEMf12 5%FBS • Diluted 1:1 with fresh RPMI-1640 0%FBS	RPMI-1640 0%FBS (75k cells)	72hr

[D=donor, BC=bottom chamber, TC=top chamber]

As before with the main flow cytometry experiment, relevant controls were analysed first to investigate assay success and potential donor variability (Figure 5.11A). Similar to previous work on activation state, it was evident that T-cell migration varied between donors, for example normalised fold migration to 10% FBS with respect to chemoattractant-free controls was 4.04-fold higher in D2 than D1 T-cells. All migration between donors was statistically significantly different as by one-way Anova testing ($p<8.05x10^{-16}$). This underscores the need to use multiple donors as biological replicates and to normalise analyses within donors for valid comparisons. Additionally, T-cell migration across all donors demonstrated a dose-response to FBS concentrations as well as limited migration towards 100ng/ml CXCL12 as a comparison. Migration between donors ($p=7.61x10^{-10}$) and chemoattractant types ($p<2.2x10^{-16}$) were significantly different following two-way Anova testing respectively. This again reiterates the potency of FBS as a chemoattractant for healthy donor T-cells and assay success in both main experiments.

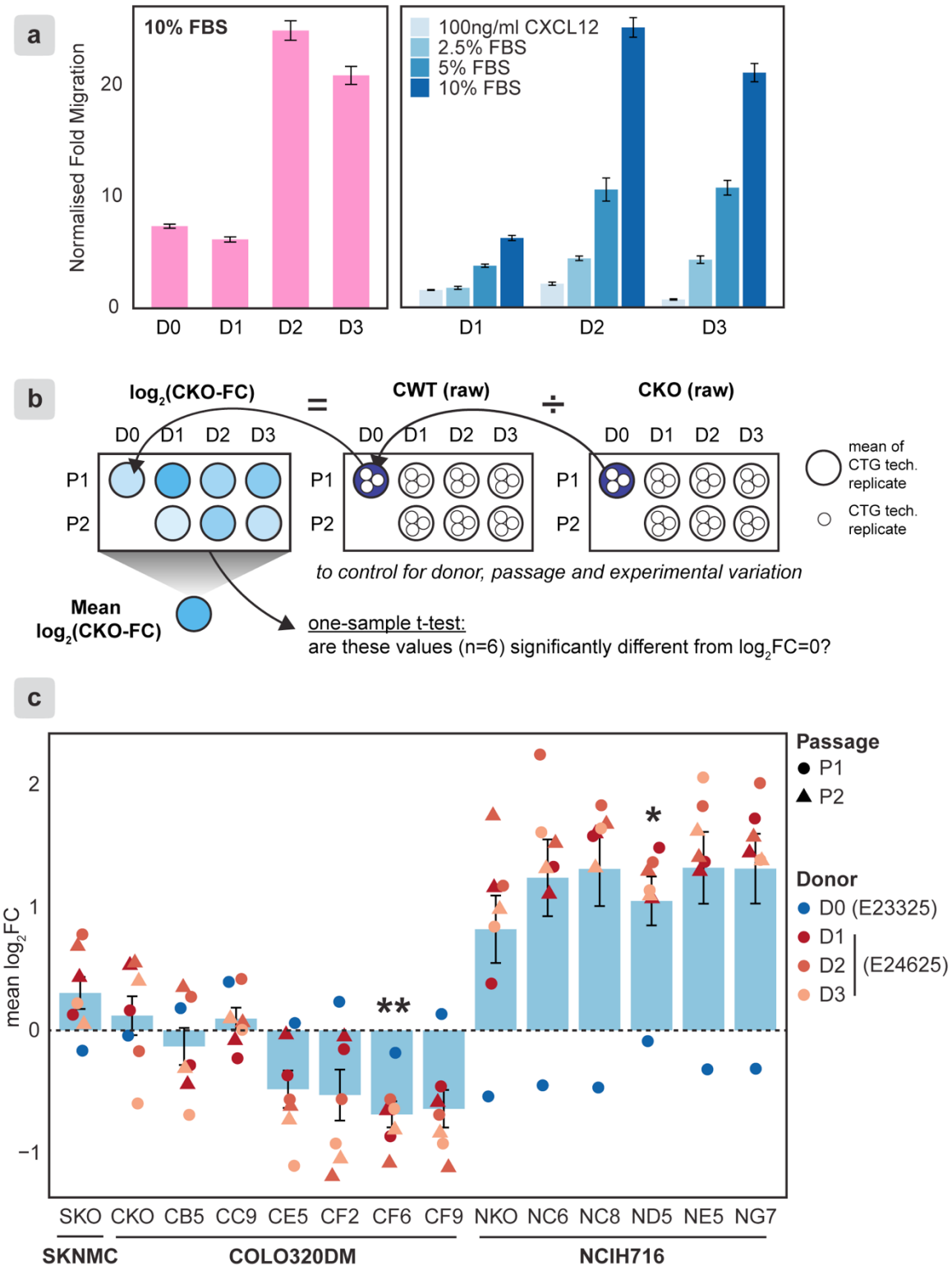


Figure 5.11 | Migration of activated healthy donor T-cells towards controls and *ZFHX4*$^{-/-}$ conditioned medium. (A) Bar plots of mean normalised fold migration calculated by dividing individual technical Transwell (n=2) and CellTiter-Glo (n=3) replicates by chemoattractant-free mean of the same experiment and donor (n=6). Error bars represent SEM. **(B)** Schematic demonstrating the calculation of log_2FC to account for experiment, donor and passage variation and allow for statistical testing using CKO as an example. **(C)** Bar plots of mean log_2FC demonstrating relative fold migration towards *ZFHX4*$^{-/-}$ CM vs respective *ZFHX4-WT* CM (n=7). Error bars represent SEM and asterisks indicate significance with BH correction following one-sample t-testing to log_2FC=0. *[D=donor, P=passage, E=experiment]*

Following a similar procedure to before, log_2FCs of KO vs WT samples were calculated using mean CellTiter-Glo technical replicates (n=3) and averaged to produce a single mean log_2FC value per sample within the same donor, passage and experiment (n=7) (Figure 5.11B).

Compared to relative WTs, most COLO320DM *ZFHX4*$^{-/-}$ CMs led to decreased fold migration, with only CF6 significantly lower than WT by one-sample t-testing (log_2FC=0; adj.p=0.00928) (Figure 5.11C). Conversely, most NCIH716 *ZFHX4*$^{-/-}$ CMs prompted enhanced fold migration, with only ND5 significantly higher than WT (adj.p=0.0256). This cell line and clonal dependent heterogeneity is not surprising considering observed heterogeneity in other assays throughout this project. The broad range of log_2FC values per condition highlights experimental noise despite normalisation, indicating additional repeats or alternative assays are needed to clarify any link between *ZFHX4*$^{-/-}$ and T-cell migration. Additionally, results when using CXCL12 suggest that even a potent chemoattractant is not sensitive enough to induce significant migration in this type of assay suggesting it would take considerable protein types and amounts within CM to induce a noticeable change outside of experimental noise. A further repeat could include a CM-concentrating protocol to support this.

5.5 Discussion

This Chapter was designed to provide *in vitro* validation of ZFHX4-related effects on immunogenicity. Using newly established and optimised T-cell activation and migration assays, I demonstrated that conditioned medium from *ZFHX4*$^{-/-}$ cell lines had variable effects on migration but consistently enhanced IFNγ expression across three independent donors, indicative of increased T-cell effector function. This provides *in vitro* evidence that genetic aberrations in *ZFHX4* have the potential to effectuate heightened immunogenicity.

There are two main limitations with my approach which will be explored below. Firstly, *ZFHX4*$^{-/-}$ models do not represent clinically relevant ZFHX4 missense mutations, the phenotype associated with enhanced immunogenicity in the TMBL setting as set out in Chapter Three. Secondly, methods used in this chapter to empirically examine immunogenicity are reductionist.

5.5.1 *ZFHX4*$^{-/-}$ doesn't model clinically relevant missense mutations

Many strengths and limitations of my *ZFHX4*$^{-/-}$ models are already outlined in Chapter Four which were largely suited to the study of ZFHX4 in tumourigenicity, considering they offered a dynamic expression axis that (in some models) nicely replicated the dynamic expression axis clinically, with *ZFHX4*^{high} contexts conferring enhanced EMT-related biology and poor prognosis. Despite not representing clinically relevant missense mutations, these models were also employed for exploring immunogenicity as a compromise, considering (a) modelling missense mutations is technically more challenging than performing complete KO editing, (b) many models would require generation due to the lack of one particularly strong hotspot and (c) complete KO editing allowed parallel exploration of the tumourigenicity theme. The observed IFNγ effect is biologically meaningful considering the robust experimental controls in my experiments, however justifiable interpretations from these models on immunogenicity are considerably limited and further work would ultimately require the generation of ZFHX4-missense mutated models. As discussed in Chapter Four,

this could be mediated through HDR-mediated CRISPR, or base/prime editing approaches, either in cell lines and/or organoid models.

Under the premise of exploring missense variants of uncertain significance (VUS) in cancer, one study using colon cell lines (HT-29, LS411N) and organoids was able to generate clinically relevant JAK1 missense mutations through base editing mutagenesis screens (Coelho et al., 2023). Investigators could then segregate these into defined LoF or GoF variant types based on functional validation of their sensitivity or resistance to IFNγ, an important determinant of immunotherapy response. Autologous T-cell organoid models and consequent killing assays provided special novelty with *in vitro* demonstration of mutational functional impact. The authors also mapped these mutations and residues to those identified in clinically annotated datasets, including immunotherapy-treated cohorts, providing further evidence that they may modulate immunogenicity in the clinical setting. Prime editing screens have also shed light on other VUS missense mutations, with studies finding some of these mutations were functional through modulating cell fitness in breast cancer cell line models (Ren et al., 2023). Furthermore, offering models more representative of colon biology and a potentially more permissive chromatin state which is likely more preferable for ZFHX4 function, colon-organoids have been successfully used to validate clinically relevant missense mutations in genes such as FBXW7 by base editing (Boretto et al., 2024, Geurts et al., 2023) and TP53 by prime editing (Geurts et al., 2021). Studies such as these emphasise the potential of all VUS-classed ZFHX4 missense mutations for immuno-oncology research. It would be extremely exciting to replicate these approaches in a colon, ZFHX4 and immunogenicity context using both cell line and organoid models.

5.5.2 Healthy donor T-cells are a minimalist approach toward immunogenicity

Whilst T-cell activation and migration assays utilised in this chapter were a compromise, considering the lack of potential methods available to explore immunogenicity *in vitro* with the models I had available, they were inherently reductionist in their approach to empirically examine immunogenicity. In addition to other model opportunities discussed below, it's also important to note other metrics reflecting immunogenicity could have been explored using the same healthy donor T-cells. Other studies have used CM to investigate effects on T-cell cell cycle and proliferative capacity (Jin et al., 2010), apoptosis and additional activation metrics such as GZMB, IL-2 and TNFα (Zhang et al., 2021, Cheng et al., 2023b) and transcriptional changes through RNAseq (Lacher et al., 2024). Furthermore, although this project's focus was on antigen-independent immunogenicity, immunopeptidomics on *ZFHX4*$^{-/-}$ models may have highlighted interesting differences in antigen processing and presentation, considering many immunoproteasome genes such as PSMB9 were among the top upregulated in clones like CC9 by RNAseq (data not shown).

Direct co-culture models offer a more promising alternative to indirect CM-based immunogenicity models, where epithelial and immune compartments grow in close proximity. Table 5.3 exemplifies the ever-expanding arsenal of epithelial-immune co-culture models to investigate immunogenicity, which largely regard T-cells as the main effectors. Whilst studies have performed co-culture of colon cancer cell lines with healthy PBMCs (Acha-Sagredo et al., 2025, Courau et al., 2019), the majority of immuno-oncology research focuses on patient-derived organoid models cultured with either peripheral mononuclear blood cells (PBMCs) or tumour-infiltrating lymphocytes (TILs) from the same (autologous) or different (allogeneic) donors. Furthermore, other flavours of epithelial models such as organ-on-a-chip (Lorenzo-Martin et al., 2025) and gel-liquid interface (Li et al., 2025b) provide a degree of novelty and reiterate the versatility of colon epithelial-immune co-culture models.

Table 5.3 | A brief summary of epithelial-T-cell co-cultures spanning mostly colon contexts

Study	Model Context		Epithelial Compartment	T-cell Compartment	
	Tissue	Setting		Origin	Donor Type
Acha-Sagredo[1]	Colon	Tumour	Cell line (HCT116, Caco2)	PBMC	Allogeneic
Courau[2]	Colon	Tumour	Cell line spheroids (HT-29, DLD1)	PBMC	Allogeneic
Zhou[3]	Bile duct	Tumour	PDO	PBMC	Allogeneic
Madrova[4]	Colon	Tumour	PDO	PBMC	Autologous
Li[5]	Colon	Tumour	PDO	PBMC + TIL	Autologous
Schreurs[6]	Colon	Healthy	PDO	TIL	Autologous
Kong[7]	Rectum	Tumour	PDO	TIL	Autologous
Recaldin[8]	Colon	Healthy	PDO	TIL	Autologous
Cattaneo[9]	Colon	Tumour	PDO	TIL	Autologous
Harter[10]	Colon	Healthy and Tumour	PDO	PBMC	Allogeneic + Autologous
Podaza[11]	Lung	Tumour	PDO	TIL	Autologous
Lorenzo-Martin[12]	Colon	Healthy and Tumour	Mini-colon (organ on a chip)	TIL	Autologous
Dijkstra[13]	Colon + Lung	Tumour	PDO	PBMC	Autologous
Voabil[14,15]	Pan-tissue	Tumour	Tumour Fragment Cultures	TIL	Autologous
Li[16]	Colon	Tumour	PDO	PBMC	Allogeneic
Li[17]	Lung	Tumour	Gel-liquid interface PDO	PBMC	Allogeneic

[1](Acha-Sagredo et al., 2025), [2](Courau et al., 2019), [3](Zhou et al., 2022), [4](Madrova et al., 2025), [5](Li et al., 2024), [6](Schreurs et al., 2021), [7](Kong et al., 2018), [8](Recaldin et al., 2024), [9](Cattaneo et al., 2020), [10](Harter et al., 2024), [11](Podaza et al., 2023), [12](Lorenzo-Martin et al., 2025), [13](Dijkstra et al., 2018), [14](Voabil et al., 2021), [15](Kaptein et al., 2025), [16](Li et al., 2025a), [17](Li et al., 2025b) [PDO=patient-derived organoid, PBMC=peripheral blood mononuclear cell, TIL=tumour-infiltrating lymphocyte]

The allogeneic setting offers easier T-cell sourcing and standardisation and are useful for approximate immunogenicity readouts such as IFNγ responsiveness (Magre et al., 2023). However, there is significant risk of alloreactivity where HLA mismatched immune activation can dominate over tumour-specific responses (Harter et al., 2024). Conversely, autologous T-cells from the same epithelial donor preserve an exact HLA match, antigen history and offer more representative models of the *in vivo* situation (Magre et al., 2023). However, in

practice this is logistically harder due to variable quality, limited yields and requirement for same-day processing. Other considerations with co-culture models involve T-cell origin. Methods extracting immune cells such as TILs from colon epithelium are well established but TILs are scarce and expansion *in vitro* is unpredictable. However, being tumour-reactive, *in vitro* expansion can permit assays such as antigen-dependent T-cell killing (Kromann et al., 2024). In contrast, PBMCs are readily obtainable in higher quantities with full lymphoid diversity, allowing hypothesis-free phenotyping and screening soluble cues, though may present with poor tumour specificity. Nevertheless, methods have been introduced to pulse PBMCs with tumour fragments and enrich tumour-reactive clones *in vitro* (Cattaneo et al., 2020, Dijkstra et al., 2018). More considerations accompany media choices and basement membrane extract formulations, both of which have the potential to hinder both epithelial and T-cell compartments. These models primarily serve to assess tumour-cell cytotoxicity but can enable mechanistic studies of immunogenicity drivers such as ZFHX4 using genome editing. One study using mouse colonic organoids of various mutational backgrounds performed co-culture with colonic fibroblasts and macrophages and following scRNAseq, demonstrated multi-layered regulatory interactions that drove distinct transcriptional changes in all compartments (Qin et al., 2023). This illustrates not only the cell-type flexibility in these co-culture systems but also their power to resolve genotype-dependent effects across the epithelial and stromal/immune interface.

One important study developed MMRp and MMRd colon cancer cells expressing equal levels of the same given antigen, enabling cognate recognition by engineered T-cells (Cattaneo et al., 2025). Here investigators demonstrated MMRp cells were intrinsically resistant to T-cell mediated killing comparative to MMRd cells suggesting this was independent of neoantigenic landscape and the MMRp-derived secretome can also modulate effective tumour recognition and killing. It would be plausible to investigate the impact of missense ZFHX4 mutations in models such as this to control for antigen-dependent interactions. This work also emphasises that immunogenicity cannot solely be

explained by mutational load. Collectively, co-culture models would offer a more biologically-relevant tool to study ZFHX4 in the context of immunogenicity and further work should prioritise using them to investigate the functional impact of missense mutations.

Another alternative to indirect CM-based models as well as *in vitro* models generally is the application of *in vivo* mouse models. These are considered gold-standard in the field of immuno-oncology for investigating immunogenicity at the most holistic level, and most studies in Table 5.3 use them for a final validation of phenotype. This is because immunocompetent mice provide a functionally intact immune system and enable assessment of interactions with immunotherapy and subsequent survival, enabling direct linking of “high immunogenicity” with “immunotherapy response” (Olson et al., 2018).

Three predominant immunocompetent models are routinely used for such investigations: (1) syngeneic models – tumour cell lines are derived from an immunocompetent mouse background (usually C57BL/6 or BALB/c), expanded and/or manipulated *in vitro* and inoculated into wild-type hosts of the same background, (2) genetically engineered mouse models (GEMMs) – modified at the germline level, these allow autochthonous tumour growth in the native microenvironment and (3) humanised patient-derived xenograft models – both immune cells and tumour cells obtained from the same patient are injected into immunodeficient NOD/SCID mice, allowing better recapitulation of human tumour heterogeneity (Olson et al., 2018). Many of these models have their pros and cons which are not further discussed here but show the potential of investigating casualty of genetic aberration (i.e. somatic mutational drivers) with changes to immunogenicity as inferred from response to immunotherapy. Specifically in colon cancer contexts, this has already been achieved in GBP2-KO MSS-CT26 cells of BALB/c mice which demonstrated reduced survival following immune checkpoint blockade (ICB), highlighting GBP2 important for immunogenicity (Wang et al., 2022). Furthermore, one study using MSI-MC38 cells of C57BL/6 background showed targeting IGF-1R with AG-538 (identified to positively modulate chemokine expression) led to enhanced anti-tumour responses following ICB (Li

et al., 2025a). Even transcriptomic investigation of MC38 and CT26 models post-ICB can validate biomarkers of sensitivity or resistance to ICB, with one study associating CD74 expression attributable to enhanced ICB response (Acha-Sagredo et al., 2025).

As an initial experiment, I could employ base or prime editing in MSS-CT26 cells to generate clinically relevant ZFHX4 missense mutations in the TMBL setting. These *ZFHX4*^{mt} lines alongside isogenic *ZFHX4*^{wt} controls could then be implanted back into wild-type BALB/c mice and allow the study of tumour growth, serially harvesting tumours for flow cytometry and/or scRNAseq to quantify TME composition and function. This would also permit mechanistic inferences about how ZFHX4 missense mutations might reshape immune landscape with respect to multiple cell types which may offer equal importance to functional T-cell infiltrate. In parallel, incorporating ICB with survival analysis will test if ZFHX4 missense mutations sensitise tumours to immunotherapy.

If intratumoural T-cell populations do appear important in these analyses, further interrogation would involve focusing on antigen-specific T-cell responses. The above studies describe investigating anti-tumour immunity in an antigen-agnostic manner where antigen-specific infiltrating T-cells responding to endogenous antigen are grouped with antigen-non-specific T-cell populations such as bystander T-cells. The former are attributable solely to tumour biology and offer greater relevance comparative to the latter in examining tumour-specific immune responses (Ceraolo et al., 2025). Disentangling these populations are crucial for characterising specific anti-tumour immunity, hence methods have been established to do this. Some describe labelling antigen-specific infiltrating T-cells in response to endogenous antigen (Li et al., 2022) but most describe overexpressing the ovalbumin (OVA) antigen SIINFEKL in tumour cell lines such as MC38 (Hua et al., 2023, Koksal et al., 2025). After implantation into respective C57BL/6 mice, this enables an OVA-specific T-cell response that arises endogenously and thus attributable to tumour biology. These tumour-specific T-cells can be detected using fluorescently-labelled tetramers, composed of four MHC-I molecules and SIINFEKL peptide which bind OVA-specific TCRs,

enabling more thorough interrogation of tumour-specific T-cell responses by methods such as flow cytometry (Porgador et al., 1997, Dersh et al., 2019). Readouts of anti-tumour immune response with or without ICB include infiltration, expansion and function of OVA-specific T-cells. Studies have used OVA-MC38 models to investigate the effect of radiotherapy (Koksal et al., 2025), β2M-KO (Hua et al., 2023), β-catenin inhibition (Wang et al., 2020) and EHMT2 inhibition (Sun et al., 2024). This could equally be applied to studying *ZFHX4*^{mt} in the setting of TMBL colon cancer using CT26 models as also reported (Sun et al., 2024).

Considering *in vivo* mouse models have already validated tumourigenicity associated with enhanced ZFHX4 expression in GBM (Chudnovsky et al., 2014), it's reasonable to also inspect whether specific missense ZFHX4 mutations modulate anti-tumour immunity, immunotherapy response and thus immunogenicity *in vivo* using similar models to those described above. Multiple syngeneic models may need to be employed considering they differ in intrinsic tumour-immunity (Jin et al., 2022) such as those employing more biologically-relevant orthotopic approaches comparative to subcutaneous and further mechanistic work would be required to demonstrate how altered ZFHX4 could confer anti-tumour immunity.

5.5.3 Predicted function of ZFHX4 missense mutations

As previously outlined, all missense mutations in ZFHX4 are currently classified as VUS. In Chapter Four it was thoroughly discussed how multiple links across ZFHX4 expression, NuRD complex activity, EMT phenotype and tumourigenesis could predict a mechanism where ZFHX4 associates with NuRD and through its function as a TF regulates EMT-related genes and subsequently promotes tumourigenesis. In contrast, this Chapter outlines ZFHX4 mutations are frequent even in a TMBL setting and are associated with an immunogenicity phenotype. Considering upregulated function by enhanced expression

prompts tumourigenesis, it's plausible to suggest missense mutation could instead *alter* rather than abrogate or enhance function of ZFHX4 for immunogenicity. This is supported by evidence from NuRD associations with immunogenicity and reported mutations in NuRD subunits that change but don't quantitively regulate its activity.

Association of NuRD with supressed immunogenicity

A vital study showed EHMT2 inhibition sensitised MSS colon cells to immunotherapy using both PDO-T-cell co-cultures and *in vivo* models (Sun et al., 2024). Mechanistic investigations revealed the subsequently secreted lectin galectin-7 (LGALS7) was upregulated and directly enhanced $CD8^+$ T-cell-mediated cytotoxicity which resulted in an efficient anti-tumour immune response. Interestingly, they found LGALS7 was epigenetically repressed by the functional association between EHMT2 and CHD4, a key NuRD member. Inhibition of NuRD members HDAC1 and HDAC2 through HDAC inhibitors such as SAHA and CXD101 potentiate anti-tumour immunity in lung adenocarcinoma (Xu et al., 2024), colon adenocarcinoma (Blaszczak et al., 2021) and hepatocellular carcinoma (Tu et al., 2025) in combination with ICB. The NuRD complex has also been inversely associated with complement expression and CD8+ T-cell infiltration in hepatocellular carcinoma (Shao et al., 2020).

Considering NuRD activity correlates with colon tumourigenesis (Chapter Four), its suppression or dissociation from TFs may indirectly promote immunogenicity through the loss of tumourigenesis. This is supported by the inhibition of β-catenin (Wang et al., 2020, Spranger et al., 2015) and Wnt signalling (DeVito et al., 2021) in different cancer types that can potentiate immunotherapy response *in vivo*. Furthermore, TGFβ in both epithelial and stromal compartments has been shown to drive immune evasion in colon cancer where its inhibition can enhance cytotoxic T-cell response and increase ICB sensitivity (Tauriello et al., 2018, Chakravarthy et al., 2018), and TGFβ is a known inducer of EMT (Miettinen et al.,

1994, Xu et al., 2009). Furthermore, lysine-specific demethylase 1 (LSD1) is a known regulator of EMT through SNAIL-mediated E-cadherin transcriptional repression (Lin et al., 2010), implicated in colon tumourigenesis (Antona et al., 2023, Ding et al., 2013, Huang et al., 2013) and demonstrates immunotherapy sensitisation upon its inhibition through stimulation of type-I IFN in tumour cells (Sheng et al., 2018). Co-targeting of LSD1 and TGFβ alongside ICB can also potentiate intratumoural CD8+ infiltration and cytotoxic function (Sheng et al., 2021). Importantly, LSD1 is a key NuRD subunit in breast cancer (Wang et al., 2009) and whilst a direct interaction has not been reported in the colon before, it suggests both LSD1 and NuRD in colon cancer could contribute to poor immunogenicity by similar mechanisms. Taken together, this suggests that NuRD-mediated EMT and tumourigenesis generally correspond with immunosuppression.

Mutations in NuRD subunits and interactors alter overall function

To understand how mutations in ZFHX4 could affect NuRD function, one must understand its architecture at the protein level. For robust biophysical and biochemical analyses, large multi-protein complexes must be isolated in their native form to accurately interpret structure and function (Saathoff et al., 2015). Methods to isolate and purify endogenous complexes of many subunits have been largely improved through protein-derived affinity baits by affinity chromatography directly on cell extracts and offer gentler elutions comparative to a typical IP (Saathoff et al., 2015). To interrogate the NuRD complex, most studies use small peptide baits constituting GST-fusions of the first <50 residues of FOG1 (Hong et al., 2005, Lejon et al., 2011, Saathoff et al., 2015, Sharifi Tabar et al., 2019, Low et al., 2020) or FOG2 (Roche et al., 2008) known to strongly bind RBBP4 and RBBP7 in NuRD. This has permitted substantial work in the field of NuRD in recent years such as the clarification of core subunits in each stoichiometry.

NuRD subunit expression has been demonstrated in some cell lines (NT2, PC3, MCF7, HeLa, U2OS, HEK293) but it would be powerful to not only demonstrate mRNA (not shown in Chapter Four) and protein expression in my COLO320DM and NCIH716 WT cell lines but examine NuRD complex composition and structure too (Sharifi Tabar et al., 2019). Proposed overall biological stoichiometry of mammalian NuRD from multiple biochemical analyses reveals NuRD is built from a histone deacetylase module ([2]MTA:[2]HDAC:[4]RBBP) and nucleosome remodelling module ([1]MBD:[1]GATAD2:[1]CHD) with asymmetry governed by singular GATAD2 and MBD subunits (Sharifi Tabar et al., 2019, Low et al., 2020, Arvindekar et al., 2022, Torrado et al., 2017). Paralogs of these subunits are context-dependent and provide alternative NuRD functions through “paralog switching” as discussed in Chapter Four. Furthermore, it’s also been shown subunit members such as CHD4 can lack in NuRD where it retains its HDAC activity but loses nucleosome remodelling activity, and this is known as the NuDe complex (Low et al., 2016, Arvindekar et al., 2022). Additional NuRD sub-complexes including MTA:HDAC:RBBP and MTA:HDAC:MBD have been identified in cells (Arvindekar et al., 2022). CDK2AP1 has been described as a key but potentially dispensable CHD-binding subunit (Leighton and Williams, 2020, Spruijt et al., 2010), and proteins such as PWWP2A and PWWP2B can compete for core subunits, displacing MBD and thereby altering transcriptional regulation (Zhang et al., 2018). This reiterates the complexity of NuRD and considering most of the work above is in non-cancer or non-colon contexts (particularly in murine erythroleukemia cells), there is a significant gap in the field for characterising true NuRD architecture in the normal colon and colon cancer.

Adding another level of complexity supplied by NuRD core subunits and their paralogs, further protein binders have been described as dynamic substoichiometric interactors and ‘context switches’ that likely recruit NuRD to target loci and retune targeting and activity (Smits et al., 2013). Some examples include previously described FOG1 (Hong et al., 2005), ZMYND8, ZNF592 and ZNF687 (Smits et al., 2013, Kloet et al., 2015, Spruijt et al.,

2016), ZNF219 and SLC25A5 (Sharifi Tabar et al., 2019) and SALL4, ZNF512B and ZNF532 (Kloet et al., 2015). Therefore, given many of these are ZF-TFs, there is substantial potential for ZFHX4 to also be regarded as a substoichiometric interactor that has likely not been identified due to its predicted context-dependent nature. Supporting evidence for this comes from ZFHX4-IPMS in NPCs (Perez Baca et al., 2025). Whilst the full list of interacting proteins was never explored, it revealed ZMYND8, ZEB1 and ZBTB2, all previously shown to bind NuRD. ZMYND8 interacts with GATAD2A in NuRD and functions to recruit NuRD to DSBs in a PARP-dependent manner (Spruijt et al., 2016). ZEB1 interacts with multiple NuRD subunits to regulate EMT as previously discussed (Manshouri et al., 2019, Perez-Oquendo et al., 2023). ZBTB2 can interact with ZF-TFs and bridge them with NuRD through a direct interaction with GATAD2A (Russo et al., 2021). Cumulatively, this gives strong evidence for direct or indirect links of ZFHX4 with NuRD as a substoichiometric interactor.

Binding of substoichiometric members to NuRD and recruitment to DNA appears to be mediated by different NuRD subunits and importantly can be affected by point mutations, with most evidence coming from the interaction with RBBP paralogs and FOG1 which supplies the foundations for the previously discussed affinity chromatography (Hong et al., 2005, Lejon et al., 2011). Using deletion mapping, one study found the substoichiometric ZNF219 can bind RBPP4, GATAD2A/B and CHD4 through its 4 N-terminal ZFs (Sharifi Tabar et al., 2019). Additionally, the FOG1 N-terminus (45aa) is conserved across mammals (Hong et al., 2005) and human FOG1, also known as ZNF89A or ZFPM1, is 1006aa and contains 9 ZFs as well as TED domains (Blum et al., 2025). One study reported point mutations switching conserved, charged aa's with glycines (R4G) or alanines (K5A) in the N-terminus of FOG1 abrogated NuRD-mediated gene repression by half through preventing MTA1 binding (Hong et al., 2005). On the other hand, some substitutions had no effect and others such as R10G altered binding to RBBP paralogs. Furthermore, ZMYND8 exclusively interacts with GATAD2A's PPPL motif through its ZF-like MYND motif, and genomic deletion of this MYND motif abrogates ZMYND8's interaction with NuRD and its functional

recruitment to sites of DNA damage (Spruijt et al., 2016). *De novo* missense ZMYND8 variants in the MYND motif (reported in neurodevelopmental disorders similar to ZFHX4) have also been shown to disrupt ZMYND8's interaction with GATAD2A (Dias et al., 2022). Mutations in ZEB2, a ZFHX4 family member, have also been shown to disrupt NuRD recruitment and downstream gene expression (Verstappen et al., 2008). Moreover, mutations in core subunits themselves also show similar trends, with *de novo* GATAD2A missense variants associating with neurodevelopmental disorders and disrupting interactions with CHD3, CHD4 and CHD5 (Werren et al., 2023). Around 80% of CHD4 missense mutations in endometrial cancer were predicted to impact protein function, with half of these residing in its ATPase and helicase domains (Le Gallo et al., 2012) compromising NuRD's DNA-repair activity without altering p53-antagonising functions (O'Shaughnessy and Hendrich, 2013). A NuRD mutagenesis screen in human adult erythroid precursors demonstrated different mutations conferred either protein loss, complex destabilisation or change of function (Sher et al., 2019). Collectively, this suggests binding motif integrity in both core subunits and substoichiometric interactors are crucial for NuRD complex formation, recruitment, binding and downstream function and missense mutations have the potential to affect all of the former, providing context-dependent effects.

Taking the above together, it's credible to suggest ZFHX4 could bind NuRD subunits through similar ZF-mediated interactions as by other validated substoichiometric interactors. Protein BLAST of FOG1 with ZFHX4 revealed 5 regions of similarity (starting from residues 695/1215/1402/1510/2305) and all are indicative of ZFHX4-ZFs. Despite not an overlay of the FOG1 N-terminus specifically, this supports potential binding of ZFHX4 to NuRD subunits through its ZFs. Furthermore, protein BLAST of ZNF219 with ZFHX4 revealed these 4 ZNF219-ZFs were a similar match to 4 ZFHX4-ZFs in the 1396-1613aa protein region, hinting at a possible similar interaction with NuRD subunits. Therefore, it's also conceivable that missense mutations in ZFHX4 can alter its interaction with NuRD given the dramatic effects observed with FOG1 point mutations. For example, the 1396aa-

1613aa ZF region described previously contains 24 different missense mutations in the combined clinical colon cancer cohort (Figure 5.1). Given the possibility of missense mutations also altering residues with PTMs, PTMs on substoichiometric interactors such as FOG1 SUMOylation (Snow et al., 2010) and ZMYND8 acetylation (Chen et al., 2018) have been shown important for function, largely predicted through changes to NuRD binding. SUMOylation in ZFHX3 has already been shown paramount for regulating its stability and function (Wu et al., 2020). ZFHX4 has many predicted PTMs and thus offers further predictions of change to function upon missense mutations.

Missense mutations in ZFHX4 may also alter its ability to bind DNA and dynamically regulate NuRD recruitment to different loci. A similar ZF-TF, CTCF is a key regulator of gene expression and somatic mutations within various ZF domains have been shown to alter its binding to some target sites, but not others, indicating a change rather than loss of function (Filippova et al., 2002). This has been postulated a result of altered affinities for specific base pairs by altered residues in ZFs such as K365T (Yang et al., 2020). Another study found somatic CTCF missense mutations were enriched pan-cancer (TCGA) particularly at residues affecting ZFs and differentially offered a spectrum in their ability to lose, change or gain CTCF function when modelled in human erythroid leukaemia cell lines (Bailey et al., 2021). Such mutations can induce functional changes affecting protein stability, folding, DNA-binding affinity and specificity at particular loci and transcriptional regulation (Bailey et al., 2021). Similar diverse functional effects have been reported with missense mutations in other ZF-TFs such as GATA3 (Takaku et al., 2018) and KLF1 (Gillinder et al., 2017) as well as in homeodomains of homeodomain-containing proteins (Kock et al., 2024). Interestingly, an analysis aggregating somatic mutations in all C2H2-ZF domains highlighted missense mutations affected motif position 9 (largely R9I) significantly 2.8 times more than expected by chance, even after controlling for mutation rate (Munro et al., 2018). Taken together this reiterates the profound and diverse effects a missense mutation can have on protein-DNA interactions despite classification as VUS.

5.5.4 Future directions

Whilst this chapter has revealed *ZFHX4*$^{-/-}$ can enhance immunogenicity *in vitro*, both model- and immunogenicity-technique-choices aren't representative of the clinical reality. This is far more nuanced as most patient mutations in *ZFHX4* are missense VUS, and immunogenicity is much more than T-cell activation and migration towards CM. To build on the associative link discussed in this thesis, a mechanistic link of these mutations to altered immune phenotypes in colon cancer is critical and further *in silico*, *in vitro* and *in vivo* efforts are warranted.

An immediate priority is systematic *in silico* analysis of recurrent *ZFHX4* missense mutations observed in colon cancer (Figure 5.1). Providing a useful triage system for selecting mutations for functional study, mutations could be chosen from the top 10 commonly altered residues across the four colon cancer cohorts (Table 5.1) and assessed in their predicted impact on protein structure, DNA binding and protein-protein interactions. Tools such as AlphaMissense (Cheng et al., 2023a), DynaMut2 (Rodrigues et al., 2021) and PolyPhen (Adzhubei et al., 2013) can predict structural or stability effects, while ZF-specific DNA binding predictors (Persikov and Singh, 2014) can help infer changes in DNA target specificity.

Next, these computational predictions should be validated *in vitro* to verify how ZFHX4 missense mutations affect transcriptional regulation and lead to a heightened immunogenicity phenotype. Base or prime editing would provide clinically-relevant models and allow many mechanistic experiments to compare *ZFHX4*^{mt} vs *ZFHX4*^{wt} contexts:

(1) IP-MS to identify changes in ZFHX4-protein interactions, including NuRD subunit and substoichiometric interactor recruitment

(2) CUT&RUN to assess changes in ZFHX4-DNA occupancy and potential shifts in chromatin targeting

(3) Transcriptomic (bulk/scRNAseq) and proteomic (WB/IF/MS) integration to identify downstream immune or tumourigenic pathways affected

(4) FOG1-affinity chromatography or similar NuRD pulldown techniques to assess changes in NuRD composition and stoichiometry

(5) *In vitro* NuRD activity assays to assess functional consequences on chromatin remodelling. Given its ATP-dependent remodelling and histone deacetylase activities (Xue et al., 1998), NuRD complex activity can be measured *in vitro* by incubating purified complex with recombinant histone octamers bound to specifically-sized DNA fragments in the presence of ATP (Saathoff et al., 2015). Band changes observed on SDS-PAGE gels are indicative of nucleosome remodelling and thus catalytic activity.

(6) *In vitro* co-culture systems and/or *in vivo* systems to phenotypically validate anti-tumour immune responses and immunotherapy sensitivity that's likely provided with *ZFHX4*^{mt}

These studies should be conducted in carefully selected model systems (normal vs cancer, cell line vs primary), given ZFHX4's context-specific roles and the issues already experienced with CUT&RUN. Cell line work would provide good starting points particularly for experimental optimisation, but ultimately primary organoid work and/or *in vivo* work would offer better representations of the colon cancer setting.

5.5.5 Conclusion

In summary, this Chapter provides compelling early evidence that *ZFHX4*$^{-/-}$ enhances immunogenicity, but to understand the clinical impact of *ZFHX4* missense mutations more refined modelling is essential. Unlike classical GoF or LoF mutations, the consequences of *ZFHX4* variants are more consistent with a 'change-of-function' mechanism where transcriptional output is rewired rather than simply increased or silenced. Several plausible mechanisms could underlie this hypothesis as predicted from existing literature. Firstly, they could alter ZFHX4-protein interactions, resulting in (a) disruption of canonical NuRD recruitment leading to incomplete/dysfunctional complexes and reduced tumourigenesis, (b) gained affinity for non-canonical cofactors that promote immunogenic gene programs or (c) altered PTMs at or near mutation sites further modifying protein interactions. Secondly, they could also alter ZFHX4-DNA binding specificity with or without NuRD recruitment, resulting in reduced occupancy at tumour-promoting loci such as EMT-related genes and/or enhanced occupancy at immune-related enhancers or promoters. Alongside a direct enhancement of immunogenicity, these mechanisms may contribute to a functional decoupling of tumourigenesis and immunogenicity: disrupting ZFHX4's role in maintaining an immunosuppressive, tumourigenic chromatin environment may indirectly enhance an immunogenicity phenotype.

Chapter Six: Discussion

In this thesis, I set out to investigate potential regulators of immunogenicity in colon cancer, focusing on somatically mutated genes predicted to positively immunosensitise a TMBL tumour through its association with multiple determinants of immunogenicity (mDOI) *in silico*. This was based on emerging concepts in the field: where immunogenicity is described by many factors other than solely mutational load, and mutation quality can just be as important as quantity for shaping immunogenicity. Following novel *in silico* analyses in Chapter 3 employing TCGA-COAD data, *ZFHX4* was identified as a candidate somatic mutational driver of immunogenicity. Interestingly, its expression was associated with poor prognosis and tumour-promoting processes such as epithelial-mesenchymal transition (EMT) across four independent colon cancer datasets. This added an unexpected but compelling layer to its relevance, further supporting its investigation *in vitro*. In addition to this, ZFHX4 remained poorly characterised in both physiological and oncogenic contexts, with most literature pertaining to germline loss-of-function (LoF) truncating heterozygous mutations in neurodevelopmental disorders. Virtually no studies address its role in the normal colon and colon cancer, highlighting the novelty of my work which is to my knowledge the first to investigate ZFHX4 in this setting and in the dual context of tumourigenicity and immunogenicity.

Chapters 4 and 5 of this thesis reflect the duality of investigations, exploring ZFHX4's role in tumourigenicity and immunogenicity respectively. A more extensive set of experiments was conducted in Chapter 4, due in part to the plethora of available and standardised *in vitro* assays for tumourigenic phenotypes compared to the relatively limited tools for studying immunogenicity. Importantly, the same *ZFHX4*$^{-/-}$ models were used across both chapters to maintain consistency and experimental feasibility. However, this approach was limited in the context of studying immunogenicity because knockout models inherently lack clinical relevance to missense mutations. Additionally, the immunogenicity assays employed were limited in scope, reflecting broader methodological constraints in modelling

tumour-immune interactions *in vitro*. Most models and techniques used throughout this project are imperfect and phenotypic results accompany trends with mixed statistical significance, however supplemented with a wealth of *in silico* data aggregated from four independent colon cancer cohorts, and existing literature on surrounding regulatory biology, a clearer understanding of ZFHX4's contribution to tumourigenicity and immunogenicity is evident.

Bringing all chapters together, this discussion sets out a role for ZFHX4 in colon cancer, where overexpression drives tumourigenesis and poor prognosis, and missense mutations promote immunogenicity, specifically in the TMBL setting (Figure 6.1). It's key to note these proposed mechanisms are speculative and require further *in silico*, *in vitro* and *in vivo* work to functionally validate tumourigenicity and immunogenicity phenotypes, as set out in each result-discussion section. Furthermore, considering ZFHX4's expression in stromal and neural compartments, and the known importance of both in tumourigenesis (Vales et al., 2019, Qin et al., 2023, Vaes et al., 2022), such future work could encompass these other cell types too.

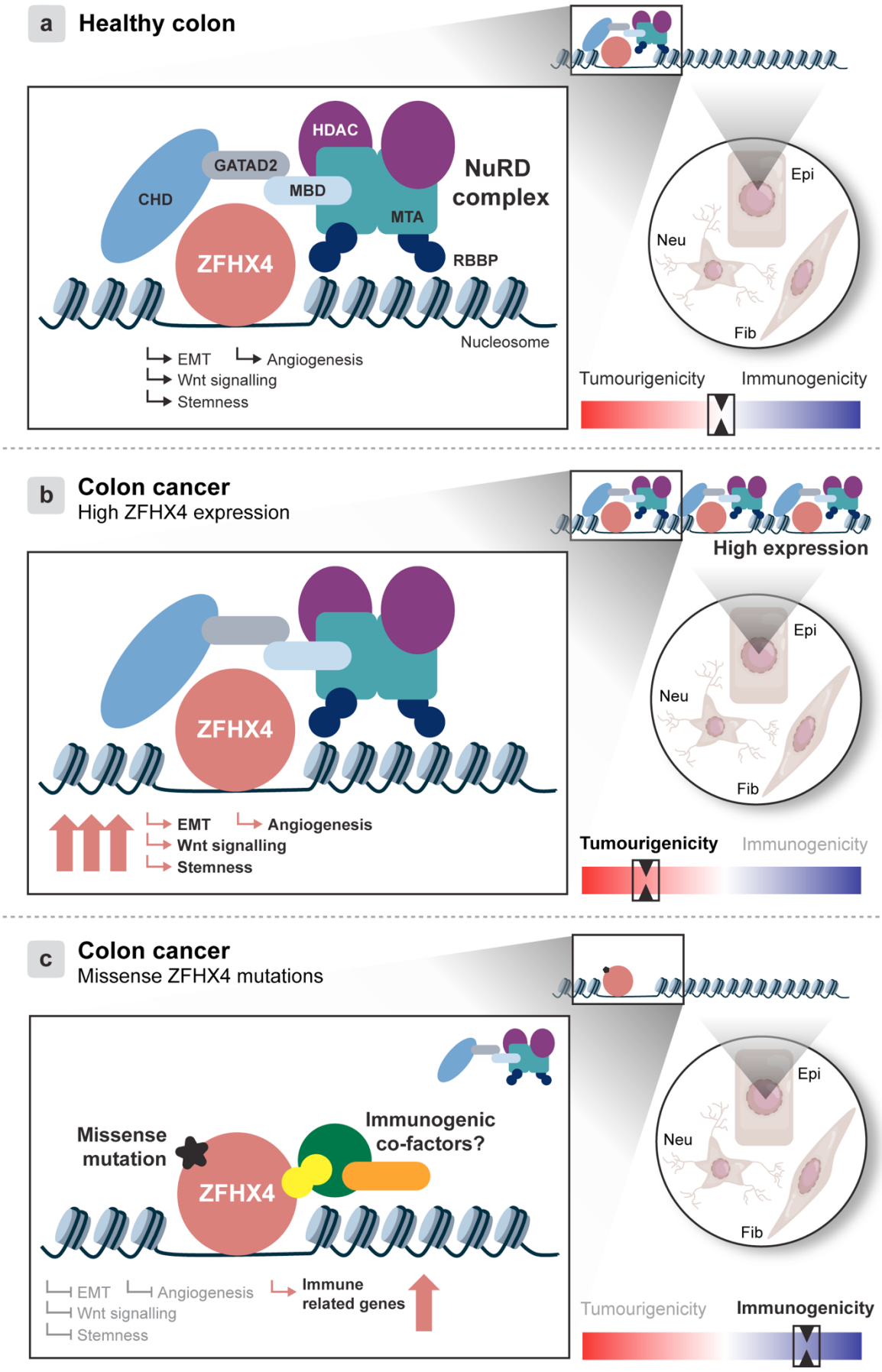

a Healthy colon
HDAC
GATAD2
MBD
MTA
CHD
NuRD complex
ZFHX4
RBBP
Nucleosome
EMT
Angiogenesis
Wnt signalling
Stemness
Epi
Neu
Fib
Tumourigenicity
Immunogenicity
b Colon cancer
High ZFHX4 expression
High expression
ZFHX4
EMT
Angiogenesis
Wnt signalling
Stemness
Epi
Neu
Fib
Tumourigenicity
Immunogenicity
c Colon cancer
Missense ZFHX4 mutations
Missense mutation
Immunogenic co-factors?
ZFHX4
EMT
Angiogenesis
Wnt signalling
Stemness
Immune related genes
Epi
Neu
Fib
Tumourigenicity
Immunogenicity

Figure 6.1 | ZFHX4 as a potential regulator of tumourigenicity and immunogenicity in colon cancer. Summarising all three results chapters with *in vitro* and *in silico* evidence, as well as surrounding literature, ZFHX4 and NuRD likely co-regulate key pathways during both embryonic development and maintenance in the adult normal colon **(A)**. High expression of *ZFHX4* **(B)** is predicted to dysregulate and thus enhance tumourigenic effects of NuRD and lead to an EMT-related phenotype associated with poor prognosis. Other phenotypes include enhanced stemness, Wnt signalling and angiogenesis as identified from the meta-analysis in Figure 4.41. Missense mutations in *ZFHX4* **(C)** likely result in 'change-of-function' which could describe (1) disruption of NuRD recruitment, (2) recruitment of pro-immunogenic co-factors and/or (3) altered DNA occupancy away from tumour-promoting loci to immune-related loci, leading to overall heightened immunogenicity particularly in the TMBL setting. *ZFHX4* can be expressed in epithelial (Epi), neuronal (Neu) and fibroblast (Fib) compartments in the colon and whilst this ZFHX4 mechanism is with respect mainly to colon epithelium, similar mechanisms may present in these other cell types. To note, this figure is purely speculative and further mechanistic work is required to validate this. Furthermore, tumourigenicity and immunogenicity are largely inversely correlated and whilst represented as a spectrum here for simplicity, it's important to note they are not binary or completely mutually exclusive. *The structure of NuRD with its subunits is accurate and based off recent stoichiometry experiments (Leighton and Williams, 2020, Basu et al., 2023, Low et al., 2020).*

Following its implication in tumourigenesis, one important question remains – **could *ZFHX4* be an oncogene or tumour suppressor gene?** Fundamentally a disease of somatic evolution, cancer is shaped by the acquisition of genetic alterations that confer a selective growth advantage to cells (Weinberg, 1994, Vogelstein and Kinzler, 2004). These alterations are classed as driver gene mutations, which by definition occur in cancer driver genes (CDGs) and are under positive selection, with a typical tumour containing 2-8 (Vogelstein et al., 2013). CDGs can be broadly classified into oncogenes or tumour suppressor genes (TSGs) depending on how their alterations promote tumourigenesis (Klein and Klein, 1985). For example, alterations in oncogenes are typically gain-of-function (GoF) and recurrent in hotspot regions, whereas TSG alterations are commonly protein-truncating indels and LoF (Vogelstein et al., 2013). Additionally, copy number alterations (CNAs), described as amplification or deletion events of large genomic regions >50bp, tend to involve oncogenes and TSGs where protein products are gained or lost respectively (Vogelstein et al., 2013, Li et al., 2020).

A key goal of cancer research is the identification and subsequent targeting of CDGs, which have supplied the basis for anti-cancer therapies available clinically (Martinez-Jimenez et

al., 2020). This identification is ever-increasing, as large-scale consortia such as TCGA (Hoadley et al., 2018), 100K genomes project (Sosinsky et al., 2024), ICGC (International Cancer Genome et al., 2010) and consequently PCAWG (Gerstung et al., 2020, Consortium, 2020) continue to catalogue mutations in many tumours across various types. Curated databases such as COSMIC (Sondka et al., 2024) and cBioPortal (Cerami et al., 2012) provide robust integration of such sequencing studies, and bioinformatic tools such as MutSigCV (Lawrence et al., 2013), dNdS (Martincorena et al., 2017), PanSoftware (Bailey et al., 2018) and IntOGen (Martinez-Jimenez et al., 2020) enable prediction of driver gene mutations and thus CDGs based on signals of positive selection *in silico*. This is important because CDGs can also possess neutral passenger mutations which confer no selective growth advantage. Such positive signals encompass mutation frequency, mutation clustering and predicted functional impact and incorporate background mutation models to account for neutral mutagenesis (Martinez-Jimenez et al., 2020). Identification of CDGs here is slightly different to those identified through expert-mediated literature curation describing functional and mechanistic approaches that have linked a gene with tumourigenesis, particularly described by the hallmarks of cancer (Futreal et al., 2004, Sondka et al., 2018, Hanahan, 2022). Therefore, the key to CDG identification lies both in *in silico* prediction of positive selection, as well as functional validation by *in vitro* and *in vivo* means. Currently to my knowledge, *ZFHX4* is not considered a CDG by the former or the latter, particularly because its predominating missense variants are of unknown significance (VUS) and there is little literature exploring a role in tumourigenesis. This contrasts with its paralog *ZFHX3*, a 10x larger gene currently classified as a CDG through both *in silico* (Gonzalez-Perez et al., 2013) and literature curation (Sondka et al., 2018) approaches. Here, *ZFHX3* predominates with truncating indel mutations and has been classed as a TSG in prostate cancer (Hu et al., 2024). Interestingly, *ZFHX4* does not display the same mutational patterns, indicating potentially more nuanced biology.

However, because CDG discovery has largely prioritised signals of positive selection, important subtle aspects of ZFHX4 biology may have been overlooked. Positive selection has been traditionally thought to outweigh negative selection in cancer development (Martincorena et al., 2017), but recent evidence suggests negative selection is equally important. A key study employing MutMatch analysed selection signals in 23,000 tumour genomes from 117 cancer types and explicitly controlled for both mutation rate and CN state, showing negative selection was prevalent for nonsynonymous mutations in oncogenes (Besedina and Supek, 2024). A novel approach toward characterising oncogenes and TSGs, this study remarkably identified *ZFHX4* in colorectal cancer as a top-ranked candidate presenting with negative selection for nonsense LoF mutations, demonstrating suggestive oncogene-like signal. Furthermore, they also demonstrate *ZFHX4* exhibits hotspot-restricted positive selection and negative selection outside of hotspots in some tumour types (lung, uterine, head and neck), an observation not previously reported. Interestingly, *CHD4* also displayed the same patterns. These findings are consistent with (1) the absence of nonsense and LoF mutations across all four independent colon cancer cohorts in this thesis, (2) the germline evidence demonstrating *ZFHX4* is highly intolerant to LoF indels and (3) *ZFHX4*$^{+/-}$ individuals exhibiting a neurodevelopmental phenotype in humans demonstrating likely haploinsufficiency. Furthermore, whilst CNAs were not investigated in depth in this thesis, TCGA-COAD data showed CNAs in *ZFHX4* were always amplifications comparative to deletions, unlike the TSG *ZFHX3* which presents with CN deletions.

Critically, taken together this suggests complete ZFHX4 LoF is detrimental to colon tumour fitness which requires baseline *ZFHX4* function as a non-canonical oncogene. This is compatible with a scenario where ZFHX4 maintains dosage-sensitive transcriptional programs required for tumour fitness and subtle missense changes could retune this program without outright loss. Importantly, this fits with the observation that *ZFHX4* expression associates with tumourigenicity, while missense mutations link instead to

immunogenicity. Furthermore, this exemplifies how the increasing wealth of sequencing datasets and bioinformatic tools to study CDGs are constantly changing traditional views on CDGs, cancer driver mutations, oncogenes and TSGs and particularly that consideration of CN is important.

Adding another layer of complexity, functional validation of oncogenes and TSGs is imperative. According to the 20/20 rule which classifies TSGs if they contain >20% truncating mutations (Vogelstein et al., 2013), SOX9 should be categorised as a TSG in the CPTAC colon cancer cohort (Vasaikar et al., 2019). However, its overexpression in cancer reflects a more oncogenic-like role, and functional work on *SOX9* is consistent with truncating mutations enhancing its stability by removing ubiquitination sites (Matheu et al., 2012). This reiterates the requirement of *in vitro* validation for both ZFHX4's non-canonical oncogenic role in tumourigenesis and regulation of immunogenicity in colon cancer.

It's important to emphasise that a mutation which doesn't drive tumour growth can still drive immunogenicity. In other words, the search for CDGs (alterations that promote tumourigenesis) is distinct from the search for immunogenicity-driver genes (IDGs; alterations that directly enhance or supress anti-tumour immune responses) though both may identify overlapping candidates. Just as the field has systematically mapped CDGs to enable targeted cancer therapies, immuno-oncology also appears to be steering towards identifying and validating IDGs for the advancement of immunotherapy, though this process if far less defined in the literature. Future work may find *ZFHX4* to be a validated IDG as well as a non-canonical CDG.

Beyond underscoring the importance of identifying somatic mutational drivers of immunogenicity in MSS colon cancer, this work establishes ZFHX4 as a compelling candidate for deeper mechanistic study. Its expression across multiple tumour microenvironment compartments, its role in NuRD complex activity and EMT regulation, and its apparent ability to modulate both tumourigenicity and immunogenicity highlight its

biological versatility. Taken together, the findings from this thesis suggest that ZFHX4 functions as a context-dependent and dosage-sensitive transcription factor whose perturbation may shift the balance between tumour progression and immune activation. Continued investigation, spanning *in silico*, *in vitro* and *in vivo* approaches, will be essential to delineate how ZFHX4 mediates these dual effects and to determine whether it represents a tractable target for therapeutic modulation in colorectal and other cancers.

There are many avenues for scientific exploration, and I hope this thesis sets the scene for future work elucidating the mysterious transcription factor that is *ZFHX4*.

References

Acha-Sagredo, A., Andrei, P., Clayton, K., Taggart, E., Antoniotti, C., Woodman, C. A., Afrache, H., Fourny, C., Armero, M., Moinudeen, H. K., Green, M., Bhardwaj, N., Mikolajczak, A., Rodriguez-Lopez, M., Crawford, M., Connick, E., Lim, S., Hobson, P., Linares, J., Ignatova, E., Pelka, D., Smyth, E. C., Diamantis, N., Sosnowska, D., Carullo, M., Ciraci, P., Bergamo, F., Intini, R., Nye, E., Barral, P., Mishto, M., Arnold, J. N., Lonardi, S., Cremolini, C., Fontana, E., Rodriguez-Justo, M. & Ciccarelli, F. D. 2025. A constitutive interferon-high immunophenotype defines response to immunotherapy in colorectal cancer. *Cancer Cell,* 43**,** 292–307 e7.

Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet,* Chapter 7**,** Unit7 20.

Aguilera, C., Nakagawa, K., Sancho, R., Chakraborty, A., Hendrich, B. & Behrens, A. 2011. c-Jun N-terminal phosphorylation antagonises recruitment of the Mbd3/NuRD repressor complex. *Nature,* 469**,** 231–5.

Akbani, R., Zhang, N., Broom, B. M., Weinstein, J. N., Casasent, T., Melott, J. M., Wakefield, C., Cleland, J. A., Wong, W. C. & Ryan, M. 2025. *TCGA Batch Effects Viewer version 2.0* [Online]. Available: https://bioinformatics.mdanderson.org/public-software/tcga-batch-effects/ [Accessed 23 June 2023].

Aleksandrova, K., Pischon, T., Jenab, M., Bueno-De-Mesquita, H. B., Fedirko, V., Norat, T., Romaguera, D., Knuppel, S., Boutron-Ruault, M. C., Dossus, L., Dartois, L., Kaaks, R., Li, K., Tjonneland, A., Overvad, K., Quiros, J. R., Buckland, G., Sanchez, M. J., Dorronsoro, M., Chirlaque, M. D., Barricarte, A., Khaw, K. T., Wareham, N. J., Bradbury, K. E., Trichopoulou, A., Lagiou, P., Trichopoulos, D., Palli, D., Krogh, V., Tumino, R., Naccarati, A., Panico, S., Siersema, P. D., Peeters, P. H., Ljuslinder, I., Johansson, I., Ericson, U., Ohlsson, B., Weiderpass, E., Skeie, G., Borch, K. B., Rinaldi, S., Romieu, I., Kong, J., Gunter, M. J., Ward, H. A., Riboli, E. & Boeing, H. 2014. Combined impact of healthy lifestyle factors on colorectal cancer: a large European cohort study. *BMC Med,* 12**,** 168.

Alitalo, K., Schwab, M., Lin, C. C., Varmus, H. E. & Bishop, J. M. 1983. Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. *Proc Natl Acad Sci U S A,* 80**,** 1707–11.

Almeida, L. G., Sakabe, N. J., Deoliveira, A. R., Silva, M. C., Mundstein, A. S., Cohen, T., Chen, Y. T., Chua, R., Gurung, S., Gnjatic, S., Jungbluth, A. A., Caballero, O. L., Bairoch, A., Kiesler, E., White, S. L., Simpson, A. J., Old, L. J., Camargo, A. A. & Vasconcelos, A. T. 2009. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res,* 37**,** D816–9.

Almeida-Marques, C., Rolfs, F., Piersma, S. R., Bijnsdorp, I. V., Pham, T. V., Knol, J. C. & Jimenez, C. R. 2024. Secretome processing for proteomics: A methods comparison. *Proteomics,* 24**,** e2300262.

Andre, T., Elez, E., Lenz, H. J., Jensen, L. H., Touchefeu, Y., Van Cutsem, E., Garcia-Carbonero, R., Tougeron, D., Mendez, G. A., Schenker, M., De La Fouchardiere, C., Limon, M. L., Yoshino, T., Li, J., Manzano Mozo, J. L., Dahan, L., Tortora, G., Chalabi, M., Goekkurt, E., Braghiroli, M. I., Joshi, R., Cil, T., Aubin, F., Cela, E., Chen, T., Lei, M., Jin, L., Blum, S. I. & Lonardi, S. 2025. Nivolumab plus ipilimumab versus nivolumab

in microsatellite instability-high metastatic colorectal cancer (CheckMate 8HW): a randomised, open-label, phase 3 trial. *Lancet,* 405**,** 383–395.

Andre, T., Lonardi, S., Wong, K. Y. M., Lenz, H. J., Gelsomino, F., Aglietta, M., Morse, M. A., Van Cutsem, E., Mcdermott, R., Hill, A., Sawyer, M. B., Hendlisz, A., Neyns, B., Abdullaev, S., Memaj, A., Lei, M., Dixon, M., Kopetz, S. & Overman, M. J. 2022. Nivolumab plus low-dose ipilimumab in previously treated patients with microsatellite instability-high/mismatch repair-deficient metastatic colorectal cancer: 4-year follow-up from CheckMate 142. *Ann Oncol,* 33**,** 1052–1060.

Andre, T., Shiu, K. K., Kim, T. W., Jensen, B. V., Jensen, L. H., Punt, C., Smith, D., Garcia-Carbonero, R., Benavides, M., Gibbs, P., De La Fouchardiere, C., Rivera, F., Elez, E., Bendell, J., Le, D. T., Yoshino, T., Van Cutsem, E., Yang, P., Farooqui, M. Z. H., Marinello, P., Diaz, L. A., Jr. & Investigators, K.-. 2020. Pembrolizumab in Microsatellite-Instability-High Advanced Colorectal Cancer. *N Engl J Med,* 383**,** 2207–2218.

Angelova, M., Charoentong, P., Hackl, H., Fischer, M. L., Snajder, R., Krogsdam, A. M., Waldner, M. J., Bindea, G., Mlecnik, B., Galon, J. & Trajanoski, Z. 2015. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol,* 16**,** 64.

Anthony, H. & Seoighe, C. 2024. Performance assessment of computational tools to detect microsatellite instability. *Brief Bioinform,* 25.

Antona, A., Leo, G., Favero, F., Varalda, M., Venetucci, J., Faletti, S., Todaro, M., Mazzucco, E., Soligo, E., Saglietti, C., Stassi, G., Manfredi, M., Pelicci, G., Cora, D., Valente, G. & Capello, D. 2023. Targeting lysine-specific demethylase 1 (KDM1A/LSD1) impairs colorectal cancer tumorigenesis by affecting cancer cells stemness, motility, and differentiation. *Cell Death Discov,* 9**,** 201.

Antoniotti, C., Rossini, D., Pietrantonio, F., Catteau, A., Salvatore, L., Lonardi, S., Boquet, I., Tamberi, S., Marmorino, F., Moretto, R., Ambrosini, M., Tamburini, E., Tortora, G., Passardi, A., Bergamo, F., Kassambara, A., Sbarrato, T., Morano, F., Ritorto, G., Borelli, B., Boccaccino, A., Conca, V., Giordano, M., Ugolini, C., Fieschi, J., Papadopulos, A., Massoue, C., Aprile, G., Antonuzzo, L., Gelsomino, F., Martinelli, E., Pella, N., Masi, G., Fontanini, G., Boni, L., Galon, J., Cremolini, C. & Investigators, G. F. 2022. Upfront FOLFOXIRI plus bevacizumab with or without atezolizumab in the treatment of patients with metastatic colorectal cancer (AtezoTRIBE): a multicentre, open-label, randomised, controlled, phase 2 trial. *Lancet Oncol,* 23**,** 876–887.

Anzalone, A. V., Koblan, L. W. & Liu, D. R. 2020. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat Biotechnol,* 38**,** 824–844.

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A. & Liu, D. R. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature,* 576**,** 149–157.

Arnaoutova, I. & Kleinman, H. K. 2010. In vitro angiogenesis: endothelial cell tube formation on gelled basement membrane extract. *Nat Protoc,* 5**,** 628–35.

Arvindekar, S., Jackman, M. J., Low, J. K. K., Landsberg, M. J., Mackay, J. P. & Viswanath, S. 2022. Molecular architecture of nucleosome remodeling and deacetylase sub-

complexes by integrative structure determination. *Protein Sci,* 31**,** e4387.

Ashworth, A., Lord, C. J. & Reis-Filho, J. S. 2011. Genetic interactions in cancer progression and treatment. *Cell,* 145**,** 30–8.

Aslan, M., Hsu, E. C., Liu, S. & Stoyanova, T. 2021. Quantifying the invasion and migration ability of cancer cells with a 3D Matrigel drop invasion assay. *Biol Methods Protoc,* 6**,** bpab014.

Atkin, W., Wooldrage, K., Parkin, D. M., Kralj-Hans, I., Macrae, E., Shah, U., Duffy, S. & Cross, A. J. 2017. Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomised controlled trial. *Lancet,* 389**,** 1299–1311.

Auslander, N., Zhang, G., Lee, J. S., Frederick, D. T., Miao, B., Moll, T., Tian, T., Wei, Z., Madan, S., Sullivan, R. J., Boland, G., Flaherty, K., Herlyn, M. & Ruppin, E. 2018. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat Med,* 24**,** 1545–1549.

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

Bagaev, A., Kotlov, N., Nomie, K., Svekolkin, V., Gafurov, A., Isaeva, O., Osokin, N., Kozlov, I., Frenkel, F., Gancharova, O., Almog, N., Tsiper, M., Ataullakhanov, R. & Fowler, N. 2021. Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell,* 39**,** 845–865 e7.

Bailey, C. G., Gupta, S., Metierre, C., Amarasekera, P. M. S., O'young, P., Kyaw, W., Laletin, T., Francis, H., Semaan, C., Sharifi Tabar, M., Singh, K. P., Mullighan, C. G., Wolkenhauer, O., Schmitz, U. & Rasko, J. E. J. 2021. Structure-function relationships explain CTCF zinc finger mutation phenotypes in cancer. *Cell Mol Life Sci,* 78**,** 7519–7536.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., Rubio-Perez, C., Nagarajan, N., Cortes-Ciriano, I., Zhou, D. C., Liang, W. W., Hess, J. M., Yellapantula, V. D., Tamborero, D., Gonzalez-Perez, A., Suphavilai, C., Ko, J. Y., Khurana, E., Park, P. J., Van Allen, E. M., Liang, H., Group, M. C. W., Cancer Genome Atlas Research, N., Lawrence, M. S., Godzik, A., Lopez-Bigas, N., Stuart, J., Wheeler, D., Getz, G., Chen, K., Lazar, A. J., Mills, G. B., Karchin, R. & Ding, L. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell,* 173**,** 371–385 e18.

Balint, S., Muller, S., Fischer, R., Kessler, B. M., Harkiolaki, M., Valitutti, S. & Dustin, M. L. 2020. Supramolecular attack particles are autonomous killing entities released from cytotoxic T cells. *Science,* 368**,** 897–901.

Barker, N., Ridgway, R. A., Van Es, J. H., Van De Wetering, M., Begthel, H., Van Den Born, M., Danenberg, E., Clarke, A. R., Sansom, O. J. & Clevers, H. 2009. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature,* 457**,** 608–11.

Barker, N., Van Es, J. H., Kuipers, J., Kujala, P., Van Den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P. J. & Clevers, H. 2007. Identification

of stem cells in small intestine and colon by marker gene Lgr5. *Nature,* 449**,** 1003–7.

Basu, S., Shukron, O., Hall, D., Parutto, P., Ponjavic, A., Shah, D., Boucher, W., Lando, D., Zhang, W., Reynolds, N., Sober, L. H., Jartseva, A., Ragheb, R., Ma, X., Cramard, J., Floyd, R., Balmer, J., Drury, T. A., Carr, A. R., Needham, L. M., Aubert, A., Communie, G., Gor, K., Steindel, M., Morey, L., Blanco, E., Bartke, T., Di Croce, L., Berger, I., Schaffitzel, C., Lee, S. F., Stevens, T. J., Klenerman, D., Hendrich, B. D., Holcman, D. & Laue, E. D. 2023. Live-cell three-dimensional single-molecule tracking reveals modulation of enhancer dynamics by NuRD. *Nat Struct Mol Biol,* 30**,** 1628–1639.

Becker, W. R., Nevins, S. A., Chen, D. C., Chiu, R., Horning, A. M., Guha, T. K., Laquindanum, R., Mills, M., Chaib, H., Ladabaum, U., Longacre, T., Shen, J., Esplin, E. D., Kundaje, A., Ford, J. M., Curtis, C., Snyder, M. P. & Greenleaf, W. J. 2022. Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *Nat Genet,* 54**,** 985–995.

Bedognetti, D., Hendrickx, W., Ceccarelli, M., Miller, L. D. & Seliger, B. 2016. Disentangling the relationship between tumor genetic programs and immune responsiveness. *Curr Opin Immunol,* 39**,** 150–8.

Bellacosa, A. 2001. Functional interactions and signaling properties of mammalian DNA mismatch repair proteins. *Cell Death Differ,* 8**,** 1076–92.

Berg, K. C. G., Eide, P. W., Eilertsen, I. A., Johannessen, B., Bruun, J., Danielsen, S. A., Bjornslett, M., Meza-Zepeda, L. A., Eknaes, M., Lind, G. E., Myklebost, O., Skotheim, R. I., Sveen, A. & Lothe, R. A. 2017. Multi-omics of 34 colorectal cancer cell lines - a resource for biomedical studies. *Mol Cancer,* 16**,** 116.

Berry, F. B., Miura, Y., Mihara, K., Kaspar, P., Sakata, N., Hashimoto-Tamaoki, T. & Tamaoki, T. 2001. Positive and negative regulation of myogenic differentiation of C2C12 cells by isoforms of the multiple homeodomain zinc finger transcription factor ATBF1. *J Biol Chem,* 276**,** 25057–65.

Bertucci, F., Finetti, P., Simeone, I., Hendrickx, W., Wang, E., Marincola, F. M., Viens, P., Mamessier, E., Ceccarelli, M., Birnbaum, D. & Bedognetti, D. 2018. The immunologic constant of rejection classification refines the prognostic value of conventional prognostic signatures in breast cancer. *Br J Cancer,* 119**,** 1383–1391.

Besedina, E. & Supek, F. 2024. Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver mutations. *Nat Commun,* 15**,** 6139.

Bettington, M., Walker, N., Rosty, C., Brown, I., Clouston, A., Mckeone, D., Pearson, S. A., Leggett, B. & Whitehall, V. 2017. Clinicopathological and molecular features of sessile serrated adenomas with dysplasia or carcinoma. *Gut,* 66**,** 97–106.

Bianchi, N. O., Bianchi, M. S. & Kere, J. 1991. DNA discontinuities in the domain of amplified human MYC oncogenes. *Genes Chromosomes Cancer,* 3**,** 136–41.

Bishop, M. R., Diaz Perez, K. K., Sun, M., Ho, S., Chopra, P., Mukhopadhyay, N., Hetmanski, J. B., Taub, M. A., Moreno-Uribe, L. M., Valencia-Ramirez, L. C., Restrepo Muneton, C. P., Wehby, G., Hecht, J. T., Deleyiannis, F., Weinberg, S. M., Wu-Chou, Y. H., Chen, P. K., Brand, H., Epstein, M. P., Ruczinski, I., Murray, J. C., Beaty, T. H., Feingold, E., Lipinski, R. J., Cutler, D. J., Marazita, M. L. & Leslie, E. J. 2020. Genome-wide Enrichment of De Novo Coding Mutations in Orofacial Cleft Trios. *Am J Hum Genet,* 107**,** 124–136.

Blankenstein, T., Coulie, P. G., Gilboa, E. & Jaffee, E. M. 2012. The determinants of tumour immunogenicity. *Nat Rev Cancer,* 12**,** 307–13.

Blaszczak, W., Liu, G., Zhu, H., Barczak, W., Shrestha, A., Albayrak, G., Zheng, S., Kerr, D., Samsonova, A. & La Thangue, N. B. 2021. Immune modulation underpins the anti-cancer activity of HDAC inhibitors. *Mol Oncol,* 15**,** 3280–3298.

Blum, M., Andreeva, A., Florentino, L. C., Chuguransky, S. R., Grego, T., Hobbs, E., Pinto, B. L., Orr, A., Paysan-Lafosse, T., Ponamareva, I., Salazar, G. A., Bordin, N., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunic, I., Llinares-Lopez, F., Marchler-Bauer, A., Meng-Papaxanthos, L., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Piovesan, D., Rivoire, C., Sigrist, C. J. A., Thanki, N., Thibaud-Nissen, F., Thomas, P. D., Tosatto, S. C. E., Wu, C. H. & Bateman, A. 2025. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res,* 53**,** D444–D456.

Boettcher, M. & Mcmanus, M. T. 2015. Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR. *Mol Cell,* 58**,** 575–85.

Boire, A., Burke, K., Cox, T. R., Guise, T., Jamal-Hanjani, M., Janowitz, T., Kaplan, R., Lee, R., Swanton, C., Vander Heiden, M. G. & Sahai, E. 2024. Why do patients with cancer die? *Nat Rev Cancer,* 24**,** 578–589.

Bonneville, R., Krook, M. A., Kautto, E. A., Miya, J., Wing, M. R., Chen, H. Z., Reeser, J. W., Yu, L. & Roychowdhury, S. 2017. Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precis Oncol,* 2017.

Boretto, M., Geurts, M. H., Gandhi, S., Ma, Z., Staliarova, N., Celotti, M., Lim, S., He, G. W., Millen, R., Driehuis, E., Begthel, H., Smabers, L., Roodhart, J., Van Es, J., Wu, W. & Clevers, H. 2024. Epidermal growth factor receptor (EGFR) is a target of the tumor-suppressor E3 ligase FBXW7. *Proc Natl Acad Sci U S A,* 121**,** e2309902121.

Bornelov, S., Reynolds, N., Xenophontos, M., Gharbi, S., Johnstone, E., Floyd, R., Ralser, M., Signolet, J., Loos, R., Dietmann, S., Bertone, P. & Hendrich, B. 2018. The Nucleosome Remodeling and Deacetylation Complex Modulates Chromatin Structure at Sites of Active Transcription to Fine-Tune Gene Expression. *Mol Cell,* 71**,** 56–72 e4.

Borras, D. M., Verbandt, S., Ausserhofer, M., Sturm, G., Lim, J., Verge, G. A., Vanmeerbeek, I., Laureano, R. S., Govaerts, J., Sprooten, J., Hong, Y., Wall, R., De Hertogh, G., Sagaert, X., Bislenghi, G., D'hoore, A., Wolthuis, A., Finotello, F., Park, W. Y., Naulaerts, S., Tejpar, S. & Garg, A. D. 2023. Single cell dynamics of tumor specificity vs bystander activity in CD8(+) T cells define the diverse immune landscapes in colorectal cancer. *Cell Discov,* 9**,** 114.

Bortolomeazzi, M., Keddar, M. R., Montorsi, L., Acha-Sagredo, A., Benedetti, L., Temelkovski, D., Choi, S., Petrov, N., Todd, K., Wai, P., Kohl, J., Denner, T., Nye, E., Goldstone, R., Ward, S., Wilson, G. A., Al Bakir, M., Swanton, C., John, S., Miles, J., Larijani, B., Kunene, V., Fontana, E., Arkenau, H. T., Parker, P. J., Rodriguez-Justo, M., Shiu, K. K., Spencer, J. & Ciccarelli, F. D. 2021. Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts. *Gastroenterology,* 161**,** 1179–1193.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I. & Jemal, A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin,* 74**,** 229–263.

Bretthauer, M., Loberg, M., Wieszczy, P., Kalager, M., Emilsson, L., Garborg, K., Rupinski, M., Dekker, E., Spaander, M., Bugajski, M., Holme, O., Zauber, A. G., Pilonis, N. D., Mroz, A., Kuipers, E. J., Shi, J., Hernan, M. A., Adami, H. O., Regula, J., Hoff, G., Kaminski, M. F. & Nord, I. C. C. S. G. 2022. Effect of Colonoscopy Screening on Risks of Colorectal Cancer and Related Death. *N Engl J Med,* 387**,** 1547–1556.

Briukhovetska, D., Dorr, J., Endres, S., Libby, P., Dinarello, C. A. & Kobold, S. 2021. Interleukins in cancer: from biology to therapy. *Nat Rev Cancer,* 21**,** 481–499.

Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H. & Holt, R. A. 2014. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res,* 24**,** 743–50.

Bullock, A. J., Schlechter, B. L., Fakih, M. G., Tsimberidou, A. M., Grossman, J. E., Gordon, M. S., Wilky, B. A., Pimentel, A., Mahadevan, D., Balmanoukian, A. S., Sanborn, R. E., Schwartz, G. K., Abou-Alfa, G. K., Segal, N. H., Bockorny, B., Moser, J. C., Sharma, S., Patel, J. M., Wu, W., Chand, D., Rosenthal, K., Mednick, G., Delepine, C., Curiel, T. J., Stebbing, J., Lenz, H. J., O'day, S. J. & El-Khoueiry, A. B. 2024. Botensilimab plus balstilimab in relapsed/refractory microsatellite stable metastatic colorectal cancer: a phase 1 trial. *Nat Med,* 30**,** 2558–2567.

Bundred, J., Lal, N., Chan, D. K. H. & Buczacki, S. J. A. 2025. Lymph node yield as a surrogate marker for tumour biology and prognosis in colon cancer. *Br J Cancer,* 132**,** 643–651.

Bürglin, T. R. 2011. *Homeodomain Subtypes and Functional Diversity*, Springer.

Burnet, M. 1957. Cancer; a biological approach. I. The processes of control. *Br Med J,* 1**,** 779–86.

C.Z.I.Cellscienceprogram, Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S. M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., Cherry, J. M., Chi, T., Chien, J., Dorman, L., Garcia-Nieto, P., Gloria, N., Hastie, M., Hegeman, D., Hilton, J., Huang, T., Infeld, A., Istrate, A. M., Jelic, I., Katsuya, K., Kim, Y. J., Liang, K., Lin, M., Lombardo, M., Marshall, B., Martin, B., Mcdade, F., Megill, C., Patel, N., Predeus, A., Raymor, B., Robatmili, B., Rogers, D., Rutherford, E., Sadgat, D., Shin, A., Small, C., Smith, T., Sridharan, P., Tarashansky, A., Tavares, N., Thomas, H., Tolopko, A., Urisko, M., Yan, J., Yeretssian, G., Zamanian, J., Mani, A., Cool, J. & Carr, A. 2025. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res,* 53**,** D886–D900.

Cai, Y., Geutjes, E. J., De Lint, K., Roepman, P., Bruurs, L., Yu, L. R., Wang, W., Van Blijswijk, J., Mohammad, H., De Rink, I., Bernards, R. & Baylin, S. B. 2014. The NuRD complex cooperates with DNMTs to maintain silencing of key colorectal tumor suppressor genes. *Oncogene,* 33**,** 2157–68.

Campbell, B. B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., De Borja, R., Davidson, S., Edwards, M., Elvin, J. A., Hodel, K. P., Zahurancik, W. J., Suo, Z., Lipman, T., Wimmer, K., Kratz, C. P., Bowers, D. C., Laetsch, T. W., Dunn, G. P., Johanns, T. M., Grimmer, M. R., Smirnov, I. V., Larouche, V., Samuel, D., Bronsema, A., Osborn, M., Stearns, D., Raman, P., Cole, K. A., Storm, P. B., Yalon, M., Opocher, E., Mason, G., Thomas, G. A., Sabel, M., George, B., Ziegler, D. S., Lindhorst, S., Issai, V. M., Constantini, S., Toledano, H., Elhasid, R., Farah, R., Dvir, R., Dirks, P., Huang, A., Galati, M. A., Chung, J., Ramaswamy, V., Irwin, M. S., Aronson, M., Durno, C., Taylor, M. D., Rechavi, G., Maris, J. M., Bouffet, E., Hawkins, C., Costello, J. F., Meyn, M. S.,

Pursell, Z. F., Malkin, D., Tabori, U. & Shlien, A. 2017. Comprehensive Analysis of Hypermutation in Human Cancer. *Cell,* 171**,** 1042–1056 e10.

Cancer Genome Atlas, N. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* 487**,** 330–7.

Cancer Research Uk. 2025. *Bowel cancer statistics* [Online]. Cancer Research UK. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#BowelCS1 [Accessed 7 Oct 2025].

Cancer Research Uk. 2025. *TNM staging for bowel cancer* [Online]. Cancer Research UK. Available: https://www.cancerresearchuk.org/about-cancer/bowel-cancer/stages-types-and-grades/tnm-staging [Accessed 7 Oct 2025].

Carethers, J. M., Chauhan, D. P., Fink, D., Nebel, S., Bresalier, R. S., Howell, S. B. & Boland, C. R. 1999. Mismatch repair proficiency and in vitro response to 5-fluorouracil. *Gastroenterology,* 117**,** 123–31.

Carpentier, G. 2010. Protein Array Analyzer for ImageJ. Available at: http://image.bio.methods.free.fr/ImageJ/?Protein-Array-Analyzer-for-ImageJ.html.

Carpentier, G., Berndt, S., Ferratge, S., Rasband, W., Cuendet, M., Uzan, G. & Albanese, P. 2020. Angiogenesis Analyzer for ImageJ - A comparative morphometric analysis of "Endothelial Tube Formation Assay" and "Fibrin Bead Assay". *Sci Rep,* 10**,** 11568.

Cattaneo, C., Scardellato, S., Mauri, G., Magnani, Z., Matafora, V., Scheper, W., Siena, S., Voest, E., Bonini, C. & Bardelli, A. 2025. *Immune surveillance is actively impaired by the secretome of mismatch repair proficient colorectal cancers* [Online]. EACR. Available: https://storage.unitedwebnetwork.com/files/991/abstract/EACR25-1534.htm [Accessed].

Cattaneo, C. M., Dijkstra, K. K., Fanchi, L. F., Kelderman, S., Kaing, S., Van Rooij, N., Van Den Brink, S., Schumacher, T. N. & Voest, E. E. 2020. Tumor organoid-T-cell coculture systems. *Nat Protoc,* 15**,** 15–39.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. & Schultz, N. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov,* 2**,** 401–4.

Ceraolo, M. G., Leccese, M., Cassotta, A., Triolo, S., Bombaci, M., Coluccio, E., Prati, D., Ungaro, R., Abrignani, S., Bandera, A., Sallusto, F., Lanzavecchia, A. & Notarbartolo, S. 2025. Dual Activation-Induced Marker Combinations Efficiently Identify and Discern Antigen-Specific and Bystander-Activated Human CD4(+) T Cells. *Eur J Immunol,* 55**,** e202451404.

Cercek, A., Lumish, M., Sinopoli, J., Weiss, J., Shia, J., Lamendola-Essel, M., El Dika, I. H., Segal, N., Shcherba, M., Sugarman, R., Stadler, Z., Yaeger, R., Smith, J. J., Rousseau, B., Argiles, G., Patel, M., Desai, A., Saltz, L. B., Widmar, M., Iyer, K., Zhang, J., Gianino, N., Crane, C., Romesser, P. B., Pappou, E. P., Paty, P., Garcia-Aguilar, J., Gonen, M., Gollub, M., Weiser, M. R., Schalper, K. A. & Diaz, L. A., Jr. 2022. PD-1 Blockade in Mismatch Repair-Deficient, Locally Advanced Rectal Cancer. *N Engl J Med,* 386**,** 2363–2376.

Chakravarthy, A., Khan, L., Bensler, N. P., Bose, P. & De Carvalho, D. D. 2018. TGF-beta-

associated extracellular matrix genes link cancer-associated fibroblasts to immune evasion and immunotherapy failure. *Nat Commun,* 9**,** 4692.

Chalabi, M., Fanchi, L. F., Dijkstra, K. K., Van Den Berg, J. G., Aalbers, A. G., Sikorska, K., Lopez-Yurda, M., Grootscholten, C., Beets, G. L., Snaebjornsson, P., Maas, M., Mertz, M., Veninga, V., Bounova, G., Broeks, A., Beets-Tan, R. G., De Wijkerslooth, T. R., Van Lent, A. U., Marsman, H. A., Nuijten, E., Kok, N. F., Kuiper, M., Verbeek, W. H., Kok, M., Van Leerdam, M. E., Schumacher, T. N., Voest, E. E. & Haanen, J. B. 2020. Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers. *Nat Med,* 26**,** 566–576.

Chan, D., Mandal, A., Zhou, Y., Collins, S., Owen, R., Bundred, J., Fuchs, H., James, S., Vendrell, I., Flannery, S., Fawkner-Corbett, D., Househam, J., Graham, T., Fischer, R., Simmons, A., Lu, X. & Buczacki, S. J. A. 2023. Mutational order and epistasis determine the consequences of FBXW7 mutations during colorectal cancer evolution. *bioRxiv*.

Chan, D. K. H., Collins, S. D. & Buczacki, S. J. A. 2023. Generation and immunofluorescent validation of gene knockouts in adult human colonic organoids using multi-guide RNA CRISPR-Cas9. *STAR Protoc,* 4**,** 101978.

Chan, D. K. H., Mandal, A., Hester, S., Yu, Z., Higgins, G. S., Kessler, B. M., Fischer, R. & Buczacki, S. J. A. 2023. Biallelic FBXW7 knockout induces AKAP8-mediated DNA damage in neighbouring wildtype cells. *Cell Death Discov,* 9**,** 200.

Chang, H. F., Schirra, C., Ninov, M., Hahn, U., Ravichandran, K., Krause, E., Becherer, U., Balint, S., Harkiolaki, M., Urlaub, H., Valitutti, S., Baldari, C. T., Dustin, M. L., Jahn, R. & Rettig, J. 2022. Identification of distinct cytotoxic granules as the origin of supramolecular attack particles in T lymphocytes. *Nat Commun,* 13**,** 1029.

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H. & Trajanoski, Z. 2017. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep,* 18**,** 248–262.

Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software,* 61**,** 1–36.

Chen, B., Scurrah, C. R., Mckinley, E. T., Simmons, A. J., Ramirez-Solano, M. A., Zhu, X., Markham, N. O., Heiser, C. N., Vega, P. N., Rolong, A., Kim, H., Sheng, Q., Drewes, J. L., Zhou, Y., Southard-Smith, A. N., Xu, Y., Ro, J., Jones, A. L., Revetta, F., Berry, L. D., Niitsu, H., Islam, M., Pelka, K., Hofree, M., Chen, J. H., Sarkizova, S., Ng, K., Giannakis, M., Boland, G. M., Aguirre, A. J., Anderson, A. C., Rozenblatt-Rosen, O., Regev, A., Hacohen, N., Kawasaki, K., Sato, T., Goettel, J. A., Grady, W. M., Zheng, W., Washington, M. K., Cai, Q., Sears, C. L., Goldenring, J. R., Franklin, J. L., Su, T., Huh, W. J., Vandekar, S., Roland, J. T., Liu, Q., Coffey, R. J., Shrubsole, M. J. & Lau, K. S. 2021. Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell,* 184**,** 6262–6280 e26.

Chen, D. S. & Mellman, I. 2013. Oncology meets immunology: the cancer-immunity cycle. *Immunity,* 39**,** 1–10.

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

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alfoldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Grant, R., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., Gabriel, S., Gupta, N., Ferriera, S., Tolonen, C., Novod, S., Bergelson, L., Roazen, D., Ruano-Rubio, V., Covarrubias, M., Llanwarne, C., Petrillo, N., Wade, G., Jeandet, T., Munshi, R., Tibbetts, K., Genome Aggregation Database, C., O'donnell-Luria, A., Solomonson, M., Seed, C., Martin, A. R., Talkowski, M. E., Rehm, H. L., Daly, M. J., Tiao, G., Neale, B. M., Macarthur, D. G. & Karczewski, K. J. 2024. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature,* 625**,** 92–100.

Chen, Y., Wang, D., Li, Y., Qi, L., Si, W., Bo, Y., Chen, X., Ye, Z., Fan, H., Liu, B., Liu, C., Zhang, L., Zhang, X., Li, Z., Zhu, L., Wu, A. & Zhang, Z. 2024. Spatiotemporal single-cell analysis decodes cellular dynamics underlying different responses to immunotherapy in colorectal cancer. *Cancer Cell,* 42**,** 1268–1285 e7.

Chen, Y., Zhang, B., Bao, L., Jin, L., Yang, M., Peng, Y., Kumar, A., Wang, J. E., Wang, C., Zou, X., Xing, C., Wang, Y. & Luo, W. 2018. ZMYND8 acetylation mediates HIF-dependent breast cancer progression and metastasis. *J Clin Invest,* 128**,** 1937–1955.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Zemgulyte, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P. & Avsec, Z. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science,* 381**,** eadg7492.

Cheng, J., Yan, J., Liu, Y., Shi, J., Wang, H., Zhou, H., Zhou, Y., Zhang, T., Zhao, L., Meng, X., Gong, H., Zhang, X., Zhu, H. & Jiang, P. 2023. Cancer-cell-derived fumarate suppresses the anti-tumor capacity of CD8(+) T cells in the tumor microenvironment. *Cell Metab,* 35**,** 961–978 e10.

Chudnovsky, Y., Kim, D., Zheng, S., Whyte, W. A., Bansal, M., Bray, M. A., Gopal, S., Theisen, M. A., Bilodeau, S., Thiru, P., Muffat, J., Yilmaz, O. H., Mitalipova, M., Woolard, K., Lee, J., Nishimura, R., Sakata, N., Fine, H. A., Carpenter, A. E., Silver, S. J., Verhaak, R. G., Califano, A., Young, R. A., Ligon, K. L., Mellinghoff, I. K., Root, D. E., Sabatini, D. M., Hahn, W. C. & Chheda, M. G. 2014. ZFHX4 interacts with the NuRD core member CHD4 and regulates the glioblastoma tumor-initiating cell state. *Cell Rep,* 6**,** 313–24.

Chung, C. R., Tang, Y., Chiu, Y. P., Li, S., Hsieh, W. K., Yao, L., Chiang, Y. C., Pang, Y., Chen, G. T., Chou, K. C., Paik, Y. S., Tran, P. L., Lin, C. P., Kao, Y. M., Chen, Y. J., Chang, W. C., Hsu, J. B., Horng, J. T. & Lee, T. Y. 2025. dbPTM 2025 update: comprehensive integration of PTMs and proteomic data for advanced insights into cancer research. *Nucleic Acids Res,* 53**,** D377–D386.

Clapier, C. R. & Cairns, B. R. 2009. The biology of chromatin remodeling complexes. *Annu Rev Biochem,* 78**,** 273–304.

Coelho, M. A., Cooper, S., Strauss, M. E., Karakoc, E., Bhosle, S., Goncalves, E., Picco, G., Burgold, T., Cattaneo, C. M., Veninga, V., Consonni, S., Dincer, C., Vieira, S. F., Gibson, F., Barthorpe, S., Hardy, C., Rein, J., Thomas, M., Marioni, J., Voest, E. E., Bassett, A. & Garnett, M. J. 2023. Base editing screens map mutations affecting interferon-gamma signaling in cancer. *Cancer Cell,* 41**,** 288–303 e6.

Collart, M. A. 2016. The Ccr4-Not complex is a key regulator of eukaryotic gene expression. *Wiley Interdiscip Rev RNA,* 7**,** 438–54.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A. & Zhang, F. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science,* 339**,** 819–23.

Consortium, I. T. P.-C. a. O. W. G. 2020. Pan-cancer analysis of whole genomes. *Nature,* 578**,** 82–93.

Coppola, D., Nebozhyn, M., Khalil, F., Dai, H., Yeatman, T., Loboda, A. & Mule, J. J. 2011. Unique ectopic lymph node-like structures present in human primary colorectal carcinoma are identified by immune gene array profiling. *Am J Pathol,* 179**,** 37–45.

Cortes-Ciriano, I., Lee, S., Park, W. Y., Kim, T. M. & Park, P. J. 2017. A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun,* 8**,** 15180.

Courau, T., Bonnereau, J., Chicoteau, J., Bottois, H., Remark, R., Assante Miranda, L., Toubert, A., Blery, M., Aparicio, T., Allez, M. & Le Bourhis, L. 2019. Cocultures of human colorectal tumor spheroids with immune cells reveal the therapeutic potential of MICA/B and NKG2A targeting for cancer treatment. *J Immunother Cancer,* 7**,** 74.

Cptac. 2025. *LinkedOMICS Website Portal (CPTAC)* [Online]. Available: https://linkedomics.org/cptac-colon/ [Accessed 23 June 2025].

Creton, M., Wagener, F., Massink, M., Fennis, W., Bloemen, M., Schols, J., Aarts, M., Van Der Molen, A. M., Van Haaften, G. & Van Den Boogaard, M. J. 2023. Concurrent de novo ZFHX4 variant and 16q24.1 deletion in a patient with orofacial clefting; a potential role of ZFHX4 and USP10. *Am J Med Genet A,* 191**,** 1083–1088.

Crockett, S. D. & Nagtegaal, I. D. 2019. Terminology, Molecular Features, Epidemiology, and Management of Serrated Colorectal Neoplasia. *Gastroenterology,* 157**,** 949–966 e4.

D'angelo, A., Shibata, K., Tokunaga, M., Furutani-Seiki, M. & Bagby, S. 2022. Generation of murine tumour-reactive T cells by co-culturing murine pancreatic cancer organoids and peripheral blood lymphocytes. *Biochem Biophys Rep,* 32**,** 101365.

De Palma, F. D. E., D'argenio, V., Pol, J., Kroemer, G., Maiuri, M. C. & Salvatore, F. 2019. The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer. *Cancers (Basel),* 11.

De Vries, N. L., Van Unen, V., Ijsselsteijn, M. E., Abdelaal, T., Van Der Breggen, R., Farina Sarasqueta, A., Mahfouz, A., Peeters, K., Hollt, T., Lelieveldt, B. P. F., Koning, F. & De Miranda, N. 2020. High-dimensional cytometric analysis of colorectal cancer reveals novel mediators of antitumour immunity. *Gut,* 69**,** 691–703.

Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. & Wallace, M. B. 2019. Colorectal cancer. *Lancet,* 394**,** 1467–1480.

Depmap Broad Institute. 2025. *DepMap Public 25Q2 Dataset* [Online]. Available: https://depmap.org/portal [Accessed].

Dersh, D., Yewdell, J. W. & Wei, J. 2019. A SIINFEKL-Based System to Measure MHC Class I Antigen Presentation Efficiency and Kinetics. *Methods Mol Biol,* 1988**,** 109–122.

Derynck, R., Turley, S. J. & Akhurst, R. J. 2021. TGFbeta biology in cancer progression

and immunotherapy. *Nat Rev Clin Oncol,* 18**,** 9–34.

Devito, N. C., Sturdivant, M., Thievanthiran, B., Xiao, C., Plebanek, M. P., Salama, A. K. S., Beasley, G. M., Holtzhausen, A., Novotny-Diermayr, V., Strickler, J. H. & Hanks, B. A. 2021. Pharmacological Wnt ligand inhibition overcomes key tumor-mediated resistance pathways to anti-PD-1 immunotherapy. *Cell Rep,* 35**,** 109071.

Dias, K. R., Carlston, C. M., Blok, L. E. R., De Hayr, L., Nawaz, U., Evans, C. A., Bayrak-Toydemir, P., Htun, S., Zhu, Y., Ma, A., Lynch, S. A., Moorwood, C., Stals, K., Ellard, S., Bainbridge, M. N., Friedman, J., Pappas, J. G., Rabin, R., Nowak, C. B., Douglas, J., Wilson, T. E., Guillen Sacoto, M. J., Mullegama, S. V., Palculict, T. B., Kirk, E. P., Pinner, J. R., Edwards, M., Montanari, F., Graziano, C., Pippucci, T., Dingmann, B., Glass, I., Mefford, H. C., Shimoji, T., Suzuki, T., Yamakawa, K., Streff, H., Schaaf, C. P., Slavotinek, A. M., Voineagu, I., Carey, J. C., Buckley, M. F., Schenck, A., Harvey, R. J. & Roscioli, T. 2022. De Novo ZMYND8 variants result in an autosomal dominant neurodevelopmental disorder with cardiac malformations. *Genet Med,* 24**,** 1952–1966.

Diaz, L. A., Jr., Shiu, K. K., Kim, T. W., Jensen, B. V., Jensen, L. H., Punt, C., Smith, D., Garcia-Carbonero, R., Benavides, M., Gibbs, P., De La Fourchardiere, C., Rivera, F., Elez, E., Le, D. T., Yoshino, T., Zhong, W. Y., Fogelman, D., Marinello, P., Andre, T. & Investigators, K.-. 2022. Pembrolizumab versus chemotherapy for microsatellite instability-high or mismatch repair-deficient metastatic colorectal cancer (KEYNOTE-177): final analysis of a randomised, open-label, phase 3 study. *Lancet Oncol,* 23**,** 659–670.

Diaz-Gay, M., Dos Santos, W., Moody, S., Kazachkova, M., Abbasi, A., Steele, C. D., Vangara, R., Senkin, S., Wang, J., Fitzgerald, S., Bergstrom, E. N., Khandekar, A., Otlu, B., Abedi-Ardekani, B., De Carvalho, A. C., Cattiaux, T., Penha, R. C. C., Gaborieau, V., Chopard, P., Carreira, C., Cheema, S., Latimer, C., Teague, J. W., Mukeriya, A., Zaridze, D., Cox, R., Albert, M., Phouthavongsy, L., Gallinger, S., Malekzadeh, R., Niavarani, A., Miladinov, M., Eric, K., Milosavljevic, S., Sangrajrang, S., Curado, M. P., Aguiar, S., Reis, R. M., Reis, M. T., Romagnolo, L. G., Guimaraes, D. P., Holcatova, I., Kalvach, J., Vaccaro, C. A., Pinero, T. A., Swiatkowska, B., Lissowska, J., Roszkowska-Purska, K., Huertas-Salgado, A., Shibata, T., Shiba, S., Sangkhathat, S., Chitapanarux, T., Roshandel, G., Ashton-Prolla, P., Damin, D. C., De Oliveira, F. H., Humphreys, L., Lawley, T. D., Perdomo, S., Stratton, M. R., Brennan, P. & Alexandrov, L. B. 2025. Geographic and age variations in mutational processes in colorectal cancer. *Nature,* 643**,** 230–240.

Dijkstra, K. K., Cattaneo, C. M., Weeber, F., Chalabi, M., Van De Haar, J., Fanchi, L. F., Slagter, M., Van Der Velden, D. L., Kaing, S., Kelderman, S., Van Rooij, N., Van Leerdam, M. E., Depla, A., Smit, E. F., Hartemink, K. J., De Groot, R., Wolkers, M. C., Sachs, N., Snaebjornsson, P., Monkhorst, K., Haanen, J., Clevers, H., Schumacher, T. N. & Voest, E. E. 2018. Generation of Tumor-Reactive T Cells by Co-culture of Peripheral Blood Lymphocytes and Tumor Organoids. *Cell,* 174**,** 1586–1598 e12.

Ding, J., Zhang, Z. M., Xia, Y., Liao, G. Q., Pan, Y., Liu, S., Zhang, Y. & Yan, Z. S. 2013. LSD1-mediated epigenetic modification contributes to proliferation and metastasis of colon cancer. *Br J Cancer,* 109**,** 994–1003.

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D. L., Weerasinghe, A., Huang, K. L., Tokheim, C., Cortes-Ciriano, I., Jayasinghe, R., Chen, F., Yu, L., Sun, S., Olsen, C., Kim, J., Taylor, A. M., Cherniack, A. D., Akbani, R., Suphavilai, C., Nagarajan, N., Stuart, J. M., Mills, G. B., Wyczalkowski, M. A., Vincent, B. G., Hutter, C. M., Zenklusen, J. C., Hoadley, K. A., Wendl, M. C.,

Shmulevich, L., Lazar, A. J., Wheeler, D. A., Getz, G. & Cancer Genome Atlas Research, N. 2018. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell,* 173**,** 305–320 e10.

Dongre, A. & Weinberg, R. A. 2019. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat Rev Mol Cell Biol,* 20**,** 69–84.

Dostert, C., Grusdat, M., Letellier, E. & Brenner, D. 2019. The TNF Family of Ligands and Receptors: Communication Modules in the Immune System and Beyond. *Physiol Rev,* 99**,** 115–160.

Drost, J., Van Jaarsveld, R. H., Ponsioen, B., Zimberlin, C., Van Boxtel, R., Buijs, A., Sachs, N., Overmeer, R. M., Offerhaus, G. J., Begthel, H., Korving, J., Van De Wetering, M., Schwank, G., Logtenberg, M., Cuppen, E., Snippert, H. J., Medema, J. P., Kops, G. J. & Clevers, H. 2015. Sequential cancer mutations in cultured human intestinal stem cells. *Nature,* 521**,** 43–7.

Du, L., Ning, Z., Zhang, H. & Liu, F. 2017. Corepressor metastasis-associated protein 3 modulates epithelial-to-mesenchymal transition and metastasis. *Chin J Cancer,* 36**,** 28.

Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. 2002. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol,* 3**,** 991–8.

Dunn, G. P., Koebel, C. M. & Schreiber, R. D. 2006. Interferons, immunity and cancer immunoediting. *Nat Rev Immunol,* 6**,** 836–48.

Dvorak, H. F. 1986. Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. *N Engl J Med,* 315**,** 1650–9.

Dyer, S. C., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Barrera-Enriquez, V. P., Becker, A., Bennett, R., Beracochea, M., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Cortes, L. A., Davidson, C., Denni, S., Dodiya, K., Donaldson, S., El Houdaigui, B., El Naboulsi, T., Falola, O., Fatima, R., Genez, T., Martinez, J. G., Gurbich, T., Hardy, M., Hollis, Z., Hunt, T., Kay, M., Kaykala, V., Lemos, D., Lodha, D., Mathlouthi, N., Merino, G. A., Merritt, R., Mirabueno, L. P., Mushtaq, A., Hossain, S. N., Perez-Silva, J. G., Perry, M., Pilizota, I., Poppleton, D., Prosovetskaia, I., Raj, S., Salam, A. I. A., Saraf, S., Saraiva-Agostinho, N., Sinha, S., Sipos, B., Sitnik, V., Steed, E., Suner, M. M., Surapaneni, L., Sutinen, K., Tricomi, F. F., Tsang, I., Urbina-Gomez, D., Veidenberg, A., Walsh, T. A., Willhoft, N. L., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Cheema, J., Da Rocha, J. B., De Silva, N. H., Giorgetti, S., Haggerty, L., Ilsley, G. R., Keatley, J., Loveland, J. E., Moore, B., Mudge, J. M., Naamati, G., Tate, J., Trevanion, S. J., Winterbottom, A., Flint, B., Frankish, A., Hunt, S. E., Finn, R. D., Freeberg, M. A., Harrison, P. W., Martin, F. J. & Yates, A. D. 2025. Ensembl 2025. *Nucleic Acids Res,* 53**,** D948–D957.

Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. 2017. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci Rep,* 7**,** 16618.

Eising, E., Carrion-Castillo, A., Vino, A., Strand, E. A., Jakielski, K. J., Scerri, T. S., Hildebrand, M. S., Webster, R., Ma, A., Mazoyer, B., Francks, C., Bahlo, M., Scheffer, I. E., Morgan, A. T., Shriberg, L. D. & Fisher, S. E. 2019. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol Psychiatry,* 24**,** 1065–1078.

Elmentaite, R., Kumasaka, N., Roberts, K., Fleming, A., Dann, E., King, H. W., Kleshchevnikov, V., Dabrowska, M., Pritchard, S., Bolt, L., Vieira, S. F., Mamanova, L., Huang, N., Perrone, F., Goh Kai'en, I., Lisgo, S. N., Katan, M., Leonard, S., Oliver, T. R. W., Hook, C. E., Nayak, K., Campos, L. S., Dominguez Conde, C., Stephenson, E., Engelbert, J., Botting, R. A., Polanski, K., Van Dongen, S., Patel, M., Morgan, M. D., Marioni, J. C., Bayraktar, O. A., Meyer, K. B., He, X., Barker, R. A., Uhlig, H. H., Mahbubani, K. T., Saeb-Parsy, K., Zilbauer, M., Clatworthy, M. R., Haniffa, M., James, K. R. & Teichmann, S. A. 2021. Cells of the human intestinal tract mapped across space and time. *Nature,* 597**,** 250–255.

Eng, C., Kim, T. W., Bendell, J., Argiles, G., Tebbutt, N. C., Di Bartolomeo, M., Falcone, A., Fakih, M., Kozloff, M., Segal, N. H., Sobrero, A., Yan, Y., Chang, I., Uyei, A., Roberts, L., Ciardiello, F. & Investigators, I. M. 2019. Atezolizumab with or without cobimetinib versus regorafenib in previously treated metastatic colorectal cancer (IMblaze370): a multicentre, open-label, phase 3, randomised, controlled trial. *Lancet Oncol,* 20**,** 849–861.

Estro. 2025. *New Study Shows Immunotherapy and Radiotherapy Could Help More Rectal Cancer Patients Avoid Surgery* [Online]. ESTRO. Available: https://www.estro.org/ESTRO/media/ESTRO/Congresses/1-3-PRIME-RT-press-release-ESTRO-2025.pdf [Accessed 7 Oct 2025].

Fakih, M., Ouyang, C., Wang, C., Tu, T. Y., Gozo, M. C., Cho, M., Sy, M., Longmate, J. A. & Lee, P. P. 2019. Immune overdrive signature in colorectal tumor subset predicts poor clinical outcome. *J Clin Invest,* 129**,** 4464–4476.

Fakih, M., Raghav, K. P. S., Chang, D. Z., Larson, T., Cohn, A. L., Huyck, T. K., Cosgrove, D., Fiorillo, J. A., Tam, R., D'adamo, D., Sharma, N., Brennan, B. J., Wang, Y. A., Coppieters, S., Zebger-Gong, H., Weispfenning, A., Seidel, H., Ploeger, B. A., Mueller, U., Oliveira, C. S. V. & Paulson, A. S. 2023. Regorafenib plus nivolumab in patients with mismatch repair-proficient/microsatellite stable metastatic colorectal cancer: a single-arm, open-label, multicentre phase 2 study. *EClinicalMedicine,* 58**,** 101917.

Fearon, E. R. & Vogelstein, B. 1990. A genetic model for colorectal tumorigenesis. *Cell,* 61**,** 759–67.

Feng, F., Sun, H., Zhao, Z., Sun, C., Zhao, Y., Lin, H., Yang, J., Xiao, Y., Wang, W. & Wu, D. 2022. Identification of APC Mutation as a Potential Predictor for Immunotherapy in Colorectal Cancer. *J Oncol,* 2022**,** 6567998.

Fevre-Montange, M., Champier, J., Szathmari, A., Wierinckx, A., Mottolese, C., Guyotat, J., Figarella-Branger, D., Jouvet, A. & Lachuer, J. 2006. Microarray analysis reveals differential gene expression patterns in tumors of the pineal region. *J Neuropathol Exp Neurol,* 65**,** 675–84.

Filippova, G. N., Qi, C. F., Ulmer, J. E., Moore, J. M., Ward, M. D., Hu, Y. J., Loukinov, D. I., Pugacheva, E. M., Klenova, E. M., Grundy, P. E., Feinberg, A. P., Cleton-Jansen, A. M., Moerland, E. W., Cornelisse, C. J., Suzuki, H., Komiya, A., Lindblom, A., Dorion-Bonnet, F., Neiman, P. E., Morse, H. C., 3rd, Collins, S. J. & Lobanenkov, V. V. 2002. Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity. *Cancer Res,* 62**,** 48–52.

Flower, J. 2025. Thesis GitHub Repository. Available at: https://github.com/tufty23/Thesis_2025.

Fontana, P., Ginevrino, M., Bejo, K., Cantalupo, G., Ciavarella, M., Lombardi, C., Maioli, M., Scarano, F., Costabile, C., Novelli, A. & Lonardo, F. 2021. A ZFHX4 mutation associated with a recognizable neuropsychological and facial phenotype. *Eur J Med Genet,* 64**,** 104321.

Frankish, A., Carbonell-Sala, S., Diekhans, M., Jungreis, I., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Arnan, C., Barnes, I., Banerjee, A., Bennett, R., Berry, A., Bignell, A., Boix, C., Calvet, F., Cerdan-Velez, D., Cunningham, F., Davidson, C., Donaldson, S., Dursun, C., Fatima, R., Giorgetti, S., Giron, C. G., Gonzalez, J. M., Hardy, M., Harrison, P. W., Hourlier, T., Hollis, Z., Hunt, T., James, B., Jiang, Y., Johnson, R., Kay, M., Lagarde, J., Martin, F. J., Gomez, L. M., Nair, S., Ni, P., Pozo, F., Ramalingam, V., Ruffier, M., Schmitt, B. M., Schreiber, J. M., Steed, E., Suner, M. M., Sumathipala, D., Sycheva, I., Uszczynska-Ratajczak, B., Wass, E., Yang, Y. T., Yates, A., Zafrulla, Z., Choudhary, J. S., Gerstein, M., Guigo, R., Hubbard, T. J. P., Kellis, M., Kundaje, A., Paten, B., Tress, M. L. & Flicek, P. 2023. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res,* 51**,** D942–D949.

Fu, C., An, N., Liu, J., A, J., Zhang, B., Liu, M., Zhang, Z., Fu, L., Tian, X., Wang, D. & Dong, J. T. 2020. The transcription factor ZFHX3 is crucial for the angiogenic function of hypoxia-inducible factor 1alpha in liver cancer cells. *J Biol Chem,* 295**,** 7060–7074.

Fu, C., Gu, H., Sun, L., Wang, Z., Zhang, Q., Luo, N., Chen, D. & Zhou, T. 2024. Predictive value of ZFHX4 mutation for the efficacy of immune checkpoint inhibitors in non-small cell lung cancer and melanoma. *Invest New Drugs,* 42**,** 623–634.

Fu, J., Qin, L., He, T., Qin, J., Hong, J., Wong, J., Liao, L. & Xu, J. 2011. The TWIST/Mi2/NuRD protein complex and its essential role in cancer metastasis. *Cell Res,* 21**,** 275–89.

Fujii, M., Shimokawa, M., Date, S., Takano, A., Matano, M., Nanki, K., Ohta, Y., Toshimitsu, K., Nakazato, Y., Kawasaki, K., Uraoka, T., Watanabe, T., Kanai, T. & Sato, T. 2016. A Colorectal Tumor Organoid Library Demonstrates Progressive Loss of Niche Factor Requirements during Tumorigenesis. *Cell Stem Cell,* 18**,** 827–838.

Fukuoka, S., Hara, H., Takahashi, N., Kojima, T., Kawazoe, A., Asayama, M., Yoshii, T., Kotani, D., Tamura, H., Mikamoto, Y., Hirano, N., Wakabayashi, M., Nomura, S., Sato, A., Kuwata, T., Togashi, Y., Nishikawa, H. & Shitara, K. 2020. Regorafenib Plus Nivolumab in Patients With Advanced Gastric or Colorectal Cancer: An Open-Label, Dose-Escalation, and Dose-Expansion Phase Ib Trial (REGONIVO, EPOC1603). *J Clin Oncol,* 38**,** 2053–2061.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. & Stratton, M. R. 2004. A census of human cancer genes. *Nat Rev Cancer,* 4**,** 177–83.

Gajewski, T. F. 2015. The Next Hurdle in Cancer Immunotherapy: Overcoming the Non-T-Cell-Inflamed Tumor Microenvironment. *Semin Oncol,* 42**,** 663–71.

Galon, J., Angell, H. K., Bedognetti, D. & Marincola, F. M. 2013. The continuum of cancer immunosurveillance: prognostic, predictive, and mechanistic signatures. *Immunity,* 39**,** 11–26.

Galon, J. & Bruni, D. 2019. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat Rev Drug Discov,* 18**,** 197–218.

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pages, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoue, F., Bruneval, P., Cugnenc, P. H., Trajanoski, Z., Fridman, W. H. & Pages, F. 2006. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science,* 313**,** 1960–4.

Galon, J., Pages, F., Marincola, F. M., Angell, H. K., Thurin, M., Lugli, A., Zlobec, I., Berger, A., Bifulco, C., Botti, G., Tatangelo, F., Britten, C. M., Kreiter, S., Chouchane, L., Delrio, P., Arndt, H., Asslaber, M., Maio, M., Masucci, G. V., Mihm, M., Vidal-Vanaclocha, F., Allison, J. P., Gnjatic, S., Hakansson, L., Huber, C., Singh-Jasuja, H., Ottensmeier, C., Zwierzina, H., Laghi, L., Grizzi, F., Ohashi, P. S., Shaw, P. A., Clarke, B. A., Wouters, B. G., Kawakami, Y., Hazama, S., Okuno, K., Wang, E., O'donnell-Tormey, J., Lagorce, C., Pawelec, G., Nishimura, M. I., Hawkins, R., Lapointe, R., Lundqvist, A., Khleif, S. N., Ogino, S., Gibbs, P., Waring, P., Sato, N., Torigoe, T., Itoh, K., Patel, P. S., Shukla, S. N., Palmqvist, R., Nagtegaal, I. D., Wang, Y., D'arrigo, C., Kopetz, S., Sinicrope, F. A., Trinchieri, G., Gajewski, T. F., Ascierto, P. A. & Fox, B. A. 2012. Cancer classification using the Immunoscore: a worldwide task force. *J Transl Med,* 10**,** 205.

Ganesh, K., Stadler, Z. K., Cercek, A., Mendelsohn, R. B., Shia, J., Segal, N. H. & Diaz, L. A., Jr. 2019. Immunotherapy in colorectal cancer: rationale, challenges and potential. *Nat Rev Gastroenterol Hepatol,* 16**,** 361–375.

Gao, J., Liu, R., Feng, D., Huang, W., Huo, M., Zhang, J., Leng, S., Yang, Y., Yang, T., Yin, X., Teng, X., Yu, H., Yuan, B. & Wang, Y. 2021. Snail/PRMT5/NuRD complex contributes to DNA hypermethylation in cervical cancer by TET1 inhibition. *Cell Death Differ,* 28**,** 2818–2836.

Gao, L., Li, Y., Wang, H., Liu, J., Zhang, R., Shan, W., Zeng, L., Zhao, Q., Li, Y. & Liu, J. 2025. SET facilitates immune escape of microsatellite stability colorectal cancer by inhibiting c-Myc degradation. *Cancer Sci,* 116**,** 29–43.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I. & Liu, D. R. 2017. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature,* 551**,** 464–471.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., Mitchell, T. J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vazquez-Garcia, I., Haase, K., Jerman, L., Sengupta, S., Macintyre, G., Malikic, S., Donmez, N., Livitz, D. G., Cmero, M., Demeulemeester, J., Schumacher, S., Fan, Y., Yao, X., Lee, J., Schlesner, M., Boutros, P. C., Bowtell, D. D., Zhu, H., Getz, G., Imielinski, M., Beroukhim, R., Sahinalp, S. C., Ji, Y., Peifer, M., Markowetz, F., Mustonen, V., Yuan, K., Wang, W., Morris, Q. D., Evolution, P., Heterogeneity Working, G., Spellman, P. T., Wedge, D. C., Van Loo, P. & Consortium, P. 2020. The evolutionary history of 2,658 cancers. *Nature,* 578**,** 122–128.

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M. G., Jadersten, M., Dolatshad, H., Verma, A., Cross, N. C., Vyas, P., Killick, S., Hellstrom-Lindberg, E., Cazzola, M., Papaemmanuil, E., Campbell, P. J. & Boultwood, J. 2015. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun,* 6**,** 5901.

Geurts, M. H., De Poel, E., Pleguezuelos-Manzano, C., Oka, R., Carrillo, L., Andersson-Rolf, A., Boretto, M., Brunsveld, J. E., Van Boxtel, R., Beekman, J. M. & Clevers, H. 2021. Evaluating CRISPR-based prime editing for cancer modeling and CFTR repair in organoids. *Life Sci Alliance,* 4.

Geurts, M. H., Gandhi, S., Boretto, M. G., Akkerman, N., Derks, L. L. M., Van Son, G., Celotti, M., Harshuk-Shabso, S., Peci, F., Begthel, H., Hendriks, D., Schurmann, P., Andersson-Rolf, A., Chuva De Sousa Lopes, S. M., Van Es, J. H., Van Boxtel, R. & Clevers, H. 2023. One-step generation of tumor models by base editor multiplexing in adult stem cell-derived organoids. *Nat Commun,* 14**,** 4998.

Giannakis, M., Mu, X. J., Shukla, S. A., Qian, Z. R., Cohen, O., Nishihara, R., Bahl, S., Cao, Y., Amin-Mansour, A., Yamauchi, M., Sukawa, Y., Stewart, C., Rosenberg, M., Mima, K., Inamura, K., Nosho, K., Nowak, J. A., Lawrence, M. S., Giovannucci, E. L., Chan, A. T., Ng, K., Meyerhardt, J. A., Van Allen, E. M., Getz, G., Gabriel, S. B., Lander, E. S., Wu, C. J., Fuchs, C. S., Ogino, S. & Garraway, L. A. 2016. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep,* 17**,** 1206.

Gillinder, K. R., Ilsley, M. D., Nebor, D., Sachidanandam, R., Lajoie, M., Magor, G. W., Tallack, M. R., Bailey, T., Landsberg, M. J., Mackay, J. P., Parker, M. W., Miles, L. A., Graber, J. H., Peters, L. L., Bieker, J. J. & Perkins, A. C. 2017. Promiscuous DNA-binding of a mutant zinc finger protein corrupts the transcriptome and diminishes cell viability. *Nucleic Acids Res,* 45**,** 1130–1143.

Gillotin, S. 2018. Isolation of Chromatin-bound Proteins from Subcellular Fractions for Biochemical Analysis. *Bio Protoc,* 8**,** e3035.

Goel, H. & O'donnell, S. 2025. A de-novo frameshift variant in ZFHX4 associated with a recognisable neurodevelopmental disorder: a case report. *Clin Dysmorphol*.

Goldman, M. J., Craft, B., Hastie, M., Repecka, K., Mcdade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J. & Haussler, D. 2020. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol,* 38**,** 675–678.

Gong, X. & Karchin, R. 2023. Clustering by antigen-presenting genes reveals immune landscapes and predicts response to checkpoint immunotherapy. *Sci Rep,* 13**,** 950.

Gonzalez-Montero, J., Rojas, C. I. & Burotto, M. 2024. Predictors of response to immunotherapy in colorectal cancer. *Oncologist,* 29**,** 824–832.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A. & Lopez-Bigas, N. 2013. IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods,* 10**,** 1081–2.

Gordon-Weeks, A. & Yuzhalin, A. E. 2020. Cancer Extracellular Matrix Proteins Regulate Tumour Immunity. *Cancers (Basel),* 12.

Goto, N., Westcott, P. M. K., Goto, S., Imada, S., Taylor, M. S., Eng, G., Braverman, J., Deshpande, V., Jacks, T., Agudo, J. & Yilmaz, O. H. 2024. SOX17 enables immune evasion of early colorectal adenomas and cancers. *Nature,* 627**,** 636–645.

Gprofiler. 2025. *gProfiler Online Tool: version e113_eg59_p19_f6a03c19* [Online]. Available: https://biit.cs.ut.ee/gprofiler/gost [Accessed 20 July 2025].

Grady, W. M. & Carethers, J. M. 2008. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology,* 135**,** 1079–99.

Graziani, V., Rodriguez-Hernandez, I., Maiques, O. & Sanz-Moreno, V. 2022. The amoeboid state as part of the epithelial-to-mesenchymal transition programme. *Trends Cell Biol,* 32**,** 228–242.

Greaves, M. 2015. Evolutionary determinants of cancer. *Cancer Discov,* 5**,** 806–20.

Greene, M. W., Abraham, P. T., Kuhlers, P. C., Lipke, E. A., Heslin, M. J., Wijaya, S. T. & Odeniyi, I. 2022. Consensus molecular subtype differences linking colon adenocarcinoma and obesity revealed by a cohort transcriptomic analysis. *PLoS One,* 17**,** e0268436.

Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., Albertsen, H., Joslyn, G., Stevens, J., Spirio, L., Robertson, M. & Et Al. 1991. Identification and characterization of the familial adenomatous polyposis coli gene. *Cell,* 66**,** 589–600.

Gtex Consortium 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet,* 45**,** 580–5.

Gtex Consortium. 2025. *GTEx Portal* [Online]. Available: https://www.gtexportal.org/home/ [Accessed 15 Aug 2025].

Gu, Z. 2022. Complex heatmap visualization. *Imeta,* 1**,** e43.

Guinney, J., Dienstmann, R., Wang, X., De Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., De Sousa, E. M. F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G. C., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J. W., Hanahan, D., Tabernero, J., Bernards, R., Friend, S. H., Laurent-Puig, P., Medema, J. P., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L. & Tejpar, S. 2015. The consensus molecular subtypes of colorectal cancer. *Nat Med,* 21**,** 1350–6.

Gunther, K., Rust, M., Leers, J., Boettger, T., Scharfe, M., Jarek, M., Bartkuhn, M. & Renkawitz, R. 2013. Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences. *Nucleic Acids Res,* 41**,** 3010–21.

Gurjao, C., Liu, D., Hofree, M., Aldubayan, S. H., Wakiro, I., Su, M. J., Felt, K., Gjini, E., Brais, L. K., Rotem, A., Rosenthal, M. H., Rozenblatt-Rosen, O., Rodig, S., Ng, K., Van Allen, E. M., Corsello, S. M., Ogino, S., Regev, A., Nowak, J. A. & Giannakis, M. 2019. Intrinsic Resistance to Immune Checkpoint Blockade in a Mismatch Repair-Deficient Colorectal Cancer. *Cancer Immunol Res,* 7**,** 1230–1236.

Ha, M., Kim, J., Park, S. M., Hong, C. M., Han, M. E., Song, P., Kang, C. D., Lee, D., Kim, Y. H., Hur, J. & Oh, S. O. 2020. Prognostic Role of Zinc Finger Homeobox 4 in Ovarian Serous Cystadenocarcinoma. *Genet Test Mol Biomarkers,* 24**,** 145–149.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., Zhou, Y., Ye, F., Jiang, M., Wu, J., Xiao, Y., Jia, X., Zhang, T., Ma, X., Zhang, Q., Bai, X., Lai, S., Yu, C., Zhu, L., Lin, R., Gao, Y., Wang, M., Wu, Y., Zhang, J., Zhan, R., Zhu, S., Hu, H., Wang, C., Chen, M., Huang, H., Liang, T., Chen, J., Wang, W., Zhang, D. & Guo, G. 2020. Construction of a human cell landscape at single-cell level. *Nature,* 581**,** 303–309.

Hanahan, D. 2022. Hallmarks of Cancer: New Dimensions. *Cancer Discov,* 12**,** 31–46.

Hanahan, D. & Weinberg, R. A. 2000. The hallmarks of cancer. *Cell,* 100**,** 57–70.

Hanahan, D. & Weinberg, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646–74.

Hanna, C. R., O'cathail, S. M., Graham, J. S., Saunders, M., Samuel, L., Harrison, M., Devlin, L., Edwards, J., Gaya, D. R., Kelly, C. A., Lewsley, L. A., Maka, N., Morrison, P., Dinnett, L., Dillon, S., Gourlay, J., Platt, J. J., Thomson, F., Adams, R. A. & Roxburgh, C. S. D. 2021. Durvalumab (MEDI 4736) in combination with extended neoadjuvant regimens in rectal cancer: a study protocol of a randomised phase II trial (PRIME-RT). *Radiat Oncol,* 16**,** 163.

Harter, M. F., Recaldin, T., Gerard, R., Avignon, B., Bollen, Y., Esposito, C., Guja-Jarosz, K., Kromer, K., Filip, A., Aubert, J., Schneider, A., Bacac, M., Bscheider, M., Stokar-Regenscheit, N., Piscuoglio, S., Beumer, J. & Gjorevski, N. 2024. Analysis of off-tumour toxicities of T-cell-engaging bispecific antibodies via donor-matched intestinal organoids and tumouroids. *Nat Biomed Eng,* 8**,** 345–360.

Hegazy, M., Cohen-Barak, E., Koetsier, J. L., Najor, N. A., Arvanitis, C., Sprecher, E., Green, K. J. & Godsel, L. M. 2020. Proximity Ligation Assay for Detecting Protein-Protein Interactions and Protein Modifications in Cells and Tissues in Situ. *Curr Protoc Cell Biol,* 89**,** e115.

Hegde, P. S. & Chen, D. S. 2020. Top 10 Challenges in Cancer Immunotherapy. *Immunity,* 52**,** 17–35.

Hegde, P. S., Karanikas, V. & Evers, S. 2016. The Where, the When, and the How of Immune Monitoring for Cancer Immunotherapies in the Era of Checkpoint Inhibition. *Clin Cancer Res,* 22**,** 1865–74.

Hellman, L. M. & Fried, M. G. 2007. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc,* 2**,** 1849–61.

Hemmi, K., Ma, D., Miura, Y., Kawaguchi, M., Sasahara, M., Hashimoto-Tamaoki, T., Tamaoki, T., Sakata, N. & Tsuchiya, K. 2006. A homeodomain-zinc finger protein, ZFHX4, is expressed in neuronal differentiation manner and suppressed in muscle differentiation manner. *Biol Pharm Bull,* 29**,** 1830–5.

Hess, G. T., Tycko, J., Yao, D. & Bassik, M. C. 2017. Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes. *Mol Cell,* 68**,** 26–43.

Hessey, S., Angelova, M., Colliver, E., Enfield, K. S., Haase, K., Leung, M. M., Richard, C., Bentham, R., Puttick, C., Pich, O., Liu, W. K., Magness, A., Bunkum, A., Grigoriadis, K., Huebner, A., Moore, D. A., Sivakumar, M., Salgado, R., Hobson, P. S., Levi, D., Ward, S., Veeriah, S., Naceur-Lombardelli, C., Rowan, A., Hiley, C. T., Zaccaria, S., Mcgranahan, N., Swanton, C. & Jamal-Hanjani, M. 2025. *The evolution of immunity and immune evasion from early- to end-stage lung cancer* [Online]. Cancer Research: AACR. Available: https://doi.org/10.1158/1538-7445.AM2025-1207 [Accessed 6th Oct 2025].

Hewish, M., Lord, C. J., Martin, S. A., Cunningham, D. & Ashworth, A. 2010. Mismatch repair deficient colorectal cancer in the era of personalized treatment. *Nat Rev Clin Oncol,* 7**,** 197–208.

Hgnc. 2025. *Gene group: Histocompatibility complex (HLA)* [Online]. Available: https://www.genenames.org/data/genegroup/#!/group/588 [Accessed 22 July 2025].

Hgnc. 2025. *Gene group: Complement system* [Online]. Available: https://www.genenames.org/data/genegroup/#!/group/492 [Accessed 20 July 2025].

Hissong, E., Jafari, M. D., Khan, S., Kasi, P., Guniganti, P., Ocean, A. J., Siolas, D., Yeo, H., Khan, U., Pigazzi, A., Lutfi, A., Chen, Z., Shah, M. A. & Hidalgo, M. 2025. *Neoadjuvant botensilimab (BOT) plus balstilimab (BAL) in resectable mismatch repair proficient (pMMR) and deficient (dMMR) colorectal cancer (CRC): NEST clinical trial update.* [Online]. ASCO Gastrointestinal Cancers Symposium: Journal of Clinical Oncology. Available: https://ascopubs.org/doi/10.1200/JCO.2025.43.4_suppl.207 [Accessed 6th Oct 2025].

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., Akbani, R., Bowlby, R., Wong, C. K., Wiznerowicz, M., Sanchez-Vega, F., Robertson, A. G., Schneider, B. G., Lawrence, M. S., Noushmehr, H., Malta, T. M., Cancer Genome Atlas, N., Stuart, J. M., Benz, C. C. & Laird, P. W. 2018. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell,* 173**,** 291–304 e6.

Hodi, F. S., O'day, S. J., Mcdermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C., Akerley, W., Van Den Eertwegh, A. J., Lutzky, J., Lorigan, P., Vaubel, J. M., Linette, G. P., Hogg, D., Ottensmeier, C. H., Lebbe, C., Peschel, C., Quirt, I., Clark, J. I., Wolchok, J. D., Weber, J. S., Tian, J., Yellin, M. J., Nichol, G. M., Hoos, A. & Urba, W. J. 2010. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med,* 363**,** 711–23.

Holland, P. W. 2013. Evolution of homeobox genes. *Wiley Interdiscip Rev Dev Biol,* 2**,** 31–45.

Holland, P. W., Booth, H. A. & Bruford, E. A. 2007. Classification and nomenclature of all human homeobox genes. *BMC Biol,* 5**,** 47.

Hong, W., Nakazawa, M., Chen, Y. Y., Kori, R., Vakoc, C. R., Rakowski, C. & Blobel, G. A. 2005. FOG-1 recruits the NuRD repressor complex to mediate transcriptional repression by GATA-1. *EMBO J,* 24**,** 2367–78.

Hoshida, Y. 2010. Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment. *PLoS One,* 5**,** e15543.

Hoshida, Y., Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. 2007. Subclass mapping: identifying common subtypes in independent disease data sets. *PLoS One,* 2**,** e1195.

Hoyos, D., Zappasodi, R., Schulze, I., Sethna, Z., De Andrade, K. C., Bajorin, D. F., Bandlamudi, C., Callahan, M. K., Funt, S. A., Hadrup, S. R., Holm, J. S., Rosenberg, J. E., Shah, S. P., Vazquez-Garcia, I., Weigelt, B., Wu, M., Zamarin, D., Campitelli, L. F., Osborne, E. J., Klinger, M., Robins, H. S., Khincha, P. P., Savage, S. A., Balachandran, V. P., Wolchok, J. D., Hellmann, M. D., Merghoub, T., Levine, A. J., Luksza, M. & Greenbaum, B. D. 2022. Fundamental immune-oncogenicity trade-offs define driver mutation fitness. *Nature,* 606**,** 172–179.

Hu, Q., Yin, J., Zhao, S., Wang, Y., Shi, R., Yan, K. & Huang, S. 2024. ZFHX3 acts as a tumor suppressor in prostate cancer by targeting FTO-mediated m(6)A demethylation. *Cell Death Discov,* 10**,** 284.

Hu, Y. & Smyth, G. K. 2009. ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J Immunol Methods,* 347**,** 70–8.

Hua, L., Nie, C., Lin, J., An, A. X., Bourre, L. & Wang, J. 2023. *STING agonist and TLR7/8 agonist overcome anti-PD-1 resistance mediated by loss of β2M* [Online]. Cancer Res. Available: https://doi.org/10.1158/1538-7445.AM2023-5118 [Accessed].

Huang, Z., Li, S., Song, W., Li, X., Li, Q., Zhang, Z., Han, Y., Zhang, X., Miao, S., Du, R. & Wang, L. 2013. Lysine-specific demethylase 1 (LSD1/KDM1A) contributes to colorectal tumorigenesis via activation of the Wnt/beta-catenin pathway by down-regulating Dickkopf-1 (DKK1) [corrected]. *PLoS One,* 8**,** e70077.

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., Seja, E., Lomeli, S., Kong, X., Kelley, M. C., Sosman, J. A., Johnson, D. B., Ribas, A. & Lo, R. S. 2016. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell,* 165**,** 35–44.

Ilyas, M. I. M. 2024. Epidemiology of Stage IV Colorectal Cancer: Trends in the Incidence, Prevalence, Age Distribution, and Impact on Life Span. *Clin Colon Rectal Surg,* 37**,** 57–61.

International Cancer Genome, C., Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T. S., Remacle, J., Schafer, A. J., Shibata, T., Stratton, M. R., Vockley, J. G., Watanabe, K., Yang, H., Yuen, M. M., Knoppers, B. M., Bobrow, M., Cambon-Thomsen, A., Dressler, L. G., Dyke, S. O., Joly, Y., Kato, K., Kennedy, K. L., Nicolas, P., Parker, M. J., Rial-Sebbag, E., Romeo-Casabona, C. M., Shaw, K. M., Wallace, S., Wiesner, G. L., Zeps, N., Lichter, P., Biankin, A. V., Chabannon, C., Chin, L., Clement, B., De Alava, E., Degos, F., Ferguson, M. L., Geary, P., Hayes, D. N., Hudson, T. J., Johns, A. L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. A., Sarin, R., Scarpa, A., Shibata, T., Van De Vijver, M., Futreal, P. A., Aburatani, H., Bayes, M., Botwell, D. D., Campbell, P. J., Estivill, X., Gerhard, D. S., Grimmond, S. M., Gut, I., Hirst, M., Lopez-Otin, C., Majumder, P., Marra, M., Mcpherson, J. D., Nakagawa, H., Ning, Z., Puente, X. S., Ruan, Y., Shibata, T., Stratton, M. R., Stunnenberg, H. G., Swerdlow, H., Velculescu, V. E., Wilson, R. K., Xue, H. H., Yang, L., Spellman, P. T., Bader, G. D., Boutros, P. C., Campbell, P. J., et al. 2010. International network of cancer genome projects. *Nature,* 464**,** 993–8.

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

Ishino, T., Kawashima, S., Tanji, E., Ueno, T., Ueda, Y., Ogasawara, S., Sato, K., Mano, H., Ishihara, S., Kato, N., Kawazu, M. & Togashi, Y. 2023. Somatic mutations can induce a noninflamed tumour microenvironment via their original gene functions, despite deriving neoantigens. *Br J Cancer,* 128**,** 1166–1175.

Ishorst, N., Holzel, S., Greve, C., Yilmaz, O., Lindenberg, T., Lambertz, J., Drichel, D., Zametica, B., Mingardo, E., Kalanithy, J. C., Channab, K., Kibris, D., Henne, S., Degenhardt, F., Siewert, A., Dixon, M., Kruse, T., Ongkosuwito, E., Girisha, K. M., Pande, S., Nowak, S., Hagelueken, G., Geyer, M., Carels, C., Van Rooij, I., Ludwig, K. U., Odermatt, B. & Mangold, E. 2025. Role of ZFHX4 in orofacial clefting based on human genetic data and zebrafish models. *Eur J Hum Genet,* 33**,** 595–606.

Jiang, J., Ding, Y., Wu, M., Chen, Y., Lyu, X., Lu, J., Wang, H. & Teng, L. 2020. Integrated genomic analysis identifies a genetic mutation model predicting response to immune checkpoint inhibitors in melanoma. *Cancer Med,* 9**,** 8498–8518.

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., Li, Z., Traugh, N., Bu, X., Li, B., Liu, J., Freeman, G. J., Brown, M. A., Wucherpfennig, K. W. & Liu, X. S. 2018. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med,* 24**,** 1550–1558.

Jiang, Y., Saunders, M., Seligmann, J., Iveson, T., Wilson, R., Graham, J., Khan, K., Militello, A., Irvine, S., Adedoyin, T., Begum, R., Bhat, R., Wilson, W., Plumb, A., Obichere, A., Rodriguez-Justo, M. & Jansen, M. 2024. NEOPRISM-CRC: Neoadjuvant pembrolizumab stratified to tumour mutation burden for high risk stage 2 or stage 3 deficient-MMR/MSI-high colorectal cancer. *Journal of Clinical Oncology,* 42.

Jimenez-Sanchez, A., Cast, O. & Miller, M. L. 2019. Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res,* 79**,** 6238–6246.

Jin, D., Fan, J., Wang, L., Thompson, L. F., Liu, A., Daniel, B. J., Shin, T., Curiel, T. J. & Zhang, B. 2010. CD73 on tumor cells impairs antitumor T-cell responses: a novel mechanism of tumor-induced immune suppression. *Cancer Res,* 70**,** 2245–55.

Jin, Y., An, X., Mao, B., Sun, R., Kumari, R., Chen, X., Shan, Y., Zang, M., Xu, L., Muntel, J., Beeler, K., Bruderer, R., Reiter, L., Guo, S., Zhou, D., Li, Q. X. & Ouyang, X. 2022. Different syngeneic tumors show distinctive intrinsic tumor-immunity and mechanisms of actions (MOA) of anti-PD-1 treatment. *Sci Rep,* 12**,** 3278.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science,* 337**,** 816–21.

Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C. L. P., Macalinao, D. C., Kahraman, M., Srinivasan, H., Lakshmanan, V., Verbandt, S., Tsantoulis, P., Gunn, N., Venkatesh, P. N., Poh, Z. W., Nahar, R., Oh, H. L. J., Loo, J. M., Chia, S., Cheow, L. F., Cheruba, E., Wong, M. T., Kua, L., Chua, C., Nguyen, A., Golovan, J., Gan, A., Lim, W. J., Guo, Y. A., Yap, C. K., Tay, B., Hong, Y., Chong, D. Q., Chok, A. Y., Park, W. Y., Han, S., Chang, M. H., Seow-En, I., Fu, C., Mathew, R., Toh, E. L., Hong, L. Z., Skanderup, A. J., Dasgupta, R., Ong, C. J., Lim, K. H., Tan, E. K. W., Koo, S. L., Leow, W. Q., Tejpar, S., Prabhakar, S. & Tan, I. B. 2022. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nat Genet,* 54**,** 963–975.

Johansen, A. F. B., Kassentoft, C. G., Knudsen, M., Laursen, M. B., Madsen, A. H., Iversen, L. H., Sunesen, K. G., Rasmussen, M. H. & Andersen, C. L. 2019. Validation of computational determination of microsatellite status using whole exome sequencing data from colorectal cancer patients. *BMC Cancer,* 19**,** 971.

Jung, P., Sato, T., Merlos-Suarez, A., Barriga, F. M., Iglesias, M., Rossell, D., Auer, H., Gallardo, M., Blasco, M. A., Sancho, E., Clevers, H. & Batlle, E. 2011. Isolation and in vitro expansion of human colonic stem cells. *Nat Med,* 17**,** 1225–7.

Kaji, K., Caballero, I. M., Macleod, R., Nichols, J., Wilson, V. A. & Hendrich, B. 2006. The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. *Nat Cell Biol,* 8**,** 285–92.

Kalluri, R. & Weinberg, R. A. 2009. The basics of epithelial-mesenchymal transition. *J Clin Invest,* 119**,** 1420–8.

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., Mcrae, J. F., Short, P. J., Torene, R. I., De Boer, E., Danecek, P., Gardner, E. J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., Reijnders, M. R. F., Yeung, A., Yntema, H. G., Deciphering Developmental Disorders, S., Vissers, L., Juusola, J., Wright, C. F., Brunner, H. G., Firth, H. V., Fitzpatrick, D. R., Barrett, J. C., Hurles, M. E., Gilissen, C. & Retterer, K. 2020. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature,* 586**,** 757–762.

Kapp, A. V. & Tibshirani, R. 2007. Are clusters found in one dataset present in another dataset? *Biostatistics,* 8**,** 9–31.

Kaptein, P., Slingerland, N., Van Der Leun, A. M., Runderkamp, E., Wagensveld, R. A., Chin, S. M., Mors, J. R., Machuca-Ostos, M., Reissig, T. M., Moynihan, K. D., Djuretic, I. M., Yeung, Y. A., Schumacher, T. N. M. & Thommen, D. S. 2025. Reinvigoration of translational activity in dysfunctional T cells initiates the early intratumoral response to PD-1 blockade. *bioRxiv*.

Kasashima, H., Duran, A., Martinez-Ordonez, A., Nakanishi, Y., Kinoshita, H., Linares, J. F., Reina-Campos, M., Kudo, Y., L'hermitte, A., Yashiro, M., Ohira, M., Bao, F., Tauriello, D. V. F., Batlle, E., Diaz-Meco, M. T. & Moscat, J. 2021. Stromal SOX2 Upregulation Promotes Tumorigenesis through the Generation of a SFRP1/2-Expressing Cancer-Associated Fibroblast Population. *Dev Cell,* 56**,** 95–110 e10.

Kautto, E. A., Bonneville, R., Miya, J., Yu, L., Krook, M. A., Reeser, J. W. & Roychowdhury, S. 2017. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget,* 8**,** 7452–7463.

Kennel, K. B. & Greten, F. R. 2025. The immune microenvironment of colorectal cancer. *Nat Rev Cancer*.

Khaliq, A. M., Erdogan, C., Kurt, Z., Turgut, S. S., Grunvald, M. W., Rand, T., Khare, S., Borgia, J. A., Hayden, D. M., Pappas, S. G., Govekar, H. R., Kam, A. E., Reiser, J., Turaga, K., Radovich, M., Zang, Y., Qiu, Y., Liu, Y., Fishel, M. L., Turk, A., Gupta, V., Al-Sabti, R., Subramanian, J., Kuzel, T. M., Sadanandam, A., Waldron, L., Hussain, A., Saleem, M., El-Rayes, B., Salahudeen, A. A. & Masood, A. 2022. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. *Genome Biol,* 23**,** 113.

Kim, J. H., Seo, M. K., Lee, J. A., Yoo, S. Y., Oh, H. J., Kang, H., Cho, N. Y., Bae, J. M., Kang, G. H. & Kim, S. 2021. Genomic and transcriptomic characterization of heterogeneous immune subgroups of microsatellite instability-high colorectal cancers. *J Immunother Cancer,* 9.

Kinzler, K. W. & Vogelstein, B. 1997. Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature,* 386**,** 761, 763.

Klein, G. & Klein, E. 1985. Evolution of tumours and the impact of molecular oncology. *Nature,* 315**,** 190–5.

Klenova, E. M., Nicolas, R. H., U, S., Carne, A. F., Lee, R. E., Lobanenkov, V. V. & Goodwin, G. H. 1997. Molecular weight abnormalities of the CTCF transcription factor: CTCF migrates aberrantly in SDS-PAGE and the size of the expressed protein is affected by

the UTRs and sequences within the coding region of the CTCF gene. *Nucleic Acids Res,* 25**,** 466–74.

Kloet, S. L., Baymaz, H. I., Makowski, M., Groenewold, V., Jansen, P. W., Berendsen, M., Niazi, H., Kops, G. J. & Vermeulen, M. 2015. Towards elucidating the stability, dynamics and architecture of the nucleosome remodeling and deacetylase complex by using quantitative interaction proteomics. *FEBS J,* 282**,** 1774–85.

Kock, K. H., Kimes, P. K., Gisselbrecht, S. S., Inukai, S., Phanor, S. K., Anderson, J. T., Ramakrishnan, G., Lipper, C. H., Song, D., Kurland, J. V., Rogers, J. M., Jeong, R., Blacklow, S. C., Irizarry, R. A. & Bulyk, M. L. 2024. DNA binding analysis of rare variants in homeodomains reveals homeodomain specificity-determining residues. *Nat Commun,* 15**,** 3110.

Kohli, K., Pillarisetty, V. G. & Kim, T. S. 2022. Key chemokines direct migration of immune cells in solid tumors. *Cancer Gene Ther,* 29**,** 10–21.

Koksal, H., Herbst, M., Perreira, P., Nater, M., Regli, N., Boudjeniba, C., Erdem Borgoni, N., Cecconi, V. & Van Den Broek, M. 2025. Pre-existing intratumoral stem-like CD8(+) T cells drive radiotherapy-induced tumor immunity. *Cell Rep,* 44**,** 115566.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature,* 533**,** 420–4.

Kong, J. C. H., Guerra, G. R., Millen, R. M., Roth, S., Xu, H., Neeson, P. J., Darcy, P. K., Kershaw, M. H., Sampurno, S., Malaterre, J., Liu, D. S. H., Pham, T. D., Narasimhan, V., Wang, M., Huang, Y. K., Visvanathan, K., Mccormick, J., Lynch, A. C., Warrier, S., Michael, M., Desai, J., Murray, W., Mitchell, C., Ngan, S., Phillips, W. A., Heriot, A. G. & Ramsay, R. G. 2018. Tumor-Infiltrating Lymphocyte Function Predicts Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. *JCO Precis Oncol,* 2**,** 1–15.

Koster, J. 2025. *R2 Genomics Analysis and Visualization Platform* [Online]. Available: https://hgserver1.amc.nl/cgi-bin/r2/main.cgi [Accessed 18 Aug 2025].

Kostich, W. A. & Sanes, J. R. 1995. Expression of zfh-4, a new member of the zinc finger-homeodomain family, in developing brain and muscle. *Dev Dyn,* 202**,** 145–52.

Krieg, C., Weber, L. M., Fosso, B., Marzano, M., Hardiman, G., Olcina, M. M., Domingo, E., El Aidy, S., Mallah, K., Robinson, M. D. & Guglietta, S. 2022. Complement downregulation promotes an inflammatory signature that renders colorectal cancer susceptible to immunotherapy. *J Immunother Cancer,* 10.

Kromann, E. H., Cearra, A. P. & Neves, J. F. 2024. Organoids as a tool to study homeostatic and pathological immune-epithelial interactions in the gut. *Clin Exp Immunol,* 218**,** 28–39.

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., Van De Velde, C. J. & Watanabe, T. 2015. Colorectal cancer. *Nat Rev Dis Primers,* 1**,** 15065.

Kumar, M., Joseph, S. R., Augsburg, M., Bogdanova, A., Drechsel, D., Vastenhouw, N. L., Buchholz, F., Gentzel, M. & Shevchenko, A. 2018. MS Western, a Method of Multiplexed Absolute Protein Quantification is a Practical Alternative to Western

Blotting. *Mol Cell Proteomics,* 17**,** 384–396.

Kurian, C. J. & Markabawi, D. 2024. CRE24-033: Complete Remission of Microsatellite Instability–High (MSI-H) Oligometastatic Poorly Differentiated Colon Adenocarcinoma After a Single Dose of Immunotherapy. *Journal of the National Comprehensive Cancer Network,* 22.

Lacher, S. B., Dorr, J., De Almeida, G. P., Honninger, J., Bayerl, F., Hirschberger, A., Pedde, A. M., Meiser, P., Ramsauer, L., Rudolph, T. J., Spranger, N., Morotti, M., Grimm, A. J., Jarosch, S., Oner, A., Gregor, L., Lesch, S., Michaelides, S., Fertig, L., Briukhovetska, D., Majed, L., Stock, S., Busch, D. H., Buchholz, V. R., Knolle, P. A., Zehn, D., Dangaj Laniti, D., Kobold, S. & Bottcher, J. P. 2024. PGE(2) limits effector expansion of tumour-infiltrating stem-like CD8(+) T cells. *Nature,* 629**,** 417–425.

Lai, A. Y. & Wade, P. A. 2011. Cancer biology and NuRD: a multifaceted chromatin remodelling complex. *Nat Rev Cancer,* 11**,** 588–96.

Lakatos, E., Gunasri, V., Zapata, L., Househam, J., Heide, T., Trahearn, N., Swinyard, O., Cisneros, L., Lynn, C., Mossner, M., Kimberley, C., Spiteri, I., Cresswell, G. D., Llibre-Palomar, G., Mitchison, M., Maley, C. C., Jansen, M., Rodriguez-Justo, M., Bridgewater, J., Baker, A. M., Sottoriva, A. & Graham, T. A. 2025. Epigenetically driven and early immune evasion in colorectal cancer evolution. *Nat Genet,* 57**,** 3039–3049.

Lal, N., Beggs, A. D., Willcox, B. E. & Middleton, G. W. 2015. An immunogenomic stratification of colorectal cancer: Implications for development of targeted immunotherapy. *Oncoimmunology,* 4**,** e976052.

Lal, N., Chan, D. K. H., Ng, M. E., Vermeulen, L. & Buczacki, S. J. A. 2022. Primary tumour immune response and lymph node yields in colon cancer. *Br J Cancer,* 126**,** 1178–1185.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. & Weirauch, M. T. 2018. The Human Transcription Factors. *Cell,* 172**,** 650–665.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., Mcdaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L. & Maglott, D. R. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res,* 46**,** D1062–D1067.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M. & Maglott, D. R. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res,* 42**,** D980–5.

Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J. J., Rutkowski, P., Lao, C. D., Cowey, C. L., Schadendorf, D., Wagstaff, J., Dummer, R., Ferrucci, P. F., Smylie, M., Hogg, D., Hill, A., Marquez-Rodas, I., Haanen, J., Guidoboni, M., Maio, M., Schoffski, P., Carlino, M. S., Lebbe, C., Mcarthur, G., Ascierto, P. A., Daniels, G. A., Long, G. V., Bastholt, L., Rizzo, J. I., Balogh, A., Moshyk, A., Hodi, F. S. & Wolchok, J. D. 2019. Five-Year Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *N Engl J Med,* 381**,** 1535–1546.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P.

S., Mckenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortes, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., Dicara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D. A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., Mccarroll, S. A., Mora, J., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S. & Getz, G. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* 499**,** 214–218.

Le, D. T., Diaz, L. A., Jr., Kim, T. W., Van Cutsem, E., Geva, R., Jager, D., Hara, H., Burge, M., O'neil, B. H., Kavan, P., Yoshino, T., Guimbaud, R., Taniguchi, H., Elez, E., Al-Batran, S. E., Boland, P. M., Cui, Y., Leconte, P., Marinello, P. & Andre, T. 2023. Pembrolizumab for previously treated, microsatellite instability-high/mismatch repair-deficient advanced colorectal cancer: final analysis of KEYNOTE-164. *Eur J Cancer,* 186**,** 185–195.

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

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., Laheru, D., Biedrzycki, B., Donehower, R. C., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Duffy, S. M., Goldberg, R. M., De La Chapelle, A., Koshiji, M., Bhaijee, F., Huebner, T., Hruban, R. H., Wood, L. D., Cuka, N., Pardoll, D. M., Papadopoulos, N., Kinzler, K. W., Zhou, S., Cornish, T. C., Taube, J. M., Anders, R. A., Eshleman, J. R., Vogelstein, B. & Diaz, L. A., Jr. 2015. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med,* 372**,** 2509–20.

Le Gallo, M., O'hara, A. J., Rudd, M. L., Urick, M. E., Hansen, N. F., O'neil, N. J., Price, J. C., Zhang, S., England, B. M., Godwin, A. K., Sgroi, D. C., Program, N. I. H. I. S. C. C. S., Hieter, P., Mullikin, J. C., Merino, M. J. & Bell, D. W. 2012. Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nat Genet,* 44**,** 1310–5.

Leach, D. R., Krummel, M. F. & Allison, J. P. 1996. Enhancement of antitumor immunity by CTLA-4 blockade. *Science,* 271**,** 1734–6.

Leighton, G. & Williams, D. C., Jr. 2020. The Methyl-CpG-Binding Domain 2 and 3 Proteins and Formation of the Nucleosome Remodeling and Deacetylase Complex. *J Mol Biol,* 432**,** 1624–1639.

Lejon, S., Thong, S. Y., Murthy, A., Alqarni, S., Murzina, N. V., Blobel, G. A., Laue, E. D. & Mackay, J. P. 2011. Insights into association of the NuRD complex with FOG-1 from the crystal structure of an RbAp48.FOG-1 complex. *J Biol Chem,* 286**,** 1196–203.

Lenz, H. J., Van Cutsem, E., Luisa Limon, M., Wong, K. Y. M., Hendlisz, A., Aglietta, M., Garcia-Alfonso, P., Neyns, B., Luppi, G., Cardin, D. B., Dragovich, T., Shah, U.,

Abdullaev, S., Gricar, J., Ledeine, J. M., Overman, M. J. & Lonardi, S. 2022. First-Line Nivolumab Plus Low-Dose Ipilimumab for Microsatellite Instability-High/Mismatch Repair-Deficient Metastatic Colorectal Cancer: The Phase II CheckMate 142 Study. *J Clin Oncol,* 40**,** 161–170.

Leonard, N. A., Corry, S. M., Reidy, E., Egan, H., O'malley, G., Thompson, K., Mcdermott, E., O'neill, A., Zakaria, N., Egan, L. J., Ritter, T., Loessner, D., Redmond, K., Sheehan, M., Canney, A., Hogan, A. M., Hynes, S. O., Treacy, O., Dunne, P. D. & Ryan, A. E. 2024. Tumor-associated mesenchymal stromal cells modulate macrophage phagocytosis in stromal-rich colorectal cancer via PD-1 signaling. *iScience,* 27**,** 110701.

Li, D., Liu, X., Gao, W., Zhao, W., Ji, S., Wang, S., Yang, J., Li, D., Chang, Z., Chen, Y., Sun, X., Zhou, J., Jiao, Y., Du, X. & Dong, G. 2025. An Immune Subtype Classification System Enables the Development of Strategies to Predict and Enhance Immunotherapy Responses in Colorectal Cancer. *Cancer Res,* 85**,** 1441–1458.

Li, G. M. 2008. Mechanisms and functions of DNA mismatch repair. *Cell Res,* 18**,** 85–98.

Li, K., Liu, C., Sui, X., Li, C., Zhang, T., Zhao, T., Zhang, D., Wu, H., Liu, Y., Wang, S., Yang, Y., Lin, B., Wang, W., Yang, F., Chen, X. & Liu, P. 2025. An organoid co-culture model for probing systemic anti-tumor immunity in lung cancer. *Cell Stem Cell,* 32**,** 1218–1234 e7.

Li, Q., Wang, C. & Wang, G. 2025. COL4A1 as a potential prognostic biomarker with functional roles in colon cancer progression. *Sci Rep,* 15**,** 31382.

Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., Kumar, K., Khurana, E., Waszak, S., Korbel, J. O., Haber, J. E., Imielinski, M., Group, P. S. V. W., Weischenfeldt, J., Beroukhim, R., Campbell, P. J. & Consortium, P. 2020. Patterns of somatic structural variation in human cancer genomes. *Nature,* 578**,** 112–121.

Li, Z., Ma, L., Gao, Z., Wang, X., Che, X., Zhang, P., Li, Y., Zhang, Q., Liu, T., Sun, Y., Bai, Y. & Deng, H. 2024. Identification and validation of tumor-specific T cell receptors from tumor infiltrating lymphocytes using tumor organoid co-cultures. *Cancer Immunol Immunother,* 73**,** 164.

Li, Z., Tuong, Z. K., Dean, I., Willis, C., Gaspal, F., Fiancette, R., Idris, S., Kennedy, B., Ferdinand, J. R., Penalver, A., Cabantous, M., Murtuza Baker, S., Fry, J. W., Carlesso, G., Hammond, S. A., Dovedi, S. J., Hepworth, M. R., Clatworthy, M. R. & Withers, D. R. 2022. In vivo labeling reveals continuous trafficking of TCF-1+ T cells between tumor and lymphoid tissue. *J Exp Med,* 219.

Liao, W., Overman, M. J., Boutin, A. T., Shang, X., Zhao, D., Dey, P., Li, J., Wang, G., Lan, Z., Li, J., Tang, M., Jiang, S., Ma, X., Chen, P., Katkhuda, R., Korphaisarn, K., Chakravarti, D., Chang, A., Spring, D. J., Chang, Q., Zhang, J., Maru, D. M., Maeda, D. Y., Zebala, J. A., Kopetz, S., Wang, Y. A. & Depinho, R. A. 2019. KRAS-IRF2 Axis Drives Immune Suppression and Immune Therapy Resistance in Colorectal Cancer. *Cancer Cell,* 35**,** 559–572 e7.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A. & Hemminki, K. 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med,* 343**,** 78–85.

Lin, T., Ponn, A., Hu, X., Law, B. K. & Lu, J. 2010. Requirement of the histone demethylase LSD1 in Snai1-mediated transcriptional repression during epithelial-mesenchymal transition. *Oncogene,* 29**,** 4896–904.

Lin, Y., Pan, X., Chen, Z., Lin, S., Shen, Z. & Chen, S. 2021. Prognostic value and immune infiltration of novel signatures in colon cancer microenvironment. *Cancer Cell Int,* 21**,** 679.

Litchfield, K., Reading, J. L., Puttick, C., Thakkar, K., Abbosh, C., Bentham, R., Watkins, T. B. K., Rosenthal, R., Biswas, D., Rowan, A., Lim, E., Al Bakir, M., Turati, V., Guerra-Assuncao, J. A., Conde, L., Furness, A. J. S., Saini, S. K., Hadrup, S. R., Herrero, J., Lee, S. H., Van Loo, P., Enver, T., Larkin, J., Hellmann, M. D., Turajlic, S., Quezada, S. A., Mcgranahan, N. & Swanton, C. 2021. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell,* 184**,** 596–614 e14.

Liu, D., Schilling, B., Liu, D., Sucker, A., Livingstone, E., Jerby-Arnon, L., Zimmer, L., Gutzmer, R., Satzger, I., Loquai, C., Grabbe, S., Vokes, N., Margolis, C. A., Conway, J., He, M. X., Elmarakeby, H., Dietlein, F., Miao, D., Tracy, A., Gogas, H., Goldinger, S. M., Utikal, J., Blank, C. U., Rauschenberg, R., Von Bubnoff, D., Krackhardt, A., Weide, B., Haferkamp, S., Kiecker, F., Izar, B., Garraway, L., Regev, A., Flaherty, K., Paschen, A., Van Allen, E. M. & Schadendorf, D. 2019. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nat Med,* 25**,** 1916–1927.

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Cancer Genome Atlas Research, N. & Hu, H. 2018. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell,* 173**,** 400–416 e11.

Liu, J., Ye, Y. & Cai, L. 2020. Supramolecular attack particle: the way cytotoxic T lymphocytes kill target cells. *Signal Transduct Target Ther,* 5**,** 210.

Liu, R., Ji, Z., Wang, X., Zhu, L., Xin, J., Ma, L., Zhang, J., Ge, S., Zhang, L., Yang, Y., Ning, T., Bai, M., Duan, J., Wang, F., Sun, Y., Li, H., Deng, T., Ba, Y. & Hao, J. 2025. Regorafenib plus sintilimab as a salvage treatment for microsatellite stable metastatic colorectal cancer: a single-arm, open-label, phase II clinical trial. *Nat Commun,* 16**,** 1481.

Liu, S., Xu, L., Kashima, M., Narumi, R., Takahata, Y., Nakamura, E., Shibuya, H., Tamura, M., Shida, Y., Inubushi, T., Nukada, Y., Miyazawa, M., Hata, K., Nishimura, R., Yamashiro, T., Tasaki, J. & Kurosaka, H. 2025. Expression analysis of genes including Zfhx4 in mice and zebrafish reveals a temporospatial conserved molecular basis underlying craniofacial development. *Dev Dyn,* 254**,** 257–271.

Liu, Z., Guo, Y., Yang, X., Chen, C., Fan, D., Wu, X., Si, C., Xu, Y., Shao, B., Chen, Z., Dang, Q., Cui, W., Han, X., Ji, Z. & Sun, Z. 2021. Immune Landscape Refines the Classification of Colorectal Cancer With Heterogeneous Prognosis, Tumor Microenvironment and Distinct Sensitivity to Frontline Therapies. *Front Cell Dev Biol,* 9**,** 784199.

Loree, J. M., Pereira, A. a. L., Lam, M., Willauer, A. N., Raghav, K., Dasari, A., Morris, V. K., Advani, S., Menter, D. G., Eng, C., Shaw, K., Broaddus, R., Routbort, M. J., Liu, Y., Morris, J. S., Luthra, R., Meric-Bernstam, F., Overman, M. J., Maru, D. & Kopetz, S. 2018. Classifying Colorectal Cancer by Tumor Location Rather than Sidedness

Highlights a Continuum in Mutation Profiles and Consensus Molecular Subtypes. *Clin Cancer Res,* 24**,** 1062–1072.

Lorenzo-Martin, L. F., Broguiere, N., Langer, J., Tillard, L., Nikolaev, M., Coukos, G., Homicsko, K. & Lutolf, M. P. 2025. Patient-derived mini-colons enable long-term modeling of tumor-microenvironment complexity. *Nat Biotechnol,* 43**,** 727–736.

Low, J. K., Webb, S. R., Silva, A. P., Saathoff, H., Ryan, D. P., Torrado, M., Brofelth, M., Parker, B. L., Shepherd, N. E. & Mackay, J. P. 2016. CHD4 Is a Peripheral Component of the Nucleosome Remodeling and Deacetylase Complex. *J Biol Chem,* 291**,** 15853–66.

Low, J. K. K., Silva, A. P. G., Sharifi Tabar, M., Torrado, M., Webb, S. R., Parker, B. L., Sana, M., Smits, C., Schmidberger, J. W., Brillault, L., Jackman, M. J., Williams, D. C., Jr., Blobel, G. A., Hake, S. B., Shepherd, N. E., Landsberg, M. J. & Mackay, J. P. 2020. The Nucleosome Remodeling and Deacetylase Complex Has an Asymmetric, Dynamic, and Modular Architecture. *Cell Rep,* 33**,** 108450.

Lu, N., Wang, J., Zhu, B., Zhang, M., Qi, F., Wang, X. & Gu, J. 2017. Whole-exome sequencing to identify novel mutations of nevoid basal cell carcinoma syndrome in a Chinese population. *Cancer Biomark,* 21**,** 161–168.

Lu, W. & Kang, Y. 2019. Epithelial-Mesenchymal Plasticity in Cancer Progression and Metastasis. *Dev Cell,* 49**,** 361–374.

Luo, R., Chyr, J., Wen, J., Wang, Y., Zhao, W. & Zhou, X. 2023. A novel integrated approach to predicting cancer immunotherapy efficacy. *Oncogene,* 42**,** 1913–1925.

Luque-Garcia, J. L., Zhou, G., Spellman, D. S., Sun, T. T. & Neubert, T. A. 2008. Analysis of electroblotted proteins by mass spectrometry: protein identification after Western blotting. *Mol Cell Proteomics,* 7**,** 308–14.

Ma, J., Setton, J., Lee, N. Y., Riaz, N. & Powell, S. N. 2018. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat Commun,* 9**,** 3292.

Madrova, M., Jiang, Y., Mckenzie, L., Oddy, C., Vlckova, P., Vendramin, R., Shiu, K., Tape, C. & Jansen, M. 2025. *Decoding Immune Evasion in Mismatch-repair Deficient Colorectal Cancer Organoid-Immune Co-culture Models* [Online]. EACR. Available: https://storage.unitedwebnetwork.com/files/991/abstract/EACR25-1442.htm [Accessed].

Magre, L., Verstegen, M. M. A., Buschow, S., Van Der Laan, L. J. W., Peppelenbosch, M. & Desai, J. 2023. Emerging organoid-immune co-culture models for cancer research: from oncoimmunology to personalized immunotherapies. *J Immunother Cancer,* 11.

Malier, M., Gharzeddine, K., Laverriere, M. H., Marsili, S., Thomas, F., Decaens, T., Roth, G. & Millet, A. 2021. Hypoxia Drives Dihydropyrimidine Dehydrogenase Expression in Macrophages and Confers Chemoresistance in Colorectal Cancer. *Cancer Res,* 81**,** 5963–5976.

Malla, S. B., Byrne, R. M., Lafarge, M. W., Corry, S. M., Fisher, N. C., Tsantoulis, P. K., Mills, M. L., Ridgway, R. A., Lannagan, T. R. M., Najumudeen, A. K., Gilroy, K. L., Amirkhah, R., Maguire, S. L., Mulholland, E. J., Belnoue-Davis, H. L., Grassi, E., Viviani, M., Rogan, E., Redmond, K. L., Sakhnevych, S., Mccooey, A. J., Bull, C., Hoey, E., Sinevici, N., Hall, H., Ahmaderaghi, B., Domingo, E., Blake, A., Richman, S. D.,

Isella, C., Miller, C., Bertotti, A., Trusolino, L., Loughrey, M. B., Kerr, E. M., Tejpar, S., Consortium, S. C., Maughan, T. S., Lawler, M., Campbell, A. D., Leedham, S. J., Koelzer, V. H., Sansom, O. J. & Dunne, P. D. 2024. Pathway level subtyping identifies a slow-cycling biological phenotype associated with poor clinical outcomes in colorectal cancer. *Nat Genet,* 56**,** 458–472.

Manshouri, R., Coyaud, E., Kundu, S. T., Peng, D. H., Stratton, S. A., Alton, K., Bajaj, R., Fradette, J. J., Minelli, R., Peoples, M. D., Carugo, A., Chen, F., Bristow, C., Kovacs, J. J., Barton, M. C., Heffernan, T., Creighton, C. J., Raught, B. & Gibbons, D. L. 2019. ZEB1/NuRD complex suppresses TBC1D2b to stimulate E-cadherin internalization and promote metastasis in lung cancer. *Nat Commun,* 10**,** 5125.

Mao, Y., Xie, H., Lv, M., Yang, Q., Shuang, Z., Gao, F., Li, S., Zhu, L. & Wang, W. 2023. The landscape of objective response rate of anti-PD-1/L1 monotherapy across 31 types of cancer: a system review and novel biomarker investigating. *Cancer Immunol Immunother,* 72**,** 2483–2498.

Mao, Y., Xu, Y., Chang, J., Chang, W., Lv, Y., Zheng, P., Zhang, Z., Li, Z., Lin, Q., Tang, W., Zhu, D., Ji, M., He, G., Feng, Q. & Xu, J. 2022. The immune phenotypes and different immune escape mechanisms in colorectal cancer. *Front Immunol,* 13**,** 968089.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. & Califano, A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics,* 7 Suppl 1**,** S7.

Marisa, L., Blum, Y., Taieb, J., Ayadi, M., Pilati, C., Le Malicot, K., Lepage, C., Salazar, R., Aust, D., Duval, A., Blons, H., Taly, V., Gentien, D., Rapinat, A., Selves, J., Mouillet-Richard, S., Boige, V., Emile, J. F., De Reynies, A. & Laurent-Puig, P. 2021. Intratumor CMS Heterogeneity Impacts Patient Prognosis in Localized Colon Cancer. *Clin Cancer Res,* 27**,** 4768–4780.

Markowitz, S. D. & Bertagnolli, M. M. 2009. Molecular origins of cancer: Molecular basis of colorectal cancer. *N Engl J Med,* 361**,** 2449–60.

Marques, A., Cavaco, P., Torre, C., Sepodes, B. & Rocha, J. 2024. Tumor mutational burden in colorectal cancer: Implications for treatment. *Crit Rev Oncol Hematol,* 197**,** 104342.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R. & Campbell, P. J. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell,* 171**,** 1029–1041 e21.

Martinez-Jimenez, F., Muinos, F., Sentis, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., Gonzalez-Perez, A. & Lopez-Bigas, N. 2020. A compendium of mutational cancer driver genes. *Nat Rev Cancer,* 20**,** 555–572.

Matano, M., Date, S., Shimokawa, M., Takano, A., Fujii, M., Ohta, Y., Watanabe, T., Kanai, T. & Sato, T. 2015. Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. *Nat Med,* 21**,** 256–62.

Matheu, A., Collado, M., Wise, C., Manterola, L., Cekaite, L., Tye, A. J., Canamero, M., Bujanda, L., Schedl, A., Cheah, K. S., Skotheim, R. I., Lothe, R. A., Lopez De Munain, A., Briscoe, J., Serrano, M. & Lovell-Badge, R. 2012. Oncogenicity of the developmental transcription factor Sox9. *Cancer Res,* 72**,** 1301–15.

Mathur, A., Ware, C., Davis, L., Gazdar, A., Pan, B. S. & Lutterbach, B. 2014. FGFR2 is amplified in the NCI-H716 colorectal cancer cell line and is required for growth and survival. *PLoS One,* 9**,** e98515.

Maung, H., Gregory, O., De Hoog, T., Hutchinson, M., Soh, D. P. B., Marino, M., Evans, T., Yeoh, A. & Turner, R. C. 2024. Primary medullary adenocarcinoma of the colon: Literature review and case series. *Surg Pract Sci,* 19**,** 100254.

Mcdermott, D. F., Huseni, M. A., Atkins, M. B., Motzer, R. J., Rini, B. I., Escudier, B., Fong, L., Joseph, R. W., Pal, S. K., Reeves, J. A., Sznol, M., Hainsworth, J., Rathmell, W. K., Stadler, W. M., Hutson, T., Gore, M. E., Ravaud, A., Bracarda, S., Suarez, C., Danielli, R., Gruenwald, V., Choueiri, T. K., Nickles, D., Jhunjhunwala, S., Piault-Louis, E., Thobhani, A., Qiu, J., Chen, D. S., Hegde, P. S., Schiff, C., Fine, G. D. & Powles, T. 2018. Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nat Med,* 24**,** 749–757.

Mcfarland, C. D., Mirny, L. A. & Korolev, K. S. 2014. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proc Natl Acad Sci U S A,* 111**,** 15138–43.

Mcfarland, C. D., Yaglom, J. A., Wojtkowiak, J. W., Scott, J. G., Morse, D. L., Sherman, M. Y. & Mirny, L. A. 2017. The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer Res,* 77**,** 4763–4772.

Mcgranahan, N., Furness, A. J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Hiley, C. T., Watkins, T. B., Shafi, S., Murugaesu, N., Mitter, R., Akarca, A. U., Linares, J., Marafioti, T., Henry, J. Y., Van Allen, E. M., Miao, D., Schilling, B., Schadendorf, D., Garraway, L. A., Makarov, V., Rizvi, N. A., Snyder, A., Hellmann, M. D., Merghoub, T., Wolchok, J. D., Shukla, S. A., Wu, C. J., Peggs, K. S., Chan, T. A., Hadrup, S. R., Quezada, S. A. & Swanton, C. 2016. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science,* 351**,** 1463–9.

Mcgranahan, N. & Swanton, C. 2017. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell,* 168**,** 613–628.

Mcmullan, T. W., Crolla, J. A., Gregory, S. G., Carter, N. P., Cooper, R. A., Howell, G. R. & Robinson, D. O. 2002. A candidate gene for congenital bilateral isolated ptosis identified by molecular analysis of a de novo balanced translocation. *Hum Genet,* 110**,** 244–50.

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

Medler, T. R., Blair, T. C., Crittenden, M. R. & Gough, M. J. 2021. Defining Immunogenic and Radioimmunogenic Tumors. *Front Oncol,* 11**,** 667075.

Meers, M. P., Tenenbaum, D. & Henikoff, S. 2019. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics Chromatin,* 12**,** 42.

Metcalf, D. 2010. The colony-stimulating factors and cancer. *Nat Rev Cancer,* 10**,** 425–34.

Miao, D., Margolis, C. A., Vokes, N. I., Liu, D., Taylor-Weiner, A., Wankowicz, S. M., Adeegbe, D., Keliher, D., Schilling, B., Tracy, A., Manos, M., Chau, N. G., Hanna, G. J., Polak, P., Rodig, S. J., Signoretti, S., Sholl, L. M., Engelman, J. A., Getz, G., Janne, P. A., Haddad, R. I., Choueiri, T. K., Barbie, D. A., Haq, R., Awad, M. M., Schadendorf, D., Hodi, F. S., Bellmunt, J., Wong, K. K., Hammerman, P. & Van Allen, E. M. 2018. Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nat Genet,* 50**,** 1271–1281.

Miao, Y., Du, Q., Zhang, H. G., Yuan, Y., Zuo, Y. & Zheng, H. 2023. Cycloheximide (CHX) Chase Assay to Examine Protein Half-life. *Bio Protoc,* 13**,** e4690.

Micalet, A., Upadhyay, A., Javanmardi, Y., De Brito, C. G., Moeendarbary, E. & Cheema, U. 2024. Patient-specific colorectal-cancer-associated fibroblasts modulate tumor microenvironment mechanics. *iScience,* 27**,** 110060.

Miettinen, P. J., Ebner, R., Lopez, A. R. & Derynck, R. 1994. TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J Cell Biol,* 127**,** 2021–36.

Millstein, J., Budden, T., Goode, E. L., Anglesio, M. S., Talhouk, A., Intermaggio, M. P., Leong, H. S., Chen, S., Elatre, W., Gilks, B., Nazeran, T., Volchek, M., Bentley, R. C., Wang, C., Chiu, D. S., Kommoss, S., Leung, S. C. Y., Senz, J., Lum, A., Chow, V., Sudderuddin, H., Mackenzie, R., George, J., Group, A., Fereday, S., Hendley, J., Traficante, N., Steed, H., Koziak, J. M., Kobel, M., Mcneish, I. A., Goranova, T., Ennis, D., Macintyre, G., Silva De Silva, D., Ramon, Y. C. T., Garcia-Donas, J., Hernando Polo, S., Rodriguez, G. C., Cushing-Haugen, K. L., Harris, H. R., Greene, C. S., Zelaya, R. A., Behrens, S., Fortner, R. T., Sinn, P., Herpel, E., Lester, J., Lubinski, J., Oszurek, O., Toloczko, A., Cybulski, C., Menkiszak, J., Pearce, C. L., Pike, M. C., Tseng, C., Alsop, J., Rhenius, V., Song, H., Jimenez-Linan, M., Piskorz, A. M., Gentry-Maharaj, A., Karpinskyj, C., Widschwendter, M., Singh, N., Kennedy, C. J., Sharma, R., Harnett, P. R., Gao, B., Johnatty, S. E., Sayer, R., Boros, J., Winham, S. J., Keeney, G. L., Kaufmann, S. H., Larson, M. C., Luk, H., Hernandez, B. Y., Thompson, P. J., Wilkens, L. R., Carney, M. E., Trabert, B., Lissowska, J., Brinton, L., Sherman, M. E., Bodelon, C., Hinsley, S., Lewsley, L. A., Glasspool, R., Banerjee, S. N., Stronach, E. A., Haluska, P., Ray-Coquard, I., Mahner, S., Winterhoff, B., Slamon, D., Levine, D. A., Kelemen, L. E., Benitez, J., Chang-Claude, J., et al. 2020. Prognostic gene expression signature for high-grade serous ovarian cancer. *Ann Oncol,* 31**,** 1240–1250.

Minsky, B. D. 2011. Unique considerations in the patient with rectal cancer. *Semin Oncol,* 38**,** 542–51.

Misetic, H., Keddar, M. R., Jeannon, J. P. & Ciccarelli, F. D. 2023. Mechanistic insights into the interactions between cancer drivers and the tumour immune microenvironment. *Genome Med,* 15**,** 40.

Mittal, D., Gubin, M. M., Schreiber, R. D. & Smyth, M. J. 2014. New insights into cancer immunoediting and its three component phases--elimination, equilibrium and escape. *Curr Opin Immunol,* 27**,** 16–25.

Mlecnik, B., Bindea, G., Angell, H. K., Maby, P., Angelova, M., Tougeron, D., Church, S. E., Lafontaine, L., Fischer, M., Fredriksen, T., Sasso, M., Bilocq, A. M., Kirilovsky, A., Obenauf, A. C., Hamieh, M., Berger, A., Bruneval, P., Tuech, J. J., Sabourin, J. C., Le Pessot, F., Mauillon, J., Rafii, A., Laurent-Puig, P., Speicher, M. R., Trajanoski, Z., Michel, P., Sesboue, R., Frebourg, T., Pages, F., Valge-Archer, V., Latouche, J. B. & Galon, J. 2016. Integrative Analyses of Colorectal Cancer Show Immunoscore Is a

Stronger Predictor of Patient Survival Than Microsatellite Instability. *Immunity,* 44**,** 698–711.

Mlecnik, B., Bindea, G., Kirilovsky, A., Angell, H. K., Obenauf, A. C., Tosolini, M., Church, S. E., Maby, P., Vasaturo, A., Angelova, M., Fredriksen, T., Mauger, S., Waldner, M., Berger, A., Speicher, M. R., Pages, F., Valge-Archer, V. & Galon, J. 2016. The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis. *Sci Transl Med,* 8**,** 327ra26.

Mlecnik, B., Tosolini, M., Kirilovsky, A., Berger, A., Bindea, G., Meatchi, T., Bruneval, P., Trajanoski, Z., Fridman, W. H., Pages, F. & Galon, J. 2011. Histopathologic-based prognostic factors of colorectal cancers are associated with the state of the local immune reaction. *J Clin Oncol,* 29**,** 610–8.

Monach, P. A., Meredith, S. C., Siegel, C. T. & Schreiber, H. 1995. A unique tumor antigen produced by a single amino acid substitution. *Immunity,* 2**,** 45–59.

Monahan, K. J., Bradshaw, N., Dolwani, S., Desouza, B., Dunlop, M. G., East, J. E., Ilyas, M., Kaur, A., Lalloo, F., Latchford, A., Rutter, M. D., Tomlinson, I., Thomas, H. J. W., Hill, J. & Hereditary, C. R. C. G. E. C. G. 2020. Guidelines for the management of hereditary colorectal cancer from the British Society of Gastroenterology (BSG)/Association of Coloproctology of Great Britain and Ireland (ACPGBI)/United Kingdom Cancer Genetics Group (UKCGG). *Gut,* 69**,** 411–444.

Monti, S. T., P.; Mesirov, J.; Golub, T. 2003. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning,* 52**,** 91–118.

Morales, J., Pujar, S., Loveland, J. E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C. M., Fatima, R., Gil, L., Goldfarb, T., Gonzalez, J. M., Haddad, D., Hardy, M., Hunt, T., Jackson, J., Joardar, V. S., Kay, M., Kodali, V. K., Mcgarvey, K. M., Mcmahon, A., Mudge, J. M., Murphy, D. N., Murphy, M. R., Rajput, B., Rangwala, S. H., Riddick, L. D., Thibaud-Nissen, F., Threadgold, G., Vatsan, A. R., Wallin, C., Webb, D., Flicek, P., Birney, E., Pruitt, K. D., Frankish, A., Cunningham, F. & Murphy, T. D. 2022. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature,* 604**,** 310–315.

Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C. J., Laversanne, M., Vignat, J., Ferlay, J., Murphy, N. & Bray, F. 2023. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut,* 72**,** 338–344.

Morris, V. K., Parseghian, C., Escano, M., Johnson, B., Raghav, K., Dasari, A., Huey, R., Overman, M. J., Willis, J., Lee, M., Wolff, R., Kee, B., Shen, J., Morelli, M., Tam, A., Foo, W., Xiao, L. & Kopetz, S. 2022. Phase I/II trial of encorafenib, cetuximab, and nivolumab in patients with microsatellite stable, BRAFV600E metastatic colorectal cancer. *Journal of Clinical Oncology,* 40.

Morris, V. K., Parseghian, C. M., Bahrambeigi, V., Abdelfattah, N., Xiao, L., Agrawal, A., Lin, K., Raghav, K. P. S., Wolff, R. A., Dasari, A., Huey, R. W., Kee, B. K., Overman, M. J., Willis, J. A., Le, P. H., Escano, M., Baig, Y. C., Pan, K., Menter, D., Tam, A. L., Foo, W. C., Shen, L., Lee, H. M., Gallup, T. D., Margain, C., Gallup, D., Rajapakshe, K. I., Guerrero, P. A., Wang, J., Corcoran, R. B., Maitra, A., Yun, K. & Kopetz, S. 2025. Phase 1/2 trial of encorafenib, cetuximab, and nivolumab in microsatellite stable BRAF(V600E) metastatic colorectal cancer. *Cancer Cell*.

Mosa, M. H., Michels, B. E., Menche, C., Nicolas, A. M., Darvishi, T., Greten, F. R. & Farin, H. F. 2020. A Wnt-Induced Phenotypic Switch in Cancer-Associated Fibroblasts Inhibits EMT in Colorectal Cancer. *Cancer Res,* 80**,** 5569–5582.

Mroz, E. A. & Rocco, J. W. 2013. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol,* 49**,** 211–5.

Munro, D., Ghersi, D. & Singh, M. 2018. Two critical positions in zinc finger domains are heavily mutated in three human cancer types. *PLoS Comput Biol,* 14**,** e1006290.

Murphy, N., Norat, T., Ferrari, P., Jenab, M., Bueno-De-Mesquita, B., Skeie, G., Dahm, C. C., Overvad, K., Olsen, A., Tjonneland, A., Clavel-Chapelon, F., Boutron-Ruault, M. C., Racine, A., Kaaks, R., Teucher, B., Boeing, H., Bergmann, M. M., Trichopoulou, A., Trichopoulos, D., Lagiou, P., Palli, D., Pala, V., Panico, S., Tumino, R., Vineis, P., Siersema, P., Van Duijnhoven, F., Peeters, P. H., Hjartaker, A., Engeset, D., Gonzalez, C. A., Sanchez, M. J., Dorronsoro, M., Navarro, C., Ardanaz, E., Quiros, J. R., Sonestedt, E., Ericson, U., Nilsson, L., Palmqvist, R., Khaw, K. T., Wareham, N., Key, T. J., Crowe, F. L., Fedirko, V., Wark, P. A., Chuang, S. C. & Riboli, E. 2012. Dietary fibre intake and risks of cancers of the colon and rectum in the European prospective investigation into cancer and nutrition (EPIC). *PLoS One,* 7**,** e39361.

Naba, A., Clauser, K. R., Whittaker, C. A., Carr, S. A., Tanabe, K. K. & Hynes, R. O. 2014. Extracellular matrix signatures of human primary metastatic colon cancers and their metastases to liver. *BMC Cancer,* 14**,** 518.

Nakamura, E., Hata, K., Takahata, Y., Kurosaka, H., Abe, M., Abe, T., Kihara, M., Komori, T., Kobayashi, S., Murakami, T., Inubushi, T., Yamashiro, T., Yamamoto, S., Akiyama, H., Kawaguchi, M., Sakata, N. & Nishimura, R. 2021. Zfhx4 regulates endochondral ossification as the transcriptional platform of Osterix in mice. *Commun Biol,* 4**,** 1258.

Nathan, E. & Tzahor, E. 2009. sFRPs: a declaration of (Wnt) independence. *Nat Cell Biol,* 11**,** 13.

Neb. 2025. *Primer Annealing Temperature Calculator* [Online]. Available: https://tmcalculator.neb.com/ [Accessed 23 Oct 2023].

Nhs England. 2025. *Bowel cancer screening standards data report 2023-24* [Online]. GOV.UK. Available: https://www.gov.uk/government/publications/bowel-cancer-screening-annual-report-2023-to-2024/bowel-cancer-screening-standards-data-report-2023-24 [Accessed 7 Oct 2025].

Nice. 2025. *Colorectal Cancer: NG151 Guidance* [Online]. NICE: National Institute for Health and Care Excellence. Available: https://www.nice.org.uk/guidance/ng151 [Accessed 7 Oct 2025].

Nicosia, L., Pranke, I., Latorre, R. V., Murray, J. B., Lonetti, L., Cavusoglu-Doran, K., Dreano, E., Costello, J. P., Carroll, M., Melotti, P., Sorio, C., Sermet-Gaudelus, I., Scallan, M. F. & Harrison, P. T. 2025. Adenine base editing with engineered virus-like particles rescues the CFTR mutation G542X in patient-derived intestinal organoids. *iScience,* 28**,** 111979.

Nie, F., Sun, X., Sun, J., Zhang, J. & Wang, Y. 2025. Epithelial-mesenchymal transition in colorectal cancer metastasis and progression: molecular mechanisms and therapeutic strategies. *Cell Death Discov,* 11**,** 336.

Nieto, M. A., Huang, R. Y., Jackson, R. A. & Thiery, J. P. 2016. Emt: 2016. *Cell,* 166**,** 21–45.

Nih National Cancer Institute. 2025. *GDC Data Portal: TCGA-COAD* [Online]. Available: https://portal.gdc.cancer.gov/projects/TCGA-COAD [Accessed 23 June 2023].

Nih National Cancer Institute. 2025. *Cancer Stat Facts: Colorectal Cancer* [Online]. NIH National Cancer Institute. Available: https://seer.cancer.gov/statfacts/html/colorect.html [Accessed 7 Oct 2025].

Ninomiya, T., Mihara, K., Fushimi, K., Hayashi, Y., Hashimoto-Tamaoki, T. & Tamaoki, T. 2002. Regulation of the alpha-fetoprotein gene by the isoforms of ATBF1 transcription factor in human hepatoma. *Hepatology,* 35**,** 82–7.

Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., Mclellan, M. D., Wendl, M. C. & Ding, L. 2014. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics,* 30**,** 1015–6.

Noel, G., Baetz, N. W., Staab, J. F., Donowitz, M., Kovbasnjuk, O., Pasetti, M. F. & Zachos, N. C. 2017. A primary human macrophage-enteroid co-culture model to investigate mucosal gut physiology and host-pathogen interactions. *Sci Rep,* 7**,** 45270.

Nogami, S., Ishii, Y., Kawaguchi, M., Sakata, N., Oya, T., Takagawa, K., Kanamori, M., Sabit, H., Obata, T., Kimura, T. & Sasahara, M. 2005. ZFH4 protein is expressed in many neurons of developing rat brain. *J Comp Neurol,* 482**,** 33–49.

Nowell, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23–8.

Nunes, L., Li, F., Wu, M., Luo, T., Hammarstrom, K., Torell, E., Ljuslinder, I., Mezheyeuski, A., Edqvist, P. H., Lofgren-Burstrom, A., Zingmark, C., Edin, S., Larsson, C., Mathot, L., Osterman, E., Osterlund, E., Ljungstrom, V., Neves, I., Yacoub, N., Guethnadottir, U., Birgisson, H., Enblad, M., Ponten, F., Palmqvist, R., Xu, X., Uhlen, M., Wu, K., Glimelius, B., Lin, C. & Sjoblom, T. 2024. Prognostic genome and transcriptome signatures in colorectal cancers. *Nature,* 633**,** 137–146.

Nusinow, D. P., Szpyt, J., Ghandi, M., Rose, C. M., Mcdonald, E. R., 3rd, Kalocsay, M., Jane-Valbuena, J., Gelfand, E., Schweppe, D. K., Jedrychowski, M., Golji, J., Porter, D. A., Rejtar, T., Wang, Y. K., Kryukov, G. V., Stegmeier, F., Erickson, B. K., Garraway, L. A., Sellers, W. R. & Gygi, S. P. 2020. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell,* 180**,** 387–402 e16.

O'shaughnessy, A. & Hendrich, B. 2013. CHD4 in the DNA-damage response and cell cycle progression: not so NuRDy now. *Biochem Soc Trans,* 41**,** 777–82.

Olcina, M. M., Balanis, N. G., Kim, R. K., Aksoy, B. A., Kodysh, J., Thompson, M. J., Hammerbacher, J., Graeber, T. G. & Giaccia, A. J. 2018. Mutations in an Innate Immunity Pathway Are Associated with Poor Overall Survival Outcomes and Hypoxic Signaling in Cancer. *Cell Rep,* 25**,** 3721–3732 e6.

Oliver, A. J., Huang, N., Bartolome-Casado, R., Li, R., Koplev, S., Nilsen, H. R., Moy, M., Cakir, B., Polanski, K., Gudino, V., Melon-Ardanaz, E., Sumanaweera, D., Dimitrov, D., Milchsack, L. M., Fitzpatrick, M. E. B., Provine, N. M., Boccacino, J. M., Dann, E., Predeus, A. V., To, K., Prete, M., Chapman, J. A., Masi, A. C., Stephenson, E., Engelbert, J., Lobentanzer, S., Perera, S., Richardson, L., Kapuge, R., Wilbrey-Clark, A., Semprich, C. I., Ellams, S., Tudor, C., Joseph, P., Garrido-Trigo, A., Corraliza, A.

M., Oliver, T. R. W., Hook, C. E., James, K. R., Mahbubani, K. T., Saeb-Parsy, K., Zilbauer, M., Saez-Rodriguez, J., Hoivik, M. L., Baekkevold, E. S., Stewart, C. J., Berrington, J. E., Meyer, K. B., Klenerman, P., Salas, A., Haniffa, M., Jahnsen, F. L., Elmentaite, R. & Teichmann, S. A. 2024. Single-cell integration reveals metaplasia in inflammatory gut diseases. *Nature,* 635**,** 699–707.

Olson, B., Li, Y., Lin, Y., Liu, E. T. & Patnaik, A. 2018. Mouse Models for Cancer Immunotherapy Research. *Cancer Discov,* 8**,** 1358–1365.

Overman, M. J., Lonardi, S., Wong, K. Y. M., Lenz, H. J., Gelsomino, F., Aglietta, M., Morse, M. A., Van Cutsem, E., Mcdermott, R., Hill, A., Sawyer, M. B., Hendlisz, A., Neyns, B., Svrcek, M., Moss, R. A., Ledeine, J. M., Cao, Z. A., Kamble, S., Kopetz, S. & Andre, T. 2018. Durable Clinical Benefit With Nivolumab Plus Ipilimumab in DNA Mismatch Repair-Deficient/Microsatellite Instability-High Metastatic Colorectal Cancer. *J Clin Oncol,* 36**,** 773–779.

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

Ozga, A. J., Chow, M. T. & Luster, A. D. 2021. Chemokines and the immune response to cancer. *Immunity,* 54**,** 859–874.

Pages, F., Mlecnik, B., Marliot, F., Bindea, G., Ou, F. S., Bifulco, C., Lugli, A., Zlobec, I., Rau, T. T., Berger, M. D., Nagtegaal, I. D., Vink-Borger, E., Hartmann, A., Geppert, C., Kolwelter, J., Merkel, S., Grutzmann, R., Van Den Eynde, M., Jouret-Mourin, A., Kartheuser, A., Leonard, D., Remue, C., Wang, J. Y., Bavi, P., Roehrl, M. H. A., Ohashi, P. S., Nguyen, L. T., Han, S., Macgregor, H. L., Hafezi-Bakhtiari, S., Wouters, B. G., Masucci, G. V., Andersson, E. K., Zavadova, E., Vocka, M., Spacek, J., Petruzelka, L., Konopasek, B., Dundr, P., Skalova, H., Nemejcova, K., Botti, G., Tatangelo, F., Delrio, P., Ciliberto, G., Maio, M., Laghi, L., Grizzi, F., Fredriksen, T., Buttard, B., Angelova, M., Vasaturo, A., Maby, P., Church, S. E., Angell, H. K., Lafontaine, L., Bruni, D., El Sissy, C., Haicheur, N., Kirilovsky, A., Berger, A., Lagorce, C., Meyers, J. P., Paustian, C., Feng, Z., Ballesteros-Merino, C., Dijkstra, J., Van De Water, C., Van Lent-Van Vliet, S., Knijn, N., Musina, A. M., Scripcariu, D. V., Popivanova, B., Xu, M., Fujita, T., Hazama, S., Suzuki, N., Nagano, H., Okuno, K., Torigoe, T., Sato, N., Furuhata, T., Takemasa, I., Itoh, K., Patel, P. S., Vora, H. H., Shah, B., Patel, J. B., Rajvik, K. N., Pandya, S. J., Shukla, S. N., Wang, Y., Zhang, G., Kawakami, Y., Marincola, F. M., Ascierto, P. A., Sargent, D. J., Fox, B. A. & Galon, J. 2018. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet,* 391**,** 2128–2139.

Palles, C., Cazier, J. B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., Kemp, Z., Spain, S. L., Guarino, E., Salguero, I., Sherborne, A., Chubb, D., Carvajal-Carmona, L. G., Ma, Y., Kaur, K., Dobbins, S., Barclay, E., Gorman, M., Martin, L., Kovac, M. B., Humphray, S., Consortium, C., Consortium, W. G. S., Lucassen, A., Holmes, C. C., Bentley, D., Donnelly, P., Taylor, J., Petridis, C., Roylance, R., Sawyer, E. J., Kerr, D. J., Clark, S., Grimes, J., Kearsey, S. E., Thomas, H. J., Mcvean, G., Houlston, R. S. & Tomlinson, I. 2013. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet,* 45**,** 136–44.

Palomares, M., Delicado, A., Mansilla, E., De Torres, M. L., Vallespin, E., Fernandez, L., Martinez-Glez, V., Garcia-Minaur, S., Nevado, J., Simarro, F. S., Ruiz-Perez, V. L., Lynch, S. A., Sharkey, F. H., Thuresson, A. C., Anneren, G., Belligni, E. F., Martinez-Fernandez, M. L., Bermejo, E., Nowakowska, B., Kutkowska-Kazmierczak, A., Bocian, E., Obersztyn, E., Martinez-Frias, M. L., Hennekam, R. C. & Lapunzina, P. 2011. Characterization of a 8q21.11 microdeletion syndrome associated with intellectual disability and a recognizable phenotype. *Am J Hum Genet,* 89**,** 295–301.

Parker, B. S., Rautela, J. & Hertzog, P. J. 2016. Antitumour actions of interferons: implications for cancer therapy. *Nat Rev Cancer,* 16**,** 131–44.

Pecina-Slaus, N., Kafka, A., Salamon, I. & Bukovac, A. 2020. Mismatch Repair Pathway, Genome Stability and Cancer. *Front Mol Biosci,* 7**,** 122.

Pei, J., Gao, Y. & Wu, A. 2024. An inflammation-related subtype classification for analyzing tumor microenvironment and clinical prognosis in colorectal cancer. *Front Immunol,* 15**,** 1369726.

Pelka, K., Hofree, M., Chen, J. H., Sarkizova, S., Pirl, J. D., Jorgji, V., Bejnood, A., Dionne, D., Ge, W. H., Xu, K. H., Chao, S. X., Zollinger, D. R., Lieb, D. J., Reeves, J. W., Fuhrman, C. A., Hoang, M. L., Delorey, T., Nguyen, L. T., Waldman, J., Klapholz, M., Wakiro, I., Cohen, O., Albers, J., Smillie, C. S., Cuoco, M. S., Wu, J., Su, M. J., Yeung, J., Vijaykumar, B., Magnuson, A. M., Asinovski, N., Moll, T., Goder-Reiser, M. N., Applebaum, A. S., Brais, L. K., Dellostritto, L. K., Denning, S. L., Phillips, S. T., Hill, E. K., Meehan, J. K., Frederick, D. T., Sharova, T., Kanodia, A., Todres, E. Z., Jane-Valbuena, J., Biton, M., Izar, B., Lambden, C. D., Clancy, T. E., Bleday, R., Melnitchouk, N., Irani, J., Kunitake, H., Berger, D. L., Srivastava, A., Hornick, J. L., Ogino, S., Rotem, A., Vigneau, S., Johnson, B. E., Corcoran, R. B., Sharpe, A. H., Kuchroo, V. K., Ng, K., Giannakis, M., Nieman, L. T., Boland, G. M., Aguirre, A. J., Anderson, A. C., Rozenblatt-Rosen, O., Regev, A. & Hacohen, N. 2021. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell,* 184**,** 4734–4752 e20.

Perez Baca, M. D. R., Jacobs, E. Z., Vantomme, L., Leblanc, P., Bogaert, E., Dheedene, A., De Cock, L., Haghshenas, S., Foroutan, A., Levy, M. A., Kerkhof, J., Mcconkey, H., Chen, C. A., Batzir, N. A., Wang, X., Palomares, M., Carels, M., Consortium, Z., Dermaut, B., Sadikovic, B., Menten, B., Yuan, B., Vergult, S. & Callewaert, B. 2024. Haploinsufficiency of ZFHX3, encoding a key player in neuronal development, causes syndromic intellectual disability. *Am J Hum Genet,* 111**,** 509–528.

Perez Baca, M. D. R., Palomares-Bralo, M., Vanhooydonck, M., Hamerlinck, L., D'haene, E., Leimbacher, S., Jacobs, E. Z., De Cock, L., D'haenens, E., Dheedene, A., Malfait, Z., Vantomme, L., Silva, A., Rooney, K., Zhao, X., Saeidian, A. H., Owen, N. M., Santos-Simarro, F., Lleuger-Pujol, R., Garcia-Minaur, S., Losantos-Garcia, I., Menten, B., Gestri, G., Ragge, N., Consortium, Z., Sadikovic, B., Bogaert, E., Vleminckx, K., Naert, T., Syx, D., Callewaert, B. & Vergult, S. 2025. Loss of function of the zinc finger homeobox 4 gene, ZFHX4, underlies a neurodevelopmental disorder. *Am J Hum Genet,* 112**,** 1388–1414.

Perez-Oquendo, M., Manshouri, R., Tian, Y., Fradette, J. J., Rodriguez, B. L., Kundu, S. T. & Gibbons, D. L. 2023. ZEB1 Is Regulated by K811 Acetylation to Promote Stability, NuRD Complex Interactions, EMT, and NSCLC Metastasis. *Mol Cancer Res,* 21**,** 779–794.

Persikov, A. V. & Singh, M. 2014. De novo prediction of DNA-binding specificities for

Cys2His2 zinc finger proteins. *Nucleic Acids Res,* 42**,** 97–108.

Podaza, E., Capuano, J., Assaad, M. A., Kuo, H., Markowitz, G., Irizarry, A., Ravichandran, H., Ackermann, S., Kane, T., Manohar, J., Sigouros, M., Moyer, J., Bhinder, B., Chandra, P., Malbari, M., Boehnke, K., Mosquera, J. M., Mittal, V., Sboner, A., Gokozan, H., Altorki, N., Elemento, O. & Martin, M. L. 2023. Novel co-culture strategies of tumor organoids with autologous T-cells reveal clinically relevant combinations of immune-checkpoint and targeted therapies. *bioRxiv*.

Porgador, A., Yewdell, J. W., Deng, Y., Bennink, J. R. & Germain, R. N. 1997. Localization, quantitation, and in situ detection of specific peptide-MHC class I complexes using a monoclonal antibody. *Immunity,* 6**,** 715–26.

Porta-Pardo, E. & Godzik, A. 2016. Mutation Drivers of Immunological Responses to Cancer. *Cancer Immunol Res,* 4**,** 789–98.

Posit Team 2023. RStudio v2023.09.1+494: Integrated Development Environment for R. Posit Software. Available at: http://www.posit.co/.

Posit Team 2025. RStudio Online: Integrated Development Environment for R. Posit Software. Available at: https://posit.cloud.

Powell, S. M., Zilz, N., Beazer-Barclay, Y., Bryan, T. M., Hamilton, S. R., Thibodeau, S. N., Vogelstein, B. & Kinzler, K. W. 1992. APC mutations occur early during colorectal tumorigenesis. *Nature,* 359**,** 235–7.

Prager, I. & Watzl, C. 2019. Mechanisms of natural killer cell-mediated cellular cytotoxicity. *J Leukoc Biol,* 105**,** 1319–1329.

Qi, W. & Zhang, Q. 2022. Identification and Validation of Immune Molecular Subtypes and Immune Landscape Based on Colon Cancer Cohort. *Front Med (Lausanne),* 9**,** 827695.

Qin, X., Cardoso Rodriguez, F., Sufi, J., Vlckova, P., Claus, J. & Tape, C. J. 2023. An oncogenic phenoscape of colonic stem cell polarization. *Cell,* 186**,** 5554–5568 e18.

Qing, T., Zhu, S., Suo, C., Zhang, L., Zheng, Y. & Shi, L. 2017. Somatic mutations in ZFHX4 gene are associated with poor overall survival of Chinese esophageal squamous cell carcinoma patients. *Sci Rep,* 7**,** 4951.

R Core Team 2022. R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/.

Ramirez, J. & Hagman, J. 2009. The Mi-2/NuRD complex: a critical epigenetic regulator of hematopoietic development, differentiation and cancer. *Epigenetics,* 4**,** 532–6.

Rath, A., Glibowicka, M., Nadeau, V. G., Chen, G. & Deber, C. M. 2009. Detergent binding explains anomalous SDS-PAGE migration of membrane proteins. *Proc Natl Acad Sci U S A,* 106**,** 1760–5.

Recaldin, T., Steinacher, L., Gjeta, B., Harter, M. F., Adam, L., Kromer, K., Mendes, M. P., Bellavista, M., Nikolaev, M., Lazzaroni, G., Krese, R., Kilik, U., Popovic, D., Stoll, B., Gerard, R., Bscheider, M., Bickle, M., Cabon, L., Camp, J. G. & Gjorevski, N. 2024. Human organoids with an autologous tissue-resident immune compartment. *Nature,* 633**,** 165–173.

Reid, X. J., Low, J. K. K. & Mackay, J. P. 2023. A NuRD for all seasons. *Trends Biochem Sci,* 48**,** 11–25.

Reis, E. S., Mastellos, D. C., Ricklin, D., Mantovani, A. & Lambris, J. D. 2018. Complement in cancer: untangling an intricate relationship. *Nat Rev Immunol,* 18**,** 5–18.

Ren, X., Yang, H., Nierenberg, J. L., Sun, Y., Chen, J., Beaman, C., Pham, T., Nobuhara, M., Takagi, M. A., Narayan, V., Li, Y., Ziv, E. & Shen, Y. 2023. High-throughput PRIME-editing screens identify functional DNA variants in the human genome. *Mol Cell,* 83**,** 4633–4645 e9.

Ribic, C. M., Sargent, D. J., Moore, M. J., Thibodeau, S. N., French, A. J., Goldberg, R. M., Hamilton, S. R., Laurent-Puig, P., Gryfe, R., Shepherd, L. E., Tu, D., Redston, M. & Gallinger, S. 2003. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *N Engl J Med,* 349**,** 247–57.

Roche, A. E., Bassett, B. J., Samant, S. A., Hong, W., Blobel, G. A. & Svensson, E. C. 2008. The zinc finger and C-terminal domains of MTA proteins are required for FOG-2-mediated transcriptional repression via the NuRD complex. *J Mol Cell Cardiol,* 44**,** 352–60.

Rodrigues, C. H. M., Pires, D. E. V. & Ascher, D. B. 2021. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci,* 30**,** 60–69.

Roelands, J., Kuppen, P. J. K., Ahmed, E. I., Mall, R., Masoodi, T., Singh, P., Monaco, G., Raynaud, C., De Miranda, N., Ferraro, L., Carneiro-Lobo, T. C., Syed, N., Rawat, A., Awad, A., Decock, J., Mifsud, W., Miller, L. D., Sherif, S., Mohamed, M. G., Rinchai, D., Van Den Eynde, M., Sayaman, R. W., Ziv, E., Bertucci, F., Petkar, M. A., Lorenz, S., Mathew, L. S., Wang, K., Murugesan, S., Chaussabel, D., Vahrmeijer, A. L., Wang, E., Ceccarelli, A., Fakhro, K. A., Zoppoli, G., Ballestrero, A., Tollenaar, R., Marincola, F. M., Galon, J., Khodor, S. A., Ceccarelli, M., Hendrickx, W. & Bedognetti, D. 2023. An integrated tumor, immune and microbiome atlas of colon cancer. *Nat Med,* 29**,** 1273–1286.

Roelands, J., Kuppen, P. J. K., Ahmed, E. I., Mall, R., Masoodi, T., Singh, P., Monaco, G., Raynaud, C., De Miranda, N., Ferraro, L., Carneiro-Lobo, T. C., Syed, N., Rawat, A., Awad, A., Decock, J., Mifsud, W., Miller, L. D., Sherif, S., Mohamed, M. G., Rinchai, D., Van Den Eynde, M., Sayaman, R. W., Ziv, E., Bertucci, F., Petkar, M. A., Lorenz, S., Mathew, L. S., Wang, K., Murugesan, S., Chaussabel, D., Vahrmeijer, A. L., Wang, E., Ceccarelli, A., Fakhro, K. A., Zoppoli, G., Ballestrero, A., Tollenaar, R., Marincola, F. M., Galon, J., Khodor, S. A., Ceccarelli, M., Hendrickx, W. & Bedognetti, D. 2023. *Supplementary Data AC-ICAM* [Online]. Figshare. Available: https://doi.org/10.6084/m9.figshare.16944775.v1 [Accessed].

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. 2015. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell,* 160**,** 48–61.

Rossi, A., Kontarakis, Z., Gerri, C., Nolte, H., Holper, S., Kruger, M. & Stainier, D. Y. 2015. Genetic compensation induced by deleterious mutations but not gene knockdowns. *Nature,* 524**,** 230–3.

Rusilowicz-Jones, E. V., Urbe, S. & Clague, M. J. 2022. Protein degradation on the global

scale. *Mol Cell,* 82**,** 1414–1423.

Russo, R., Russo, V., Cecere, F., Valletta, M., Gentile, M. T., Colucci-D'amato, L., Angelini, C., Riccio, A., Pedone, P. V., Chambery, A. & Baglivo, I. 2021. ZBTB2 protein is a new partner of the Nucleosome Remodeling and Deacetylase (NuRD) complex. *Int J Biol Macromol,* 168**,** 67–76.

Rutledge, W. C., Kong, J., Gao, J., Gutman, D. A., Cooper, L. A., Appin, C., Park, Y., Scarpace, L., Mikkelsen, T., Cohen, M. L., Aldape, K. D., Mclendon, R. E., Lehman, N. L., Miller, C. R., Schniederjan, M. J., Brennan, C. W., Saltz, J. H., Moreno, C. S. & Brat, D. J. 2013. Tumor-infiltrating lymphocytes in glioblastoma are associated with specific genomic alterations and related to transcriptional class. *Clin Cancer Res,* 19**,** 4951–60.

Saathoff, H., Brofelth, M., Trinh, A., Parker, B. L., Ryan, D. P., Low, J. K., Webb, S. R., Silva, A. P., Mackay, J. P. & Shepherd, N. E. 2015. A peptide affinity reagent for isolating an intact and catalytically active multi-protein complex from mammalian cells. *Bioorg Med Chem,* 23**,** 960–5.

Saez-Ibanez, A. R., Upadhaya, S. & Campbell, J. 2023. Immuno-oncology clinical trials take a turn beyond PD1/PDL1 inhibitors. *Nat Rev Drug Discov,* 22**,** 442–443.

Sakata, N., Hemmi, K., Kawaguchi, M., Miura, Y., Noguchi, S., Ma, D., Sasahara, M., Kato, T., Hori, M. & Tamaoki, T. 2000. The mouse ZFH-4 protein contains four homeodomains and twenty-two zinc fingers. *Biochem Biophys Res Commun,* 273**,** 686–93.

Saleh, R., Sasidharan Nair, V., Toor, S. M., Taha, R. Z., Murshed, K., Al-Dhaheri, M., Khawar, M., Petkar, M. A., Abu Nada, M., Al-Ejeh, F. & Elkord, E. 2020. Differential gene expression of tumor-infiltrating CD8(+) T cells in advanced versus early-stage colorectal cancer and identification of a gene signature of poor prognosis. *J Immunother Cancer,* 8.

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K. R., Zhao, T., Batiste, R., Van Arnam, J., Cancer Genome Atlas Research, N., Shmulevich, I., Rao, A. U. K., Lazar, A. J., Sharma, A. & Thorsson, V. 2018. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep,* 23**,** 181–193 e7.

Samstein, R. M., Lee, C. H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., Barron, D. A., Zehir, A., Jordan, E. J., Omuro, A., Kaley, T. J., Kendall, S. M., Motzer, R. J., Hakimi, A. A., Voss, M. H., Russo, P., Rosenberg, J., Iyer, G., Bochner, B. H., Bajorin, D. F., Al-Ahmadie, H. A., Chaft, J. E., Rudin, C. M., Riely, G. J., Baxi, S., Ho, A. L., Wong, R. J., Pfister, D. G., Wolchok, J. D., Barker, C. A., Gutin, P. H., Brennan, C. W., Tabar, V., Mellinghoff, I. K., Deangelis, L. M., Ariyan, C. E., Lee, N., Tap, W. D., Gounder, M. M., D'angelo, S. P., Saltz, L., Stadler, Z. K., Scher, H. I., Baselga, J., Razavi, P., Klebanoff, C. A., Yaeger, R., Segal, N. H., Ku, G. Y., Dematteo, R. P., Ladanyi, M., Rizvi, N. A., Berger, M. F., Riaz, N., Solit, D. B., Chan, T. A. & Morris, L. G. T. 2019. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet,* 51**,** 202–206.

Sansom, O. J., Berger, J., Bishop, S. M., Hendrich, B., Bird, A. & Clarke, A. R. 2003. Deficiency of Mbd2 suppresses intestinal tumorigenesis. *Nat Genet,* 34**,** 145–7.

Sansom, O. J., Maddison, K. & Clarke, A. R. 2007. Mechanisms of disease: methyl-binding domain proteins as potential therapeutic targets in cancer. *Nat Clin Pract Oncol,* 4**,**

305–15.

Sato, T., Stange, D. E., Ferrante, M., Vries, R. G., Van Es, J. H., Van Den Brink, S., Van Houdt, W. J., Pronk, A., Van Gorp, J., Siersema, P. D. & Clevers, H. 2011. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology,* 141**,** 1762–72.

Saul, D., Kosinsky, R. L., Atkinson, E. J., Doolittle, M. L., Zhang, X., Lebrasseur, N. K., Pignolo, R. J., Robbins, P. D., Niedernhofer, L. J., Ikeno, Y., Jurk, D., Passos, J. F., Hickson, L. J., Xue, A., Monroe, D. G., Tchkonia, T., Kirkland, J. L., Farr, J. N. & Khosla, S. 2022. A new gene set identifies senescent cells and predicts senescence-associated pathways across tissues. *Nat Commun,* 13**,** 4827.

Saunders, M. P., Graham, J., Cunningham, D., Plummer, R., Church, D., Kerr, R., Cook, S., Zheng, S., La Thangue, N. & Kerr, D. 2022. CXD101 and nivolumab in patients with metastatic microsatellite-stable colorectal cancer (CAROSELL): a multicentre, open-label, single-arm, phase II trial. *ESMO Open,* 7**,** 100594.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P. & Cardona, A. 2012. Fiji: an open-source platform for biological-image analysis. *Nat Methods,* 9**,** 676–82.

Schoen, R. E., Razzak, A., Yu, K. J., Berndt, S. I., Firl, K., Riley, T. L. & Pinsky, P. F. 2015. Incidence and mortality of colorectal cancer in individuals with a family history of colorectal cancer. *Gastroenterology,* 149**,** 1438–1445 e1.

Schreiber, R. D., Old, L. J. & Smyth, M. J. 2011. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science,* 331**,** 1565–70.

Schreurs, R., Baumdick, M. E., Drewniak, A. & Bunders, M. J. 2021. In vitro co-culture of human intestinal organoids and lamina propria-derived CD4(+) T cells. *STAR Protoc,* 2**,** 100519.

Schug, J., Schuller, W. P., Kappen, C., Salbaum, J. M., Bucan, M. & Stoeckert, C. J., Jr. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol,* 6**,** R33.

Schumacher, T. N., Scheper, W. & Kvistborg, P. 2019. Cancer Neoantigens. *Annu Rev Immunol,* 37**,** 173–200.

Senbabaoglu, Y., Michailidis, G. & Li, J. Z. 2014. Critical limitations of consensus clustering in class discovery. *Sci Rep,* 4**,** 6207.

Shankaran, V., Ikeda, H., Bruce, A. T., White, J. M., Swanson, P. E., Old, L. J. & Schreiber, R. D. 2001. IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature,* 410**,** 1107–11.

Shao, S., Cao, H., Wang, Z., Zhou, D., Wu, C., Wang, S., Xia, D. & Zhang, D. 2020. CHD4/NuRD complex regulates complement gene expression and correlates with CD8 T cell infiltration in human hepatocellular carcinoma. *Clin Epigenetics,* 12**,** 31.

Shapiro, I. E. & Bassani-Sternberg, M. 2023. The impact of immunopeptidomics: From basic research to clinical implementation. *Semin Immunol,* 66**,** 101727.

Sharifi Tabar, M., Mackay, J. P. & Low, J. K. K. 2019. The stoichiometry and interactome of the Nucleosome Remodeling and Deacetylase (NuRD) complex are conserved across multiple cell lines. *FEBS J,* 286**,** 2043–2061.

Sharma, P. & Allison, J. P. 2015. The future of immune checkpoint therapy. *Science,* 348**,** 56–61.

Shen, J., Choi, Y. L., Lee, T., Kim, H., Chae, Y. K., Dulken, B. W., Bogdan, S., Huang, M., Fisher, G. A., Park, S., Lee, S. H., Hwang, J. E., Chung, J. H., Kim, L., Song, H., Pereira, S., Shin, S., Lim, Y., Ahn, C. H., Kim, S., Oum, C., Kim, S., Park, G., Song, S., Jung, W., Kim, S., Bang, Y. J., Mok, T. S. K., Ali, S. M. & Ock, C. Y. 2024. Inflamed immune phenotype predicts favorable clinical outcomes of immune checkpoint inhibitor therapy across multiple cancer types. *J Immunother Cancer,* 12.

Shen, P., Zeng, D., Lin, Z., Lin, P., Luo, M., Zhao, L., Chai, H., Huang, L. & Xiao, M. 2022. ZFHX4 mutation as a novel predictive biomarker for immune checkpoint blockade in non-small cell lung cancer. *Journal of Clinical Oncology,* 40.

Sheng, W., Lafleur, M. W., Nguyen, T. H., Chen, S., Chakravarthy, A., Conway, J. R., Li, Y., Chen, H., Yang, H., Hsu, P. H., Van Allen, E. M., Freeman, G. J., De Carvalho, D. D., He, H. H., Sharpe, A. H. & Shi, Y. 2018. LSD1 Ablation Stimulates Anti-tumor Immunity and Enables Checkpoint Blockade. *Cell,* 174**,** 549–563 e19.

Sheng, W., Liu, Y., Chakraborty, D., Debo, B. & Shi, Y. 2021. Simultaneous Inhibition of LSD1 and TGFbeta Enables Eradication of Poorly Immunogenic Tumors with Anti-PD-1 Treatment. *Cancer Discov,* 11**,** 1970–1981.

Sher, F., Hossain, M., Seruggia, D., Schoonenberg, V. a. C., Yao, Q., Cifani, P., Dassama, L. M. K., Cole, M. A., Ren, C., Vinjamur, D. S., Macias-Trevino, C., Luk, K., Mcguckin, C., Schupp, P. G., Canver, M. C., Kurita, R., Nakamura, Y., Fujiwara, Y., Wolfe, S. A., Pinello, L., Maeda, T., Kentsis, A., Orkin, S. H. & Bauer, D. E. 2019. Rational targeting of a NuRD subcomplex guided by comprehensive in situ mutagenesis. *Nat Genet,* 51**,** 1149–1159.

Skoufou-Papoutsaki, N., Adler, S., D'santos, P., Mannion, L., Mehmed, S., Kemp, R., Smith, A., Perrone, F., Nayak, K., Russell, A., Zilbauer, M. & Winton, D. J. 2023. Efficient genetic editing of human intestinal organoids using ribonucleoprotein-based CRISPR. *Dis Model Mech,* 16.

Smits, A. H., Jansen, P. W., Poser, I., Hyman, A. A. & Vermeulen, M. 2013. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res,* 41**,** e28.

Snow, J. W., Kim, J., Currie, C. R., Xu, J. & Orkin, S. H. 2010. Sumoylation regulates interaction of FOG1 with C-terminal-binding protein (CTBP). *J Biol Chem,* 285**,** 28064–75.

So, T. & Ishii, N. 2019. The TNF-TNFR Family of Co-signal Molecules. *Adv Exp Med Biol,* 1189**,** 53–84.

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I. & Forbes, S. A. 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer,* 18**,** 696–705.

Sondka, Z., Dhir, N. B., Carvalho-Silva, D., Jupe, S., Madhumita, Mclaren, K., Starkey, M.,

Ward, S., Wilding, J., Ahmed, M., Argasinska, J., Beare, D., Chawla, M. S., Duke, S., Fasanella, I., Neogi, A. G., Haller, S., Hetenyi, B., Hodges, L., Holmes, A., Lyne, R., Maurel, T., Nair, S., Pedro, H., Sangrador-Vegas, A., Schuilenburg, H., Sheard, Z., Yong, S. Y. & Teague, J. 2024. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res,* 52**,** D1210–D1217.

Sorrentino, U., Fedrigo, M., Calo, A. P., Perin, M., Veronese, P. & Salviati, L. 2024. ZFHX4 truncating variant and orofacial clefting. *Am J Med Genet A,* 194**,** 115–116.

Sosinsky, A., Ambrose, J., Cross, W., Turnbull, C., Henderson, S., Jones, L., Hamblin, A., Arumugam, P., Chan, G., Chubb, D., Noyvert, B., Mitchell, J., Walker, S., Bowman, K., Pasko, D., Buongermino Pereira, M., Volkova, N., Rueda-Martin, A., Perez-Gil, D., Lopez, J., Pullinger, J., Siddiq, A., Zainy, T., Choudhury, T., Yavorska, O., Fowler, T., Bentley, D., Kingsley, C., Hing, S., Deans, Z., Rendon, A., Hill, S., Caulfield, M. & Murugaesu, N. 2024. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat Med,* 30**,** 279–289.

Spranger, S., Bao, R. & Gajewski, T. F. 2015. Melanoma-intrinsic beta-catenin signalling prevents anti-tumour immunity. *Nature,* 523**,** 231–5.

Spruijt, C. G., Bartels, S. J., Brinkman, A. B., Tjeertes, J. V., Poser, I., Stunnenberg, H. G. & Vermeulen, M. 2010. CDK2AP1/DOC-1 is a bona fide subunit of the Mi-2/NuRD complex. *Mol Biosyst,* 6**,** 1700–6.

Spruijt, C. G., Luijsterburg, M. S., Menafra, R., Lindeboom, R. G., Jansen, P. W., Edupuganti, R. R., Baltissen, M. P., Wiegant, W. W., Voelker-Albert, M. C., Matarese, F., Mensinga, A., Poser, I., Vos, H. R., Stunnenberg, H. G., Van Attikum, H. & Vermeulen, M. 2016. ZMYND8 Co-localizes with NuRD on Target Genes and Regulates Poly(ADP-Ribose)-Dependent Recruitment of GATAD2A/NuRD to Sites of DNA Damage. *Cell Rep,* 17**,** 783–798.

Stacey, S. N., Helgason, H., Gudjonsson, S. A., Thorleifsson, G., Zink, F., Sigurdsson, A., Kehr, B., Gudmundsson, J., Sulem, P., Sigurgeirsson, B., Benediktsdottir, K. R., Thorisdottir, K., Ragnarsson, R., Fuentelsaz, V., Corredera, C., Gilaberte, Y., Grasa, M., Planelles, D., Sanmartin, O., Rudnai, P., Gurzau, E., Koppova, K., Nexo, B. A., Tjonneland, A., Overvad, K., Jonasson, J. G., Tryggvadottir, L., Johannsdottir, H., Kristinsdottir, A. M., Stefansson, H., Masson, G., Magnusson, O. T., Halldorsson, B. V., Kong, A., Rafnar, T., Thorsteinsdottir, U., Vogel, U., Kumar, R., Nagore, E., Mayordomo, J. I., Gudbjartsson, D. F., Olafsson, J. H. & Stefansson, K. 2015. New basal cell carcinoma susceptibility loci. *Nat Commun,* 6**,** 6825.

Steele, C. D., Abbasi, A., Islam, S. M. A., Bowes, A. L., Khandekar, A., Haase, K., Hames-Fathi, S., Ajayi, D., Verfaillie, A., Dhami, P., Mclatchie, A., Lechner, M., Light, N., Shlien, A., Malkin, D., Feber, A., Proszek, P., Lesluyes, T., Mertens, F., Flanagan, A. M., Tarabichi, M., Van Loo, P., Alexandrov, L. B. & Pillay, N. 2022. Signatures of copy number alterations in human cancer. *Nature,* 606**,** 984–991.

Stratton, M. R., Campbell, P. J. & Futreal, P. A. 2009. The cancer genome. *Nature,* 458**,** 719–24.

Stubbs, L., Sun, Y. & Caetano-Anolles, D. 2011. *Function and Evolution of C2H2 Zinc Finger Arrays,* Subcelllular Biochemistry, Springer.

Sugawara, T., Miya, F., Ishikawa, T., Lysenko, A., Nishino, J., Kamatani, T., Takemoto, A.,

Boroevich, K. A., Kakimi, K., Kinugasa, Y., Tanabe, M. & Tsunoda, T. 2022. Immune subtypes and neoantigen-related immune evasion in advanced colorectal cancer. *iScience,* 25**,** 103740.

Sun, L., Liu, R., Wu, Z. J., Liu, Z. Y., Wan, A. H., Yan, S., Liu, C., Liang, H., Xiao, M., You, N., Lou, Y., Deng, Y., Bu, X., Chen, D., Huang, J., Zhang, X., Kuang, D. M. & Wan, G. 2024. Galectin-7 Induction by EHMT2 Inhibition Enhances Immunity in Microsatellite Stability Colorectal Cancer. *Gastroenterology,* 166**,** 466–482.

Sun, T., Chen, Y. & Chen, Y. X. 2025. Single-cell and bulk transcriptome analyses reveal elevated amino acid metabolism promoting tumor-directed immune evasion in colorectal cancer. *Front Immunol,* 16**,** 1575829.

Sun, X., Fu, X., Li, J., Xing, C., Martin, D. W., Zhang, H. H., Chen, Z. & Dong, J. T. 2012. Heterozygous deletion of Atbf1 by the Cre-loxP system in mice causes preweaning mortality. *Genesis,* 50**,** 819–27.

Sun, Y., Li, H., Ma, Z., Wang, J., Yang, H., Zhang, X. & Liu, B. 2022. Identification of immune subtypes and their prognosis and molecular implications in colorectal cancer. *PLoS One,* 17**,** e0278114.

Suraweera, N., Duval, A., Reperant, M., Vaury, C., Furlan, D., Leroy, K., Seruca, R., Iacopetta, B. & Hamelin, R. 2002. Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. *Gastroenterology,* 123**,** 1804–11.

Suzuki, H., Watkins, D. N., Jair, K. W., Schuebel, K. E., Markowitz, S. D., Chen, W. D., Pretlow, T. P., Yang, B., Akiyama, Y., Van Engeland, M., Toyota, M., Tokino, T., Hinoda, Y., Imai, K., Herman, J. G. & Baylin, S. B. 2004. Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. *Nat Genet,* 36**,** 417–22.

Synthego. 2019. *Synthego Performance Analysis, ICE Analysis v3.0* [Online]. Available: https://ice.synthego.com/ [Accessed 23 Oct 2023].

Tabula Sapiens, C., Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., Harper, W., Hemenez, M., Ponnusamy, R., Salehi, A., Sanagavarapu, B. A., Spallino, E., Aaron, K. A., Concepcion, W., Gardner, J. M., Kelly, B., Neidlinger, N., Wang, Z., Crasta, S., Kolluru, S., Morri, M., Tan, S. Y., Travaglini, K. J., Xu, C., Alcantara-Hernandez, M., Almanzar, N., Antony, J., Beyersdorf, B., Burhan, D., Calcuttawala, K., Carter, M. M., Chan, C. K. F., Chang, C. A., Chang, S., Colville, A., Culver, R. N., Cvijovic, I., D'amato, G., Ezran, C., Galdos, F. X., Gillich, A., Goodyer, W. R., Hang, Y., Hayashi, A., Houshdaran, S., Huang, X., Irwin, J. C., Jang, S., Juanico, J. V., Kershner, A. M., Kim, S., Kiss, B., Kong, W., Kumar, M. E., Kuo, A. H., Li, B., Loeb, G. B., Lu, W. J., Mantri, S., Markovic, M., Mcalpine, P. L., De Morree, A., Mrouj, K., Mukherjee, S., Muser, T., Neuhofer, P., Nguyen, T. D., Perez, K., Puluca, N., Qi, Z., Rao, P., Raquer-Mckay, H., Schaum, N., Scott, B., Seddighzadeh, B., Segal, J., Sen, S., Sikandar, S., Spencer, S. P., Steffes, L. C., Subramaniam, V. R., Swarup, A., Swift, M., Van Treuren, W., Trimm, E., Veizades, S., Vijayakumar, S., Vo, K. C., Vorperian, S. K., Wang, W., Weinstein, H. N. W., Winkler, J., Wu, T. T. H., Xie, J., Yung, A. R., Zhang, Y., et al. 2022. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science,* 376**,** eabl4896.

Tajima, A., Hess, M. T., Cabrera, B. L., Kolodner, R. D. & Carethers, J. M. 2004. The

mismatch repair complex hMutS alpha recognizes 5-fluorouracil-modified DNA: implications for chemosensitivity and resistance. *Gastroenterology,* 127**,** 1678–84.

Takaku, M., Grimm, S. A., Roberts, J. D., Chrysovergis, K., Bennett, B. D., Myers, P., Perera, L., Tucker, C. J., Perou, C. M. & Wade, P. A. 2018. GATA3 zinc finger 2 mutations reprogram the breast cancer transcriptional network. *Nat Commun,* 9**,** 1059.

Takei, S., Tanaka, Y., Lin, Y. T., Koyama, S., Fukuoka, S., Hara, H., Nakamura, Y., Kuboki, Y., Kotani, D., Kojima, T., Bando, H., Mishima, S., Ueno, T., Kojima, S., Wakabayashi, M., Sakamoto, N., Kojima, M., Kuwata, T., Yoshino, T., Nishikawa, H., Mano, H., Endo, I., Shitara, K. & Kawazoe, A. 2024. Multiomic molecular characterization of the response to combination immunotherapy in MSS/pMMR metastatic colorectal cancer. *J Immunother Cancer,* 12.

Tang, Y. Q., Chen, T. F., Zhang, Y., Zhao, X. C., Zhang, Y. Z., Wang, G. Q., Huang, M. L., Cai, S. L., Zhao, J., Wei, B. & Huang, J. 2020. The tumor immune microenvironment transcriptomic subtypes of colorectal cancer for prognosis and development of precise immunotherapy. *Gastroenterol Rep (Oxf),* 8**,** 381–389.

Taraborrelli, L., Senbabaoglu, Y., Wang, L., Lim, J., Blake, K., Kljavin, N., Gierke, S., Scherl, A., Ziai, J., Mcnamara, E., Owyong, M., Rao, S., Calviello, A. K., Oreper, D., Jhunjhunwala, S., Argiles, G., Bendell, J., Kim, T. W., Ciardiello, F., Wongchenko, M. J., De Sauvage, F. J., De Sousa, E. M. F., Yan, Y., West, N. R. & Murthy, A. 2023. Tumor-intrinsic expression of the autophagy gene Atg16l1 suppresses anti-tumor immunity in colorectal cancer. *Nat Commun,* 14**,** 5945.

Tauriello, D. V. F., Palomo-Ponce, S., Stork, D., Berenguer-Llergo, A., Badia-Ramentol, J., Iglesias, M., Sevillano, M., Ibiza, S., Canellas, A., Hernando-Momblona, X., Byrom, D., Matarin, J. A., Calon, A., Rivas, E. I., Nebreda, A. R., Riera, A., Attolini, C. S. & Batlle, E. 2018. TGFbeta drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature,* 554**,** 538–543.

Thakkar, K., Castro, A., Qian, D., Swanton, C. & Litchfield, K. 2024. Drivers of immunotherapy response within the CPI3000+ cohort. *Cancer Res,* 84.

Thiery, J. P., Acloque, H., Huang, R. Y. & Nieto, M. A. 2009. Epithelial-mesenchymal transitions in development and disease. *Cell,* 139**,** 871–90.

Thompson, J. C., Davis, C., Deshpande, C., Hwang, W. T., Jeffries, S., Huang, A., Mitchell, T. C., Langer, C. J. & Albelda, S. M. 2020. Gene signature of antigen processing and presentation machinery predicts response to checkpoint blockade in non-small cell lung cancer (NSCLC) and melanoma. *J Immunother Cancer,* 8.

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., Ziv, E., Culhane, A. C., Paull, E. O., Sivakumar, I. K. A., Gentles, A. J., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J. S., Mose, L. E., Vo, N. S., Liu, J., Liu, Y., Rader, J., Dhankani, V., Reynolds, S. M., Bowlby, R., Califano, A., Cherniack, A. D., Anastassiou, D., Bedognetti, D., Mokrab, Y., Newman, A. M., Rao, A., Chen, K., Krasnitz, A., Hu, H., Malta, T. M., Noushmehr, H., Pedamallu, C. S., Bullman, S., Ojesina, A. I., Lamb, A., Zhou, W., Shen, H., Choueiri, T. K., Weinstein, J. N., Guinney, J., Saltz, J., Holt, R. A., Rabkin, C. S., Cancer Genome Atlas Research, N., Lazar, A. J., Serody, J. S., Demicco, E. G., Disis, M. L., Vincent, B. G. & Shmulevich, I. 2018. The Immune Landscape of Cancer. *Immunity,* 48**,** 812–830 e14.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A,* 99**,** 6567–72.

Tiwari, P., Kaila, P. & Guptasarma, P. 2019. Understanding anomalous mobility of proteins on SDS-PAGE with special reference to the highly acidic extracellular domains of human E- and N-cadherins. *Electrophoresis,* 40**,** 1273–1281.

Tong, X., Patel, A. S., Kim, E., Li, H., Chen, Y., Li, S., Liu, S., Dilly, J., Kapner, K. S., Zhang, N., Xue, Y., Hover, L., Mukhopadhyay, S., Sherman, F., Myndzar, K., Sahu, P., Gao, Y., Li, F., Li, F., Fang, Z., Jin, Y., Gao, J., Shi, M., Sinha, S., Chen, L., Chen, Y., Kheoh, T., Yang, W., Yanai, I., Moreira, A. L., Velcheti, V., Neel, B. G., Hu, L., Christensen, J. G., Olson, P., Gao, D., Zhang, M. Q., Aguirre, A. J., Wong, K. K. & Ji, H. 2024. Adeno-to-squamous transition drives resistance to KRAS inhibition in LKB1 mutant lung cancer. *Cancer Cell,* 42**,** 413–428 e7.

Torrado, M., Low, J. K. K., Silva, A. P. G., Schmidberger, J. W., Sana, M., Sharifi Tabar, M., Isilak, M. E., Winning, C. S., Kwong, C., Bedward, M. J., Sperlazza, M. J., Williams, D. C., Jr., Shepherd, N. E. & Mackay, J. P. 2017. Refinement of the subunit interaction network within the nucleosome remodelling and deacetylase (NuRD) complex. *FEBS J,* 284**,** 4216–4232.

Tougeron, D., Mouillet, G., Trouilloud, I., Lecomte, T., Coriat, R., Aparicio, T., Des Guetz, G., Lecaille, C., Artru, P., Sickersen, G., Cauchin, E., Sefrioui, D., Boussaha, T., Ferru, A., Matysiak-Budnik, T., Silvain, C., Karayan-Tapon, L., Pages, J. C., Vernerey, D., Bonnetain, F., Michel, P., Taieb, J. & Zaanan, A. 2016. Efficacy of Adjuvant Chemotherapy in Colon Cancer With Microsatellite Instability: A Large Multicenter AGEO Study. *J Natl Cancer Inst,* 108.

Tu, Y., Wu, H., Zhong, C., Liu, Y., Xiong, Z., Chen, S., Wang, J., Wong, P. P., Yang, W., Liang, Z., Lu, J., Chen, S., Zhang, L., Feng, Y., Si-Tou, W. W., Yin, B., Lin, Y., Liang, J., Liang, L., Vong, J. S. L., Ren, W., Kwong, T. T., Leung, H., To, K. F., Ma, S., Tong, M., Sun, H., Xia, Q., Zhou, J., Kerr, D., La Thangue, N., Sung, J. J. Y., Chan, S. L. & Cheng, A. S. 2025. Pharmacological activation of STAT1-GSDME pyroptotic circuitry reinforces epigenetic immunotherapy for hepatocellular carcinoma. *Gut,* 74**,** 613–627.

Tumeh, P. C., Harview, C. L., Yearley, J. H., Shintaku, I. P., Taylor, E. J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V., West, A. N., Carmona, M., Kivork, C., Seja, E., Cherry, G., Gutierrez, A. J., Grogan, T. R., Mateus, C., Tomasic, G., Glaspy, J. A., Emerson, R. O., Robins, H., Pierce, R. H., Elashoff, D. A., Robert, C. & Ribas, A. 2014. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature,* 515**,** 568–71.

Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., Mcgranahan, N., Reading, J. L., Wong, Y. N. S., Rowan, A., Kanu, N., Al Bakir, M., Chambers, T., Salgado, R., Savas, P., Loi, S., Birkbak, N. J., Sansregret, L., Gore, M., Larkin, J., Quezada, S. A. & Swanton, C. 2017. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol,* 18**,** 1009–1021.

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., Von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., Von Heijne, G., Nielsen, J. & Ponten, F. 2015. Tissue-based map of

the human proteome. *Science,* 347**,** 1260419.

Uniprot, C. 2025. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res,* 53**,** D609–D617.

Upadhaya, S., Hubbard-Lucey, V. M. & Yu, J. X. 2020. Immuno-oncology drug development forges on despite COVID-19. *Nat Rev Drug Discov,* 19**,** 751–752.

Usefi, H. 2022. Clustering, multicollinearity, and singular vectors. *Computational Statistics and Data Analysis,* 173.

Usset, J., Rosendahl Huber, A., Andrianova, M. A., Batlle, E., Carles, J., Cuppen, E., Elez, E., Felip, E., Gomez-Rey, M., Lo Giacco, D., Martinez-Jimenez, F., Munoz-Couselo, E., Siu, L. L., Tabernero, J., Vivancos, A., Muinos, F., Gonzalez-Perez, A. & Lopez-Bigas, N. 2024. Five latent factors underlie response to immunotherapy. *Nat Genet,* 56**,** 2112–2120.

Uwccc Flow Cytometry Laboratory. 2016. *FlowJo for Antibody Titrations: Separation Index and Concatenation* [Online]. Available: https://cancer.wisc.edu/research/wp-content/uploads/2017/03/Flow_TechNotes_FlowJo-Antibody-Titrations_20170918.pdf [Accessed].

Vaes, N., Idris, M., Boesmans, W., Alves, M. M. & Melotte, V. 2022. Nerves in gastrointestinal cancer: from mechanism to modulations. *Nat Rev Gastroenterol Hepatol,* 19**,** 768–784.

Vales, S., Bacola, G., Biraud, M., Touvron, M., Bessard, A., Geraldo, F., Dougherty, K. A., Lashani, S., Bossard, C., Flamant, M., Duchalais, E., Marionneau-Lambot, S., Oullier, T., Oliver, L., Neunlist, M., Vallette, F. M. & Van Landeghem, L. 2019. Tumor cells hijack enteric glia to activate colon cancer stem cells and stimulate tumorigenesis. *EBioMedicine,* 49**,** 172–188.

Van De Wetering, M., Francies, H. E., Francis, J. M., Bounova, G., Iorio, F., Pronk, A., Van Houdt, W., Van Gorp, J., Taylor-Weiner, A., Kester, L., Mclaren-Douglas, A., Blokker, J., Jaksani, S., Bartfeld, S., Volckman, R., Van Sluis, P., Li, V. S., Seepo, S., Sekhar Pedamallu, C., Cibulskis, K., Carter, S. L., Mckenna, A., Lawrence, M. S., Lichtenstein, L., Stewart, C., Koster, J., Versteeg, R., Van Oudenaarden, A., Saez-Rodriguez, J., Vries, R. G., Getz, G., Wessels, L., Stratton, M. R., Mcdermott, U., Meyerson, M., Garnett, M. J. & Clevers, H. 2015. Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell,* 161**,** 933–45.

Vasaikar, S., Huang, C., Wang, X., Petyuk, V. A., Savage, S. R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O. A., Gritsenko, M. A., Zimmerman, L. J., Mcdermott, J. E., Clauss, T. R., Moore, R. J., Zhao, R., Monroe, M. E., Wang, Y. T., Chambers, M. C., Slebos, R. J. C., Lau, K. S., Mo, Q., Ding, L., Ellis, M., Thiagarajan, M., Kinsinger, C. R., Rodriguez, H., Smith, R. D., Rodland, K. D., Liebler, D. C., Liu, T., Zhang, B. & Clinical Proteomic Tumor Analysis, C. 2019. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell,* 177**,** 1035–1049 e19.

Vasen, H. F., Watson, P., Mecklin, J. P. & Lynch, H. T. 1999. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology,* 116**,** 1453–6.

Verstappen, G., Van Grunsven, L. A., Michiels, C., Van De Putte, T., Souopgui, J., Van Damme, J., Bellefroid, E., Vandekerckhove, J. & Huylebroeck, D. 2008. Atypical

Mowat-Wilson patient confirms the importance of the novel association between ZFHX1B/SIP1 and NuRD corepressor complex. *Hum Mol Genet,* 17**,** 1175–83.

Voabil, P., De Bruijn, M., Roelofsen, L. M., Hendriks, S. H., Brokamp, S., Van Den Braber, M., Broeks, A., Sanders, J., Herzig, P., Zippelius, A., Blank, C. U., Hartemink, K. J., Monkhorst, K., Haanen, J., Schumacher, T. N. & Thommen, D. S. 2021. An ex vivo tumor fragment platform to dissect response to PD-1 blockade in cancer. *Nat Med,* 27**,** 1250–1261.

Vogelstein, B. & Kinzler, K. W. 2004. Cancer genes and the pathways they control. *Nat Med,* 10**,** 789–99.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr. & Kinzler, K. W. 2013. Cancer genome landscapes. *Science,* 339**,** 1546–58.

Von Locquenghien, M., Rozalen, C. & Celia-Terrassa, T. 2021. Interferons in cancer immunoediting: sculpting metastasis and immunotherapy response. *J Clin Invest,* 131.

Vulto-Van Silfhout, A. T., Hehir-Kwa, J. Y., Van Bon, B. W., Schuurs-Hoeijmakers, J. H., Meader, S., Hellebrekers, C. J., Thoonen, I. J., De Brouwer, A. P., Brunner, H. G., Webber, C., Pfundt, R., De Leeuw, N. & De Vries, B. B. 2013. Clinical significance of de novo and inherited copy-number variation. *Hum Mutat,* 34**,** 1679–87.

Waldman, A. D., Fritz, J. M. & Lenardo, M. J. 2020. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat Rev Immunol,* 20**,** 651–668.

Wang, C., Gu, Y., Zhang, K., Xie, K., Zhu, M., Dai, N., Jiang, Y., Guo, X., Liu, M., Dai, J., Wu, L., Jin, G., Ma, H., Jiang, T., Yin, R., Xia, Y., Liu, L., Wang, S., Shen, B., Huo, R., Wang, Q., Xu, L., Yang, L., Huang, X., Shen, H., Sha, J. & Hu, Z. 2016. Systematic identification of genes with a cancer-testis expression pattern in 19 cancer types. *Nat Commun,* 7**,** 10499.

Wang, C., Yan, J., Yin, P., Gui, L., Ji, L., Ma, B. & Gao, W. Q. 2020. beta-Catenin inhibition shapes tumor immunity and synergizes with immunotherapy in colorectal cancer. *Oncoimmunology,* 9**,** 1809947.

Wang, H., Li, S., Wang, Q., Jin, Z., Shao, W., Gao, Y., Li, L., Lin, K., Zhu, L., Wang, H., Liao, X. & Wang, D. 2021. Tumor immunological phenotype signature-based high-throughput screening for the discovery of combination immunotherapy compounds. *Sci Adv,* 7.

Wang, H., Zhou, Y., Zhang, Y., Fang, S., Zhang, M., Li, H., Xu, F., Liu, L., Liu, J., Zhao, Q. & Wang, F. 2022. Subtyping of microsatellite stability colorectal cancer reveals guanylate binding protein 2 (GBP2) as a potential immunotherapeutic target. *J Immunother Cancer,* 10.

Wang, J., Xiu, J., Farrell, A., Baca, Y., Arai, H., Battaglin, F., Kawanishi, N., Soni, S., Zhang, W., Millstein, J., Shields, A. F., Grothey, A., Weinberg, B. A., Marshall, J. L., Lou, E., Khushman, M., Sohal, D. P. S., Hall, M. J., Liu, T., Oberley, M., Spetzler, D., Korn, W. M., Shen, L. & Lenz, H. J. 2023. Mutational analysis of microsatellite-stable gastrointestinal cancer with high tumour mutational burden: a retrospective cohort study. *Lancet Oncol,* 24**,** 151–161.

Wang, P., Chen, Y. & Wang, C. 2021. Beyond Tumor Mutation Burden: Tumor Neoantigen Burden as a Biomarker for Immunotherapy and Other Types of Therapy. *Front Oncol,*

11**,** 672677.

Wang, S., He, Z., Wang, X., Li, H. & Liu, X. S. 2019. Antigen presentation and tumor immunogenicity in cancer immunotherapy response prediction. *Elife,* 8.

Wang, Y., Li, C., Wang, Z., Wang, Z., Wu, R., Wu, Y., Song, Y. & Liu, H. 2022. Comparison between immunotherapy efficacy in early non-small cell lung cancer and advanced non-small cell lung cancer: a systematic review. *BMC Med,* 20**,** 426.

Wang, Y., Qiu, X., Li, Q., Qin, J., Ye, L., Zhang, X., Huang, X., Wen, X., Wang, Z., He, W., Di, Y. & Zhou, Q. 2025. Single-cell and spatial-resolved profiling reveals cancer-associated fibroblast heterogeneity in colorectal cancer metabolic subtypes. *J Transl Med,* 23**,** 175.

Wang, Y., Zhang, H., Chen, Y., Sun, Y., Yang, F., Yu, W., Liang, J., Sun, L., Yang, X., Shi, L., Li, R., Li, Y., Zhang, Y., Li, Q., Yi, X. & Shang, Y. 2009. LSD1 is a subunit of the NuRD complex and targets the metastasis programs in breast cancer. *Cell,* 138**,** 660–72.

Ward-Kavanagh, L. K., Lin, W. W., Sedy, J. R. & Ware, C. F. 2016. The TNF Receptor Superfamily in Co-stimulating and Co-inhibitory Responses. *Immunity,* 44**,** 1005–19.

Watanabe, T., Wu, T. T., Catalano, P. J., Ueki, T., Satriano, R., Haller, D. G., Benson, A. B., 3rd & Hamilton, S. R. 2001. Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *N Engl J Med,* 344**,** 1196–206.

Weinberg, R. A. 1994. Oncogenes and tumor suppressor genes. *CA Cancer J Clin,* 44**,** 160–70.

Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. A., Kang, G. H., Widschwendter, M., Weener, D., Buchanan, D., Koh, H., Simms, L., Barker, M., Leggett, B., Levine, J., Kim, M., French, A. J., Thibodeau, S. N., Jass, J., Haile, R. & Laird, P. W. 2006. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet,* 38**,** 787–93.

Weng, Y., Feng, Y., Li, Z., Xu, S., Wu, D., Huang, J., Wang, H. & Wang, Z. 2024. Zfp260 choreographs the early stage osteo-lineage commitment of skeletal stem cells. *Nat Commun,* 15**,** 10186.

Werren, E. A., Guxholli, A., Jones, N., Wagner, M., Hannibal, I., Granadillo, J. L., Tyndall, A. V., Moccia, A., Kuehl, R., Levandoski, K. M., Day-Salvatore, D. L., Wheeler, M., University of Washington Center for Mendelian, G., Chong, J. X., Bamshad, M. J., Innes, A. M., Pierson, T. M., Mackay, J. P., Bielas, S. L. & Martin, D. M. 2023. De novo variants in GATAD2A in individuals with a neurodevelopmental disorder: GATAD2A-related neurodevelopmental disorder. *HGG Adv,* 4**,** 100198.

Westendorp, B. F., Buller, N., Karpus, O. N., Van Dop, W. A., Koster, J., Versteeg, R., Koelink, P. J., Snel, C. Y., Meisner, S., Roelofs, J., Uhmann, A., Ver Loren Van Themaat, E., Heijmans, J., Hahn, H., Muncan, V., Wildenberg, M. E. & Van Den Brink, G. R. 2018. Indian Hedgehog Suppresses a Stromal Cell-Driven Intestinal Immune Response. *Cell Mol Gastroenterol Hepatol,* 5**,** 67–82 e1.

Westermann, L., Li, Y., Gocmen, B., Niedermoser, M., Rhein, K., Jahn, J., Cascante, I., Scholer, F., Moser, N., Neubauer, B., Hofherr, A., Behrens, Y. L., Gohring, G., Kottgen,

A., Kottgen, M. & Busch, T. 2022. Wildtype heterogeneity contributes to clonal variability in genome edited cells. *Sci Rep,* 12**,** 18211.

Whyte, W. A., Bilodeau, S., Orlando, D. A., Hoke, H. A., Frampton, G. M., Foster, C. T., Cowley, S. M. & Young, R. A. 2012. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature,* 482**,** 221–5.

Wiel, L., Baakman, C., Gilissen, D., Veltman, J. A., Vriend, G. & Gilissen, C. 2019. MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. *Hum Mutat,* 40**,** 1030–1038.

Wilkerson, M. D. & Hayes, D. N. 2010. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics,* 26**,** 1572–3.

Wolchok, J. D., Chiarion-Sileni, V., Rutkowski, P., Cowey, C. L., Schadendorf, D., Wagstaff, J., Queirolo, P., Dummer, R., Butler, M. O., Hill, A. G., Postow, M. A., Gaudy-Marqueste, C., Medina, T., Lao, C. D., Walker, J., Marquez-Rodas, I., Haanen, J., Guidoboni, M., Maio, M., Schoffski, P., Carlino, M. S., Sandhu, S., Lebbe, C., Ascierto, P. A., Long, G. V., Ritchings, C., Nassar, A., Askelson, M., Benito, M. P., Wang, W., Hodi, F. S., Larkin, J. & Checkmate, I. 2025. Final, 10-Year Outcomes with Nivolumab plus Ipilimumab in Advanced Melanoma. *N Engl J Med,* 392**,** 11–22.

World Health Organisation 2019. *International statistical classification of diseases and related health problems: 10th revision*, World Health Organisation.

Wu, J., Chen, W., Zhou, Y., Chi, Y., Hua, X., Wu, J., Gu, X., Chen, S. & Zhou, Z. 2023. TSNAdb v2.0: The Updated Version of Tumor-specific Neoantigen Database. *Genomics Proteomics Bioinformatics,* 21**,** 259–266.

Wu, L., Bai, R., Zhang, Y., Chen, H., Wu, J., Chen, Z., Wang, H. & Zhao, L. 2024. METTL3-VISTA axis-based combination immunotherapy for APC truncation colorectal cancer. *J Immunother Cancer,* 12.

Wu, R., Fang, J., Liu, M., A, J., Liu, J., Chen, W., Li, J., Ma, G., Zhang, Z., Zhang, B., Fu, L. & Dong, J. T. 2020. SUMOylation of the transcription factor ZFHX3 at Lys-2806 requires SAE1, UBC9, and PIAS2 and enhances its stability and function in cell proliferation. *J Biol Chem,* 295**,** 6741–6753.

Wu, X., Lin, H. & Li, S. 2019. Prognoses of different pathological subtypes of colorectal cancer at different stages: A population-based retrospective cohort study. *BMC Gastroenterol,* 19**,** 164.

Wu, X., Yan, H., Qiu, M., Qu, X., Wang, J., Xu, S., Zheng, Y., Ge, M., Yan, L. & Liang, L. 2023. Comprehensive characterization of tumor microenvironment in colorectal cancer via molecular analysis. *Elife,* 12.

Wu, Y., Zhong, A., Sidharta, M., Kim, T. W., Ramirez, B., Persily, B., Studer, L. & Zhou, T. 2024. Robust and inducible genome editing via an all-in-one prime editor in human pluripotent stem cells. *Nat Commun,* 15**,** 10824.

Xia, W., Mao, Q., Chen, B., Wang, L., Ma, W., Liang, Y., Zhang, T., Dong, G., Xu, L. & Jiang, F. 2019. The TWIST1-centered competing endogenous RNA network promotes proliferation, invasion, and migration of lung adenocarcinoma. *Oncogenesis,* 8**,** 62.

Xiao, A., Li, X., Wang, C., Ye, J. & Fakih, M. 2024. Updated survival outcome of regorafenib,

ipilimumab, and nivolumab in refractory microsatellite stable non-liver metastatic colorectal cancer: A phase I nonrandomized clinical trial. *Eur J Cancer,* 213**,** 115111.

Xu, C., Prete, M., Webb, S., Jardine, L., Stewart, B. J., Hoo, R., He, P., Meyer, K. B. & Teichmann, S. A. 2023. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell,* 186**,** 5876–5891 e20.

Xu, J., Lamouille, S. & Derynck, R. 2009. TGF-beta-induced epithelial to mesenchymal transition. *Cell Res,* 19**,** 156–72.

Xu, M., Chang, J., Wang, W., Wang, X., Wang, X., Weng, W., Tan, C., Zhang, M., Ni, S., Wang, L., Huang, Z., Deng, Z., Li, W., Huang, D. & Sheng, W. 2022. Classification of colon adenocarcinoma based on immunological characterizations: Implications for prognosis and immunotherapy. *Front Immunol,* 13**,** 934083.

Xu, T., Fang, Y., Gu, Y., Xu, D., Hu, T., Yu, T., Xu, Y. Y., Shen, H. Y., Ma, P. & Shu, Y. 2024. HDAC inhibitor SAHA enhances antitumor immunity via the HDAC1/JAK1/FGL1 axis in lung adenocarcinoma. *J Immunother Cancer,* 12.

Xue, Y., Wong, J., Moreno, G. T., Young, M. K., Cote, J. & Wang, W. 1998. NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. *Mol Cell,* 2**,** 851–61.

Yan, C., Yang, H., Chen, L., Liu, R., Shang, W., Yuan, W., Yang, F., Sun, Q. & Xia, L. 2021. Clinical significance of mucinous component in colorectal adenocarcinoma: a propensity score-matched study. *BMC Cancer,* 21**,** 1286.

Yang, J., Antin, P., Berx, G., Blanpain, C., Brabletz, T., Bronner, M., Campbell, K., Cano, A., Casanova, J., Christofori, G., Dedhar, S., Derynck, R., Ford, H. L., Fuxe, J., Garcia De Herreros, A., Goodall, G. J., Hadjantonakis, A. K., Huang, R. Y. J., Kalcheim, C., Kalluri, R., Kang, Y., Khew-Goodall, Y., Levine, H., Liu, J., Longmore, G. D., Mani, S. A., Massague, J., Mayor, R., Mcclay, D., Mostov, K. E., Newgreen, D. F., Nieto, M. A., Puisieux, A., Runyan, R., Savagner, P., Stanger, B., Stemmler, M. P., Takahashi, Y., Takeichi, M., Theveneau, E., Thiery, J. P., Thompson, E. W., Weinberg, R. A., Williams, E. D., Xing, J., Zhou, B. P., Sheng, G. & Association, E. M. T. I. 2020. Guidelines and definitions for research on epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol,* 21**,** 341–352.

Yang, J., Horton, J. R., Liu, B., Corces, V. G., Blumenthal, R. M., Zhang, X. & Cheng, X. 2023. Structures of CTCF-DNA complexes including all 11 zinc fingers. *Nucleic Acids Res,* 51**,** 8447–8462.

Yang, L., Wang, X., Zhou, X., Chen, H., Song, S., Deng, L., Yao, Y. & Yin, X. 2025. A tunable human intestinal organoid system achieves controlled balance between self-renewal and differentiation. *Nat Commun,* 16**,** 315.

Yayli, G., Bernardini, A., Mendoza Sanchez, P. K., Scheer, E., Damilot, M., Essabri, K., Morlet, B., Negroni, L., Vincent, S. D., Timmers, H. T. M. & Tora, L. 2023. ATAC and SAGA co-activator complexes utilize co-translational assembly, but their cellular localization properties and functions are distinct. *Cell Rep,* 42**,** 113099.

Yi, H., Li, G., Long, Y., Liang, W., Cui, H., Zhang, B., Tan, Y., Li, Y., Shen, L., Deng, D., Tang, Y., Mao, C., Tian, S., Cai, Y., Zhu, Q., Hu, Y., Chen, W. & Fang, L. 2020. Integrative multi-omics analysis of a colon cancer cell line with heterogeneous Wnt activity revealed RUNX2 as an epigenetic regulator of EMT. *Oncogene,* 39**,** 5152–

5164.

Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Trevino, V., Shen, H., Laird, P. W., Levine, D. A., Carter, S. L., Getz, G., Stemke-Hale, K., Mills, G. B. & Verhaak, R. G. 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun,* 4**,** 2612.

Yu, X., He, T., Tong, Z., Liao, L., Huang, S., Fakhouri, W. D., Edwards, D. P. & Xu, J. 2023. Molecular mechanisms of TWIST1-regulated transcription in EMT and cancer metastasis. *EMBO Rep,* 24**,** e56902.

Zambelli, F., Mastropasqua, F., Picardi, E., D'erchia, A. M., Pesole, G. & Pavesi, G. 2018. RNentropy: an entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Res,* 46**,** e46.

Zapata, L., Caravagna, G., Williams, M. J., Lakatos, E., Abduljabbar, K., Werner, B., Chowell, D., James, C., Gourmet, L., Milite, S., Acar, A., Riaz, N., Chan, T. A., Graham, T. A. & Sottoriva, A. 2023. Immune selection determines tumor antigenicity and influences response to checkpoint inhibitors. *Nat Genet,* 55**,** 451–460.

Zaretsky, J. M., Garcia-Diaz, A., Shin, D. S., Escuin-Ordinas, H., Hugo, W., Hu-Lieskovan, S., Torrejon, D. Y., Abril-Rodriguez, G., Sandoval, S., Barthly, L., Saco, J., Homet Moreno, B., Mezzadra, R., Chmielowski, B., Ruchalski, K., Shintaku, I. P., Sanchez, P. J., Puig-Saus, C., Cherry, G., Seja, E., Kong, X., Pang, J., Berent-Maoz, B., Comin-Anduix, B., Graeber, T. G., Tumeh, P. C., Schumacher, T. N., Lo, R. S. & Ribas, A. 2016. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N Engl J Med,* 375**,** 819–29.

Zegerman, P., Canas, B., Pappin, D. & Kouzarides, T. 2002. Histone H3 lysine 4 methylation disrupts binding of nucleosome remodeling and deacetylase (NuRD) repressor complex. *J Biol Chem,* 277**,** 11621–4.

Zeve, D., Stas, E., De Sousa Casal, J., Mannam, P., Qi, W., Yin, X., Dubois, S., Shah, M. S., Syverson, E. P., Hafner, S., Karp, J. M., Carlone, D. L., Ordovas-Montanes, J. & Breault, D. T. 2022. Robust differentiation of human enteroendocrine cells from intestinal stem cells. *Nat Commun,* 13**,** 261.

Zhang, J. & Zhang, Q. 2019. Using Seahorse Machine to Measure OCR and ECAR in Cancer Cells. *Methods Mol Biol,* 1928**,** 353–363.

Zhang, M., Du, S., Ou, H., Cui, R., Jiang, N., Lin, Y., Ge, R., Ma, D. & Zhang, J. 2021. Ablation of Zfhx4 results in early postnatal lethality by disrupting the respiratory center in mice. *J Mol Cell Biol,* 13**,** 210–224.

Zhang, N. & Bevan, M. J. 2011. CD8(+) T cells: foot soldiers of the immune system. *Immunity,* 35**,** 161–8.

Zhang, N., Ng, A. S., Cai, S., Li, Q., Yang, L. & Kerr, D. 2021. Novel therapeutic strategies: targeting epithelial-mesenchymal transition in colorectal cancer. *Lancet Oncol,* 22**,** e358–e368.

Zhang, T., Wei, G., Millard, C. J., Fischer, R., Konietzny, R., Kessler, B. M., Schwabe, J. W. R. & Brockdorff, N. 2018. A variant NuRD complex containing PWWP2A/B excludes MBD2/3 to regulate transcription at active genes. *Nat Commun,* 9**,** 3798.

Zhang, X., Liu, X., Zhou, W., Du, Q., Yang, M., Ding, Y. & Hu, R. 2021. Blockade of IDO-Kynurenine-AhR Axis Ameliorated Colitis-Associated Colon Cancer via Inhibiting Immune Tolerance. *Cell Mol Gastroenterol Hepatol,* 12**,** 1179–1199.

Zhang, Y. P., Guo, Z. Q., Cai, X. T., Rong, Z. X., Fang, Y., Chen, J. Q., Zhuang, K. M., Ruan, M. J., Ma, S. C., Lin, L. Y., Han, D. D., Li, Y. S., Wang, Y. Y., Wang, J., Cao, C. H., Tang, X. R., Xie, Q. K., Chen, Y., Lin, Y., Tan, J. L., Yu, Z. H., Wu, Z. N., Wei, W., Zheng, D. Y., Zeng, Y. J., Ruan, Y. C., Xu, Z. P., Gu, J. Z., Xiao, L. S., Liu, L., Guan, J., Bai, X., Wu, D. H. & Dong, Z. Y. 2025. PAI-1-driven SFRP2(high) cancer-associated fibroblasts hijack the abscopal effect of radioimmunotherapy. *Cancer Cell,* 43**,** 856–874 e9.

Zheng, X., Gao, Y., Yu, C., Fan, G., Li, P., Zhang, M., Yu, J. & Xu, M. 2021. Identification of immune-related subtypes of colorectal cancer to improve antitumor immunotherapy. *Sci Rep,* 11**,** 19432.

Zhou, G., Lieshout, R., Van Tienderen, G. S., De Ruiter, V., Van Royen, M. E., Boor, P. P. C., Magre, L., Desai, J., Koten, K., Kan, Y. Y., Ge, Z., Campos Carrascosa, L., Geuijen, C., Sprengers, D., Van Der Laan, L. J. W., Verstegen, M. M. A. & Kwekkeboom, J. 2022. Modelling immune cytotoxicity for cholangiocarcinoma with tumour-derived organoids and effector T cells. *Br J Cancer,* 127**,** 649–660.

Zhou, Y., Zhang, Y., Li, M., Ming, T., Zhang, C., Huang, C., Li, J., Li, F., Li, H., Zhao, E., Shu, F., Liu, L., Pan, X., Gao, Y., Tian, L., Song, L., Huang, H. & Liao, W. 2024. Oncogenic KRAS drives immunosuppression of colorectal cancer by impairing DDX60-mediated dsRNA accumulation and viral mimicry. *Sci Immunol,* 9**,** eado8758.

Zhou, Z., Zhang, P., Hu, X., Kim, J., Yao, F., Xiao, Z., Zeng, L., Chang, L., Sun, Y. & Ma, L. 2017. USP51 promotes deubiquitination and stabilization of ZEB1. *Am J Cancer Res,* 7**,** 2020–2031.

Zhu, D., Fang, J., Li, Y. & Zhang, J. 2009. Mbd3, a component of NuRD/Mi-2 complex, helps maintain pluripotency of mouse embryonic stem cells by repressing trophectoderm differentiation. *PLoS One,* 4**,** e7684.

Zong, S., Xu, P. P., Xu, Y. H. & Guo, Y. 2022. A bioinformatics analysis: ZFHX4 is associated with metastasis and poor survival in ovarian cancer. *J Ovarian Res,* 15**,** 90.

Supplementary materials

Supplementary Methods, Notes, Figures and Tables have been ordered by respective chapter, denoted by the header on each page.

---Chapter Two---

Supplementary Methods 2.1	Calculation of proliferation rates
Supplementary Methods 2.2	Quantification of protein in cytokine arrays
Supplementary Methods 2.3	Quantification of western blotting
Supplementary Methods 2.4	Other bioinformatic pipelines employed throughout this thesis
Supplementary Figure S2.1	Calculation of doubling times from normalised viable cell number data obtained from CTG luminescence
Supplementary Table 2.1	Summary of accessed intestine-based single-cell datasets for healthy tissue investigation

---Chapter Three---

Supplementary Table 3.1	Linear and logistic regression results on genes identified following mutation comparison of Imm^{high} and Imm^{low} groups in TCGA-COAD all-sample and nHM cohorts as well as an independent pan-cancer immunotherapy treated cohort
Supplementary Table 3.2	Statistical summaries produced from TCGA-COAD survival data with respect to ICS group
Supplementary Table 3.3	Statistical analyses performed on TCGA-COAD survival data with respect to first comparator
Supplementary Table 3.4	Top 20 positive and negative PC loadings of PC1, PC2 and PC3 following PCA of the scaled clustering matrix used for identifying ICS groups
Supplementary Table 3.5	Mutation frequencies of the top 15 genes in each ICS group
Supplementary Figure S3.1	Tumour microenvironment deconvolution of TCGA-COAD samples coloured by ICS group and segregated by (A) immune-adaptive, (B) immune-innate and (C) stromal cell types

---Chapter Four---

Supplementary Note 4.1	Antibody optimisation for IHC
Supplementary Note 4.2	A comment on available genome editing technologies
Supplementary Note 4.3	Migration and invasion assays
Supplementary Figure S4.1	Optimisation of staining ZFHX4 protein expression by IHC using small intestine tissue (n=1)
Supplementary Figure S4.2	Schematics describing Transwell (A) and 3D Matrigel Droplet (B) assays to measure cell line migration and invasion

Supplementary Figure S4.3 Initial optimisation of chemoattractants used for migration and invasion assays of COLO320DM, NCIH716 and SKNMC cells

Supplementary Figure S4.4 Microscope images of crystal-violet stained Transwell-insert membranes following a Transwell migration assay

Supplementary Figure S4.5 Quantification of the Transwell migration assay

Supplementary Figure S4.6 *ZFHX4* expression in healthy intestine scRNAseq datasets

Supplementary Figure S4.7 ZFHX4 protein expression across healthy adult human tissues, with a specific focus on the colon and small intestine

Supplementary Figure S4.8 ZFHX4 protein expression across multiple cancer tissues, with a specific focus on COAD

Supplementary Figure S4.9 IHC controls across matched normal colon, small intestine and COAD tissue from two patients

Supplementary Figure S4.10 Proportions of protein-coding (PC) and non-protein-coding (non-PC) *ZFHX4* mRNA transcripts detected in the chosen cell lines for ZFHX4 study

Supplementary Figure S4.11 Genomic validation of polyclonal *ZFHX4*$^{-/-}$ cell lines by sanger sequencing

Supplementary Figure S4.12 The correlation of *ZFHX4* expression at the transcript-level with indel size and indel type following CRISPR-mediated gDNA disruption

Supplementary Figure S4.13 Gene set enrichment analysis (GSEA) results of interesting pathways in ZFHX4 monoKO cell lines and respective controls

Supplementary Table 4.1 Top 30 upregulated and downregulated genes in the metastatic organoid nosgRNA vs noCRISPR comparison, with respect to nosgRNA and ordered by t-statistic

Supplementary Table 4.2 Top 30 upregulated and downregulated genes in the metastatic organoid *ZFHX4*$^{-/-}$ vs nosgRNA comparison, with respect to *ZFHX4*$^{-/-}$

---Chapter Five---

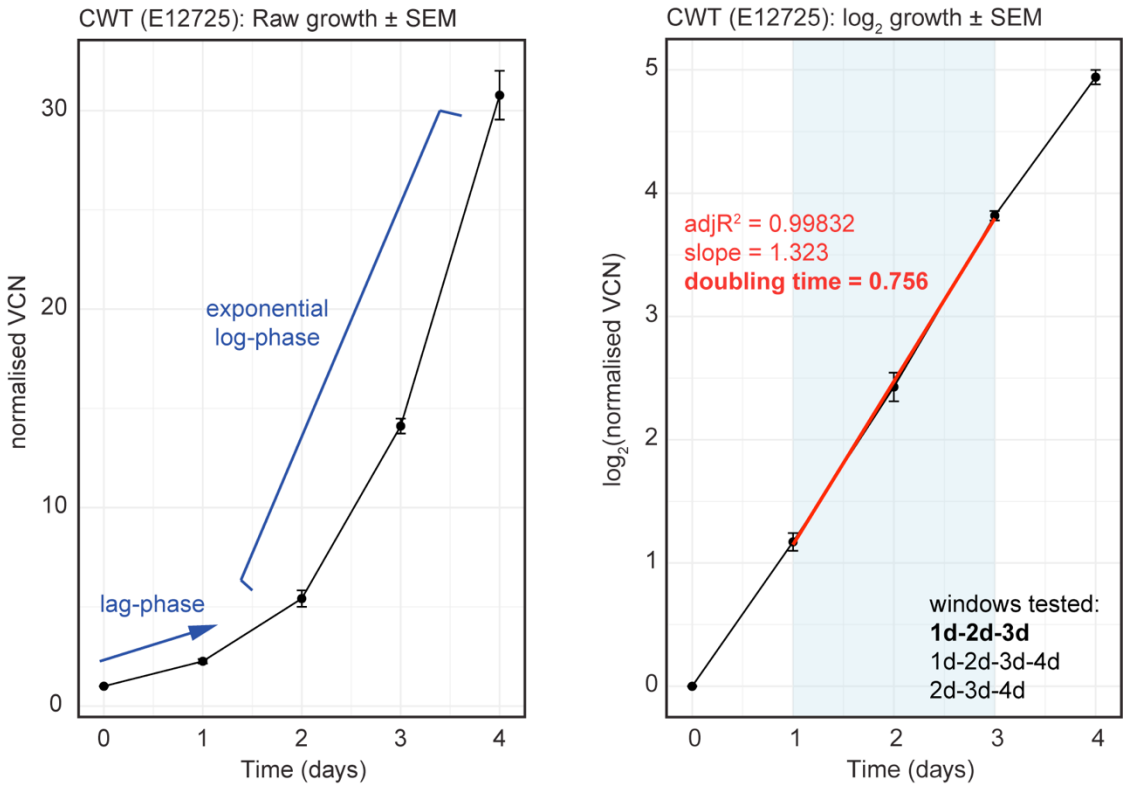
Supplementary Figure S5.1 Channel spillover of single-stained controls before compensation

Supplementary Methods 2.1: Calculation of proliferation rates

For measurements of proliferation rates, identical plates representing different time points were prepared analogous to Section 2.4.1.1, plating 1000 cells in 100ul respective media per well and maintaining thorough resuspension of master cell suspensions. Media-only controls were added to every plate and an extra plate was prepared to represent a day-0 timepoint (d_0), usually read following an overnight rest and setting baseline viable cell numbers. As multiple experiments on various samples of various timepoints were performed during the project, 'best exponential growth' per sample within one experiment was investigated to allow comparison across experiments and statistical analysis. Firstly, normalised VCN at timepoint d_x was calculated per replicate as follows:

$$normalised\ VCN_{d_x} = \frac{VCN_{d_x}}{mean\ VCN_{d_0}}$$

Normalised VCN values were averaged at each time point and SEM was calculated. For visualisation, normalised VCN was plotted across time. For calculation of growth rates per sample using R-programming, normalised VCN was $\log_2$-transformed and the contiguous time window during which the $\log_2$-transformed signal increased most linearly was identified, representing the exponential growth phase. This "best exponential window" was selected based on the highest adjusted R^2 from linear regression of $\log_2$(normalised VCN) versus time, testing windows of ≥3 timepoints. Slopes from these regressions (representing the rate of doubling in $\log_2$ space) were converted into doubling times (representing days taken for the cell population to double). d_0 was excluded from the fitting to avoid underestimation of growth due to the initial attachment phase but included in the plots to anchor the curves. For experiments with only two post-d_0 timepoints, the slope was calculated and flagged as low-confidence. This approach allowed for consistent comparison of growth rates between samples, while accounting for variability in plating, lag phases and the limited linear range of exponential growth. This method is depicted in Supplementary Figure S2.1 using CWT from experiment E12725 as an example.



Supplementary Figure S2.1 | Calculation of doubling times from normalised viable cell number data obtained from CTG luminescence. CWT line from experiment E12725 used an example. The linear growth curve on the left was log_2-transformed to allow identification of the 'best exponential growth' from multiple regression fittings. Doubling time is in days.

Supplementary Methods 2.2: Quantification of protein in cytokine arrays

For HRP signal and thus protein quantification, Fiji (Fiji is just ImageJ, v2.16.0; (Schindelin et al., 2012)) and the "Protein Array Analyser" macro (v1.1.b; (Carpentier, 2010)) were employed. This macro is also available to view on my GitHub Repository (Flower, 2025). Alongside base settings of a radius of 5 with no background subtraction, quantification was performed as follows:

$$normalised\ PI_i\ (NPI) = \frac{PI_i - meanPI(negative\ control\ spots)}{meanPI(positive\ reference\ spots)}$$

where PI_i is raw pixel intensity (PI) of dot i area calculated by the macro. Duplicate NPI values were averaged and proteins with averaged values greater than the maximum NPI

of negative controls and a visible dot on the array were considered expressed. Fold change and delta normalised pixel intensity (dNPI) were calculated for each protein as:

$$\Delta NPI = \frac{NPI\ (KO)}{NPI\ (WT)}$$

where WT is respective to the KO sample. For data visualisation, only dNPI values > 0.01 were considered and heatmaps were plotted using RStudio (v4.2.2, (R Core Team, 2022)).

Supplementary Methods 2.3: Quantification of western blotting

TIFF images were rotated (with no interpolation), converted to 8-bit grayscale images and inverted on Fiji. Per membrane, region-of-interest (ROI) boxes were drawn around a single band for housekeepers or rough band spread >268kDa for ZFHX4, kept identically sized for quantifying each lane. One ROI was drawn in background to measure background signal. Integrated density (intden) was measured for each ROI and normalised protein signal was quantified as follows:

$$normalised\ intden = \frac{ZFHX4(sample_{intden} - background_{intden})}{Vinculin(sample_{intden} - background_{intden})}$$

Per membrane, WT samples were also run alongside KOs and thus protein expression fold change was calculated as follows:

$$fold\ change = \frac{KO_{normalised\ intden}}{WT_{normalised\ intden}}$$

Supplementary Methods 2.4: Other bioinformatic pipelines employed throughout this thesis

This section is to provide greater detail than to what's been reported in the main results chapters, which do largely remain stand-alone.

Classifying mutational load in TCGA-COAD samples

Representing % mutated microsatellite loci, MSIsensor-score (Ding et al., 2018) and MSImantis-score (Cerami et al., 2012) values were obtained respectively and both are concordant for predicting microsatellite instability (MSI) status (Anthony and Seoighe, 2024). A sample was regarded as microsatellite unstable (MSI) if MSIsensor-score was >3.5 (Niu et al., 2014, Johansen et al., 2019) but if MSIsensor-score wasn't available, a >0.4 threshold for MSImantis-score was used instead (Kautto et al., 2017). For quantifying tumour mutational burden (TMB), MAF SNV counts were divided by the GRCh38 coding exon size (36Mb) (GENCODE rel.43; (Frankish et al., 2023)) as an estimate. TMB thresholding for TMBL (<10mut/Mb) and TMBH (≥10mut/Mb) were set as by a previous study, additionally defining ultra-hypermutant (uHM) cases as ≥100mut/Mb (Campbell et al., 2017).

As clinical mismatch repair (MMR) and polymerase proofreading (PP) status by IHC methodology was not available for these patients, novel scores were produced to predict them based on mutational signatures. Firstly, trinucleotide context was derived using sigminer (a wrapper for SigProfiler) and the number of *de novo* signatures was calculated using Non-Matrix Factorisation (NMF) (n=8 for the 343 cohort), based on >0.9 cophenetic stability and minimal RSS clustering validation measures. SBS mutational signatures were then extracted and analysed using sigminer's sig_fit() function and the GRCh38 reference genome (BSgenome.Hsapiens.UCSC.hg38). MMRd score represents the sum of relative SBS signature contributions (6, 14, 15, 20, 21, 26, 44) associated with MMRd as by COSMIC (Sondka et al., 2024). PPd score represents the sum of POLE (10a, 10b, 14) and POLD1 (10c, 10d, 20) SBS signatures as by the same derivation strategy. MMRd score thresholding was deduced using Receiver Operator Curve (ROC) analysis and Youden's Index using the pROC package (MS status as response, MMRd score as predictor, threshold = 0.4774415) whilst PPd score thresholding was deduced visually using density plots (threshold = 0.6).

Classifying mutational load in non-TCGA-COAD samples

For the three other colon cancer cohorts, TMB was either annotated from an available respective supplementary table or generated *de novo* from the WES data and dividing total SNV counts by reported coding exon size similar to the above. The same thresholds for TMB status and hypermutant phenotype were applied as with TCGA-COAD data, but MMRd and PPd prediction was not performed with these datasets.

Survival analyses

All survival data was downloaded from respective publications except for TCGA-COAD, which was obtained from an integrated TCGA survival dataset (Liu et al., 2018) as by another publication (Thorsson et al., 2018). Kaplan-Meir plots and downstream log-rank and cox proportional hazards testing were performed in R using ggsurvfit, survival and survminer packages.

Chapter Three: The $Imm^{grouping}$ approach

Immune Score was produced by the ConsensusTME package (Jimenez-Sanchez et al., 2019) representing an overall enrichment of TCGA-COAD-specific immune cell population ssGSEA values. The consensusTMEAnalysis function was employed on the HUGOsymbol gene expression matrix, using cancerType = “COAD” and statMethod “ssgsea”. TIL scores were annotated from a previous publication (Saltz et al., 2018) which represent estimates of %TIL infiltrate derived from deep learning on H&E stains. Cytotoxicity scores represented the geometric mean of GZMA and PRF1 expression as by a previous publication (Rooney et al., 2015). Immune genes such as CD8A and PDL1 were obtained from the gene expression matrix. Mutation comparison between groups was performed with maftools. Imm^{high} and Imm^{low} groups were derived from Patient IDs uniquely present in at least 3 top or bottom 15^{th} percentiles respectively of an immune related parameter (Immune Score, Cytotoxicity, TIL Percent, PDL1, CD8A), common biomarkers for immunotherapy response.

Linear regression was performed to test if Gene X was associated with immune score, irrespective of TMB as by the following formula: $lm(Immune_{Score} \sim GeneMutation + TMB)$. Logistic regression was performed to test if Gene X was associated with samples in the Imm^{high} group, irrespective of TMB as by the following formulae:

$$immgroupbin = 1\ if\ Imm^{high}, 0\ if\ Imm^{low},\ glm(immgroupbin \sim GeneMutation + TMB)$$

Logistic regression was also performed in a similar way to test if Gene X was associated with samples in CB group, irrespective of TMB. Regressions were only performed on genes mutated in >5 samples in the respective group. Gene set enrichment analysis (GSEA) for the hot tumour signature (Wang et al., 2021a) was performed using fgsea.

Chapter Four: GTEx and scRNAseq data

GTEx data was accessed and downloaded directly from the GTEx portal (GTEx Consortium, 2013, GTEx Consortium, 2025). Single-cell data (Supplementary Table 2.1) was accessed online using the cellxgene visual portal (C.Z.I.CellScienceProgram et al., 2025). Furthermore, another scRNAseq dataset for normal (n=112864) vs COAD (n=257251) comparisons was also accessed through cellxgene (Pelka et al., 2021).

Supplementary Table 2.1 | Summary of accessed intestine-based single-cell datasets for healthy tissue investigation

Study	Description	N*
Space-Time Gut Atlas (Teichmann-A)[1]	Healthy foetal, paediatric and adult	401 305
PanGI Cell Atlas (Teichmann-B)[2]	Healthy foetal, preterm and adult	1 285 530
HTAN Cell Atlas[3]	Healthy adult (biopsies)	30 340 (set A) 34 008 (set B)
Human Cell Atlas - Intestine (Teichmann-C)[4]	Healthy adult (integration from 4 separate intestine datasets)	263 275
Tabula Sapiens Human Cell Atlas - Intestine[5]	Healthy adult	80 103
Human Cell Landscape - Intestine[6]	Healthy foetal and adult	59 414

**=number of cells post-filtering for healthy tissue only. [1](Elmentaite et al., 2021), [2](Oliver et al., 2024), [3](Chen et al., 2021), [4](Xu et al., 2023), [5](Tabula Sapiens et al., 2022), [6](Han et al., 2020)*

Supplementary Table 3.1 | Linear and logistic regression results on genes identified following mutation comparison of Imm^{high} and Imm^{low} groups in TCGA-COAD all-sample and nHM cohorts as well as an independent pan-cancer immunotherapy treated cohort

All samples in TCGA-COAD cohort (n=341), with Imm^{high} (n=58) and Imm^{low} (n=57)			**All samples in immunotherapy cohort[a]** (n=249)	**Canonical nHM in TCGA-COAD cohort** (n=294), with Imm^{high} (n=48) and Imm^{low} (n=50)			**TMBL samples in immunotherapy cohort[a]** (n=149)
Gene	**Linear Reg.* (Effect size [sig.])**	**Logistic Reg. (Odds Ratio [sig.])**	**Logistic Reg. (Odds Ratio [sig.])**	**Gene**	**Linear Reg.* (Effect size [sig.])**	**Logistic Reg. (Odds Ratio [sig.])**	**Logistic Reg. (Odds Ratio [sig.])**
ZFHX4	0.0173 [0.1342]	1.65 [0.4073]	2.99 **[0.0046]**	**DNAH5**	0.0095 [0.4697]	4.99 [0.0514]	0.85 [0.7946]
FAT4	0.0101 [0.3665]	2.51 [0.0839]	1.82 [0.0965]	**ZFHX4**	0.0161 [0.2192]	4.28 [0.0811]	2.56 [0.2007]
PIK3CA	0.0136 [0.1395]	3.94 [**0.0057**]	2.62 [0.1134]	**FBXW7**	0.0095 [0.4773]	3.10 [0.0761]	NA
OBSCN	-0.0007 [0.9521]	0.78 [0.7312]	1.00 [0.9909]	**MUC16**	0.0213 [0.0563]	3.23 [0.0617]	0.80 [0.6745]
SYNE1	0.0178 [0.0818]	1.46 [0.4742]	2.45 **[0.0308]**	**TTN**	0.0051 [0.5804]	1.69 [0.2359]	0.79 [0.6291]
MUC16	0.0236 **[0.0214]**	1.71 [0.2834]	1.58 [0.1671]	**PIK3CA**	0.0165 [0.0865]	1.63 [0.3087]	2.37 [0.3280]
TTN	0.0154 [0.0810]	1.49 [0.3700]	1.45 [0.2718]	**TP53**	0.0052 [0.5699]	1.31 [0.5506]	1.08 [0.8759]
KRAS	-0.0015 [0.8564]	1.07 [0.8757]	6.97 **[0.0041]**	**KRAS**	-0.0008 [0.9283]	0.96 [0.9296]	7.74 **[0.0233]**
TP53	-0.0009 [0.9203]	0.88 [0.7592]	1.70 [0.1035]	**SYNE1**	0.0133 [0.2374]	0.87 [0.7935]	2.08 [0.3539]
APC	-0.0294 **[0.0014]**	0.63 [0.3055]	1.01 [0.9884]	**APC**	-0.0199 [0.0568]	0.31 **[0.0180]**	1.30 [0.7483]

*Linear regressions were performed with respect to immune score. Logistic regressions were performed with respect to either Imm^{high} or CB. All regression analyses controlled for TMB. Bold indicates significance ($p<0.05$). *effect size values are raw change in outcome per unit of predictor. [a](Miao et al., 2018) [sig. = significance, Reg. = regression, TMB/L = tumour mutation burden/low, CB = clinical benefit]*

Supplementary Table 3.2 | Statistical summaries produced from TCGA-COAD survival data with respect to ICS group

Survival Type	ICS Type	Median survival (years)	Probability of 1-year survival (95%CI)	Probability of 2-year survival (95%CI)	Probability of 5-year survival (95%CI)
OS	**ICS1**	1.83	64% (41-99)	42% (21-87)	42% (21-87)
	ICS2	6.93	91% (80-100)	91% (80-100)	65% (40-100)
	ICS3	5.06	91% (85-98)	86% (78-95)	55% (37-82)
	ICS4	6.78	90% (84-96)	84% (77-92)	68% (56-83)
	ICS5	3.89	81% (70-95)	74% (61-91)	48% (28-83)
PFS	**ICS1**	1.83	60% (36-100)	36% (15-86)	36% (15-86)
	ICS2	NA	90% (79-100)	90% (79-100)	72% (46-100)
	ICS3	NA	88% (82-96)	82% (74-92)	66% (52-82)
	ICS4	5.26	84% (77-91)	68% (59-78)	55% (43-69)
	ICS5	5.96	73% (61-89)	58% (43-79)	51% (34-76)
DSS	**ICS1**	1.83	70% (47-100)	47% (23-94)	47% (23-94)
	ICS2	NA	95% (87-100)	95% (87-100)	95% (87-100)
	ICS3	NA	92% (86-98)	88% (79-96)	72% (52-100)
	ICS4	NA	92% (87-98)	87% (81-94)	82% (73-92)
	ICS5	NA	84% (73-97)	76% (63-93)	60% (40-89)

Supplementary Table 3.3 | Statistical analyses performed on TCGA-COAD survival data with respect to first comparator

Comp.	Survival	Statistical Analysis	HR (95% CI)	p-value
All ICS types	OS	Log-rank	NA	0.130
	PFS		NA	**0.0452**
	DSS		NA	**0.0242**
ICS3 vs ICS5	OS	Log-rank	NA	0.400
		Cox proportional hazards	0.8546 (0.5925 - 1.233)	0.401
	PFS	Log-rank	NA	**0.0486**
		Cox proportional hazards	0.7164 (0.5112 – 1.004)	0.0527
	DSS	Log-rank	NA	0.156
		Cox proportional hazards	0.7309 (0.4709 – 1.134)	0.162
ICS1 vs rest	OS	Cox proportional hazards	2.5714 (1.1091 – 5.9595)	**0.0277**
	PFS		2.0088 (0.8741 – 4.6168)	0.100
	DSS		3.0836 (1.2127 – 7.8370)	**0.018**
ICS2 vs rest	OS		0.8134 (0.3267 – 2.0251)	0.657
	PFS		0.3660 (0.1155 – 1.1604)	0.0878
	DSS		0.2332 (0.0321 – 1.6946)	0.150
ICS3 vs rest	OS		0.9681 (0.5631 – 1.6631)	0.906
	PFS		0.6627 (0.3923 – 1.1191)	0.124
	DSS		0.8203 (0.4037 – 1.6672)	0.584
ICS4 vs rest	OS		0.7315 (0.4482 – 1.1939)	0.211
	PFS		1.1787 (0.7634 – 1.8202)	0.458
	DSS		0.7825 (0.4239 – 1.4445)	0.433
ICS5 vs rest	OS		1.4198 (0.7605 – 2.6504)	0.271
	PFS		1.5449 (0.8945 – 2.6681)	0.119
	DSS		1.8325 (0.9025 – 3.7216)	0.0937

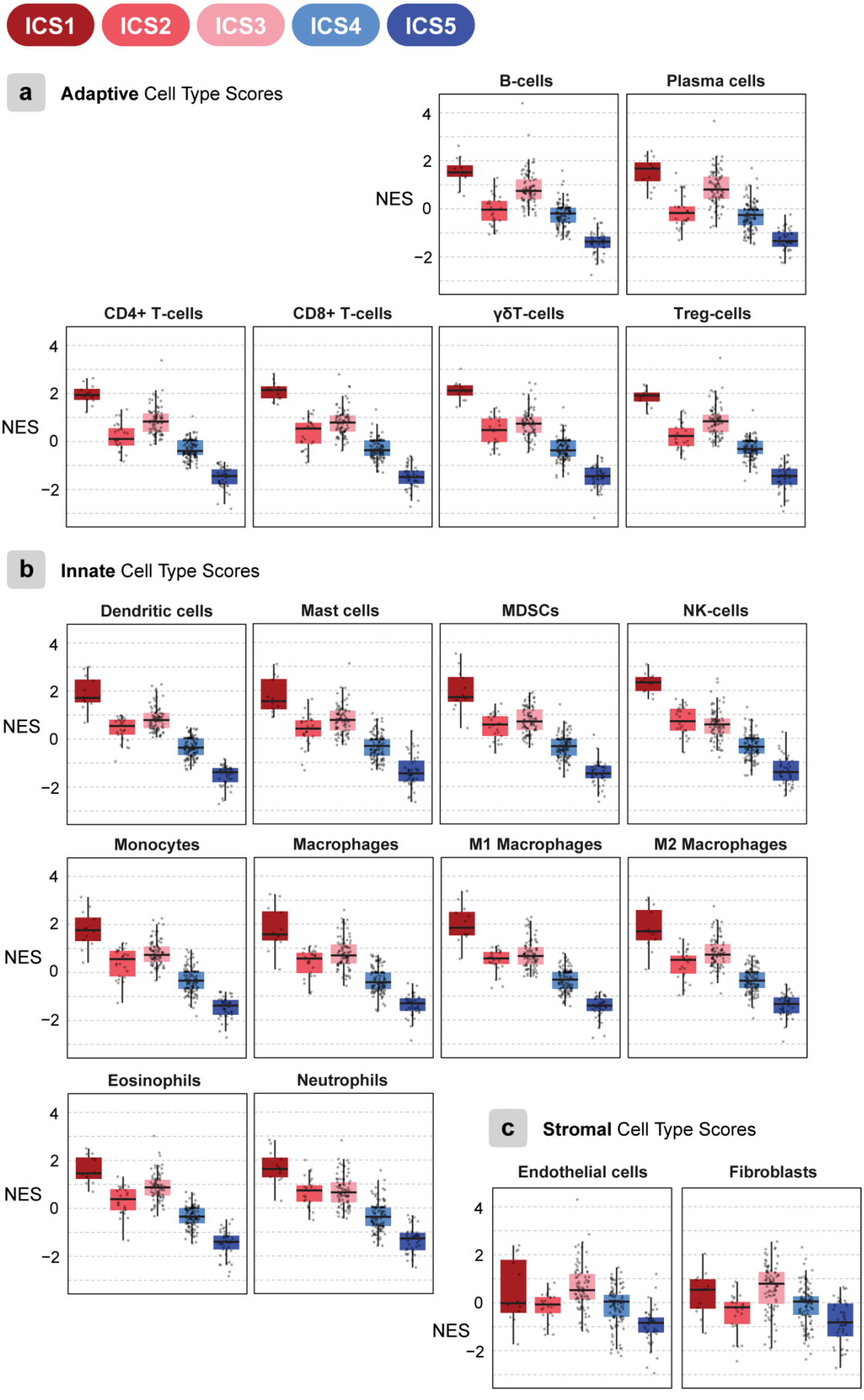
Supplementary Table 3.4 | Top 20 positive and negative PC loadings of PC1, PC2 and PC3 following PCA of the scaled clustering matrix used for identifying ICS groups. Manual internet searches were performed to generally summarise a biological meaning for each positive and negative PC driver and coloured text was used as a key. Loadings rounded to 5 decimal places.

PC	Top 20 positive loadings			Top 20 negative loadings		
PC1	ESTIMATE_immune	0.09672	Immune infiltration of both pro and anti-tumour, programs relating to transcriptional immune activation, antigen presentation and T-cell recruitment/ activation	IL17RE	-0.02987	Genomic instability and aneuploidy which correlates with immune evasion, chemokines involved in monocyte/myeloid recruitment and immune suppression
	Dendritic_cells	0.09533		Frac_Gen_Altered	-0.02785	
	Macrophages_M1	0.09375		IL22RA1	-0.02767	
	T_cells_CD4	0.09340		CXCL14	-0.02739	
	T_cells_CD8	0.09327		GI_Score	-0.02162	
	T_cells_gam_delta	0.09217		IFNLR1	-0.02124	
	TcellGEPsig	0.09207		IL17D	-0.01989	
	Eosinophils	0.09158		IL23R	-0.01974	
	MDSC	0.09109		Ploidy_Score	-0.01947	
	Monocytes	0.09064		CCL15	-0.01719	
	T_regulatory_cells	0.09055		IL17RB	-0.01666	
	IL2RB	0.09010		LAMP1	-0.01527	
	Macrophages_M2	0.08946		VEGFA	-0.01519	
	IL10RA	0.08929		CCL24	-0.01482	
	ESTIMATE_purity	0.08920		IL13RA2	-0.01384	
	HLA-DPB1	0.08905		IL37	-0.01227	
	Macrophages	0.08882		HMGB1	-0.01127	
	hot_tumour_sig	0.08863		IL20RA	-0.01032	
	APM_sig	0.08847		TNFRSF25	-0.01025	
	CCR5	0.08829		CCL14	-0.00958	
PC2	IL34	0.12108	CAF and stromal enriched, ECM remodelling, fibrosis and TGFb signalling, immunosuppression and recruitment of pro-tumour TME (i.e.macrophages)	TIGS_Score	-0.13946	Hypermutation and mutational processes, neoantigen generation, antigen processing and presentation
	TGFBR2	0.11994		MSI_SENS_Score	-0.13151	
	Fibroblasts	0.11868		Frame_Shift_Ins	-0.12188	
	core_mat_sig	0.11229		Frame_Shift_Del	-0.12181	
	EDNRB	0.10792		Splice_Site	-0.11838	
	TNFSF11	0.10685		INDEL_AIM_count	-0.11715	
	ACKR3	0.10551		NLS	-0.11631	
	CXCL14	0.10263		AIM_count	-0.11544	
	CX3CR1	0.10191		pTNB	-0.11511	
	ACKR1	0.10019		MMRd_Score	-0.11175	
	SELP	0.10017		PSME1	-0.10914	
	CCL11	0.09531		PSMB10	-0.10578	
	ESTIMATE_stromal	0.09485		In_Frame_Del	-0.10377	
	Endothelial	0.09388		SNV_AIM_count	-0.10336	
	THBS1	0.09129		Missense_Mut	-0.10323	
	CXCL12	0.09043		TMB	-0.10101	
	primECMsig	0.08870		APM_Score	-0.09681	
	IL1R1	0.08809		INDEL_AIM_rate	-0.09128	
	TNFSF18	0.08607		IL15RA	-0.09107	
	OSMR	0.08490		PSMB9	-0.09104	
PC3	TGFB3	0.14372	Myeloid enriched, fibroblast enriched, ECM remodelling, fibrosis and TGFb signalling, immunosuppression and recruitment of pro-tumour TME (ie macro), mutational processes/ neoantigen generation	CD40LG	-0.14639	Th17/Th22 responses, costimulatory receptors for T/B activation and migration, antigen processing and presentation, Treg recruitment
	CCL7	0.13523		IL26	-0.14156	
	core_mat_sig	0.11218		IL17A	-0.13990	
	Fibroblasts	0.10912		IL17F	-0.11934	
	pTNB	0.10830		TMIGD2	-0.11479	
	NLS	0.10679		CCL28	-0.11418	
	primECMsig	0.10615		TNFRSF17	-0.10795	
	AIM_count	0.10518		CXCR6	-0.10255	
	ESTIMATE_stromal	0.10327		CCL20	-0.10249	
	Splice_Site	0.10240		TNFRSF13B	-0.09561	
	TGFB1	0.10076		HLA-F	-0.09260	
	INDEL_AIM_count	0.10004		CCL15	-0.08339	
	SNV_AIM_count	0.09673		CD27	-0.08321	
	Frame_Shift_Del	0.09643		PSMB10	-0.08178	
	Frame_Shift_Ins	0.09640		TNFRSF14	-0.08047	
	OSMR	0.09559		CCR7	-0.07849	
	CCL2	0.09442		IL22RA2	-0.07744	
	CCL3	0.09378		HLA-A	-0.07743	
	MSI_SENS_Score	0.09269		CXCL14	-0.07661	
	CCL8	0.09076		CCR4	-0.07645	

Supplementary Table 3.5 | Mutation frequencies of the top 15 genes in each ICS group. Values are % (ICS1 n=12, ICS2 n=25, ICS3 n=82, ICS4 n=115, ICS5 n=41)

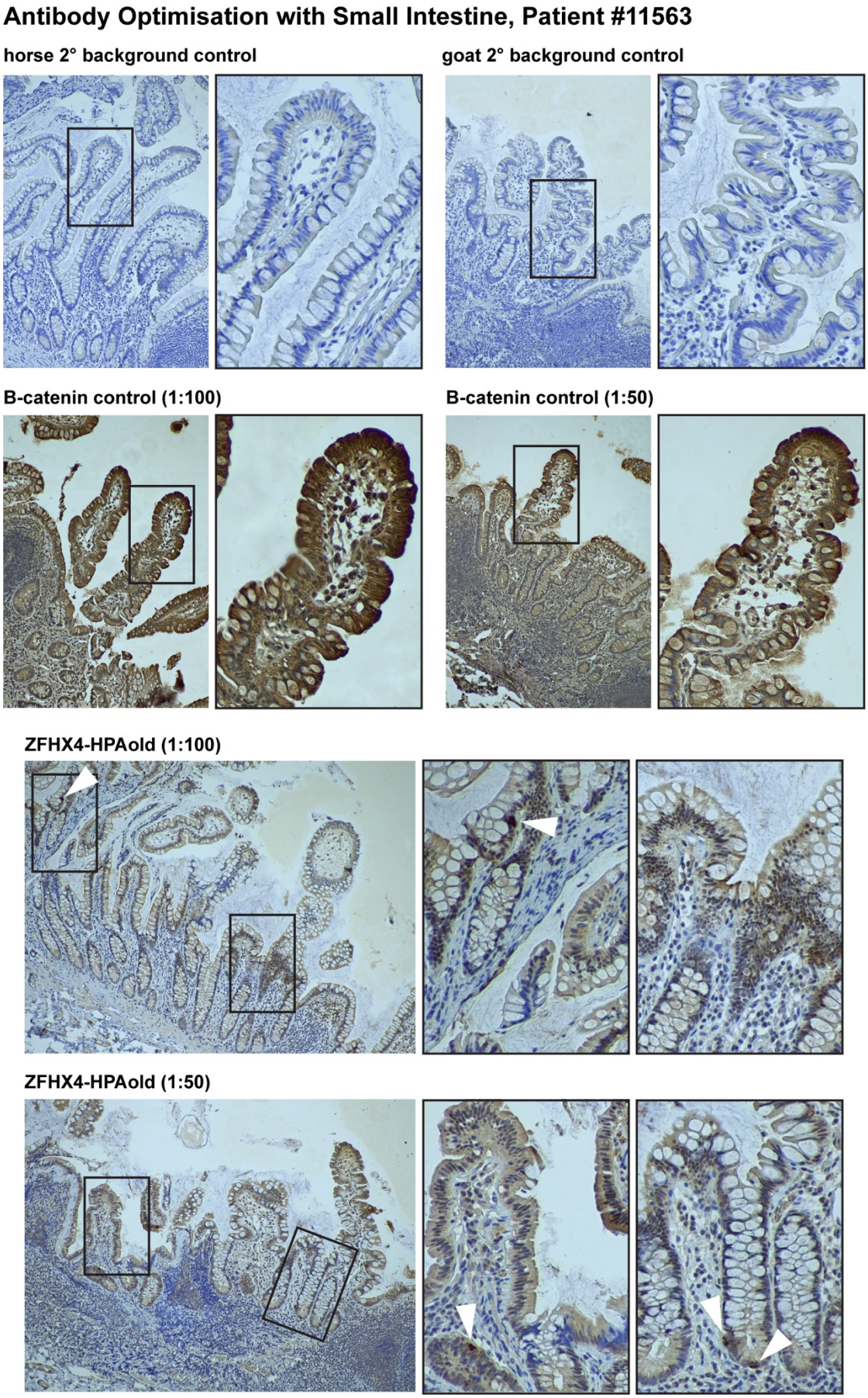
Gene	ICS1	Gene	ICS2	Gene	ICS3	Gene	ICS4	Gene	ICS5
TTN	100	**TTN**	84	**APC**	72	**APC**	84	**APC**	90
KMT2D	83	**RNF43**	76	**TP53**	71	**TP53**	67	**TP53**	68
MXRA5	75	**MUC16**	72	**TTN**	43	**KRAS**	48	**KRAS**	46
DST	75	**OBSCN**	72	**KRAS**	41	**TTN**	37	**TTN**	34
RYR1	75	**ADGRV1**	72	**PIK3CA**	35	**PIK3CA**	25	**SYNE1**	22
MUC5B	75	**SYNE1**	68	**SYNE1**	29	**MUC16**	23	**PIK3CA**	20
PREX2	67	**FAT4**	68	**MUC16**	22	**FAT4**	17	**USH2A**	17
ANKRD11	67	**PCLO**	68	**SOX9**	21	**DNAH5**	17	**DST**	15
CHD9	67	**KMT2B**	68	**FAT4**	17	**ZFHX4**	15	**HECW1**	15
ROBO2	67	**DOCK3**	64	**FBXW7**	16	**OBSCN**	15	**SPAG17**	15
NEB	67	**ZFHX4**	64	**ZFHX4**	16	**SYNE1**	13	**ADGRV1**	15
DNAH3	67	**RYR2**	60	**DNAH5**	16	**AMER1**	13	**FLG**	15
NBEA	67	**DNAH5**	60	**RYR2**	15	**SDK1**	13	**PCLO**	15
USH2A	67	**MACF1**	60	**LRP1B**	15	**LRP1B**	12	**CSMD3**	15
MDB6	67	**HMCN1**	60	**DNAH11**	13	**DSCAM**	12	**DMD**	15

Supplementary Figure S3.1 | Tumour microenvironment deconvolution of TCGA-COAD samples coloured by ICS group and segregated by (A) immune-adaptive, (B) immune-innate and (C) stromal cell types. *[NES = normalised enrichment score from consensusTME]*


ICS1
ICS2
ICS3
ICS4
ICS5
a
Adaptive Cell Type Scores
B-cells
Plasma cells
CD4+ T-cells
CD8+ T-cells
γδT-cells
Treg-cells
NES
b
Innate Cell Type Scores
Dendritic cells
Mast cells
MDSCs
NK-cells
Monocytes
Macrophages
M1 Macrophages
M2 Macrophages
Eosinophils
Neutrophils
c
Stromal Cell Type Scores
Endothelial cells
Fibroblasts

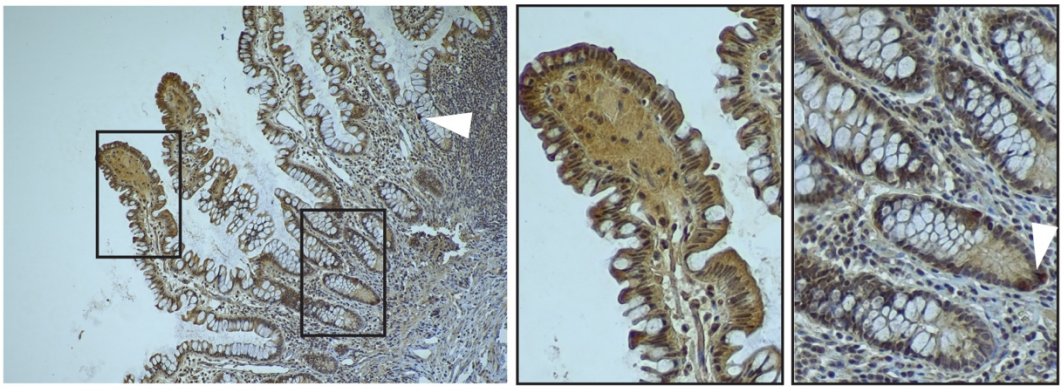
Supplementary Note 4.1: Antibody optimisation for IHC

Small intestine tissue (from patient #11563) was used to optimise IHC with all five commercially available ZFHX4 antibodies at two concentrations (Supplementary Figure S4.1). No-primary controls successfully demonstrated absent background staining and the positive control (ß-catenin staining) demonstrated expected cytoplasmic staining in epithelium. However, all ZFHX4 antibodies displayed divergent patterns (summarised in Table 4.3). At 1:100 and 1:50, staining was similar across antibodies. ZFHX4 localised to nuclei in epithelium and stroma with variable cytoplasmic background (highest for M02/M11, lowest for HPA/Abc). HPA and Abc highlighted a subset of epithelial nuclei with striking black signal, supporting specificity as different epitopes with the same pattern. All antibodies marked stromal nuclei, but M02/M11 showed poor specificity, staining nearly all nuclei and obscuring haematoxylin (HPA/Abc preserved clear counterstain). Patterns differed from Human Protein Atlas images which were stained with concentrated 1:30 HPA antibody suggesting this could be a methodological difference. Overall, the HPA antibody at 1:100 offered the best balance of specificity and reproducibility, avoiding M02/M11 background while confirming nuclear localisation and an interesting subset also identified by abc.

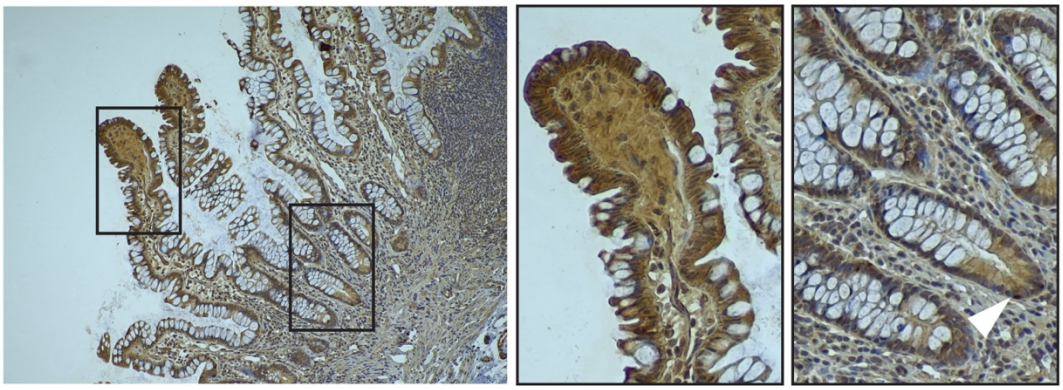

Antibody Optimisation with Small Intestine, Patient #11563
horse 2° background control
goat 2° background control
B-catenin control (1:100)
B-catenin control (1:50)
ZFHX4-HPAold (1:100)
ZFHX4-HPAold (1:50)

Antibody Optimisation with Small Intestine, Patient #11563

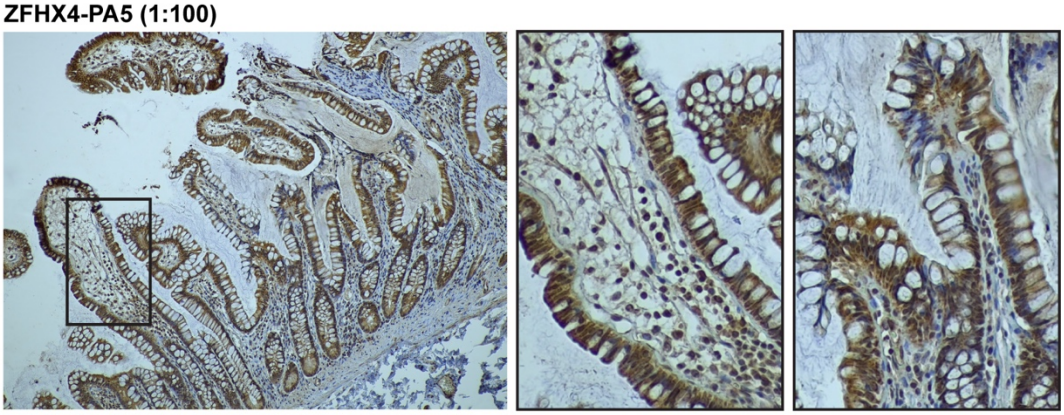
ZFHX4-HPAnew (1:100)



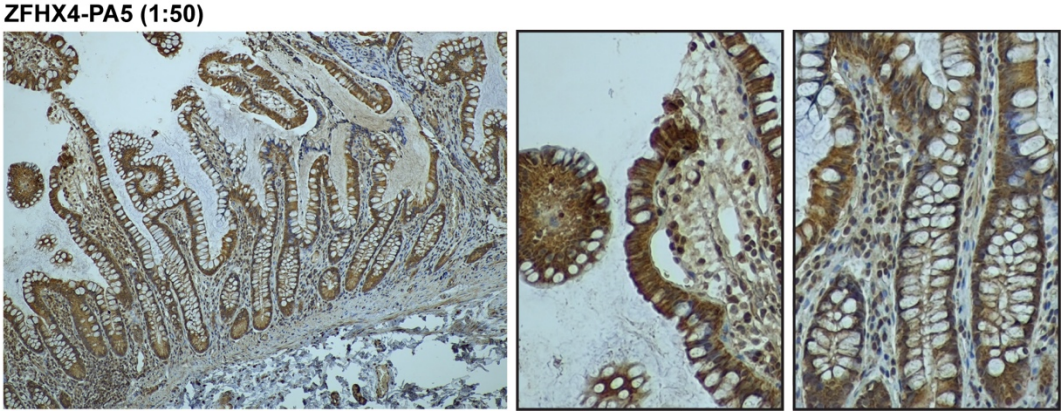
ZFHX4-HPAnew (1:50)



ZFHX4-PA5 (1:100)

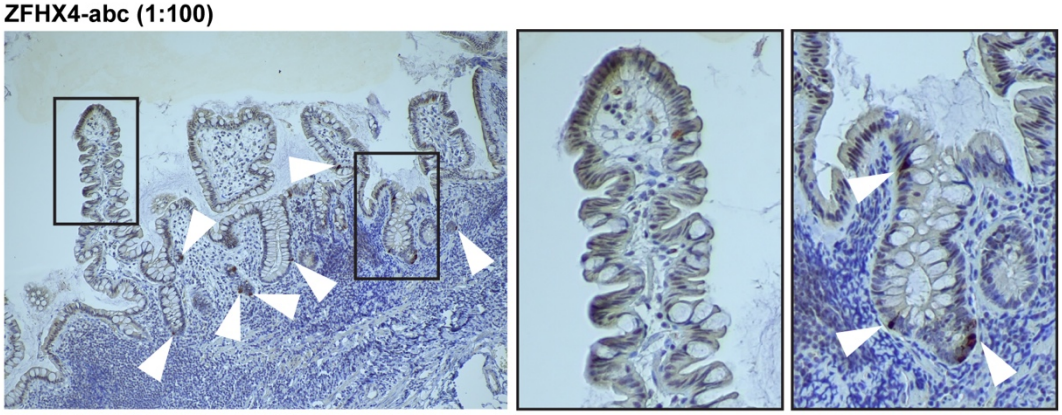


ZFHX4-PA5 (1:50)

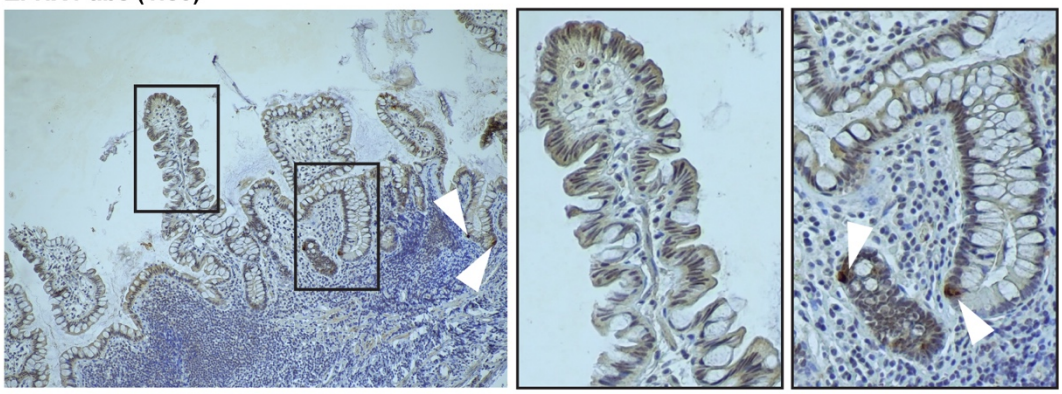


Antibody Optimisation with Small Intestine, Patient #11563

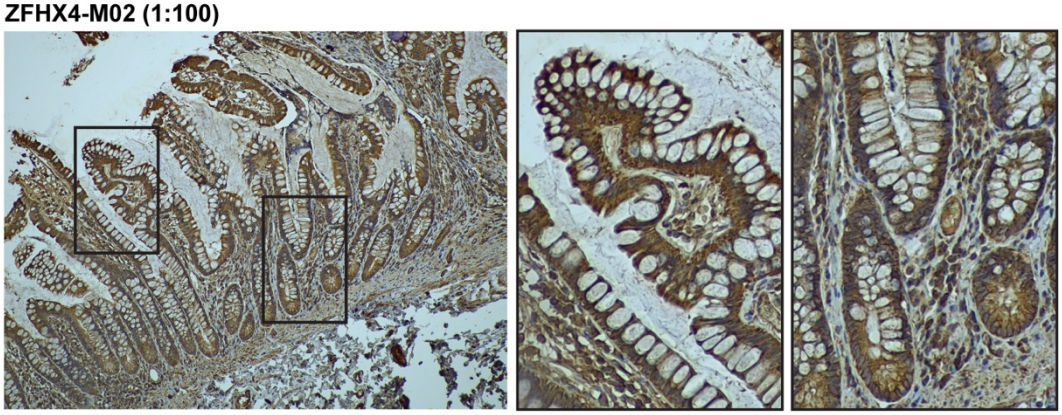
ZFHX4-abc (1:100)



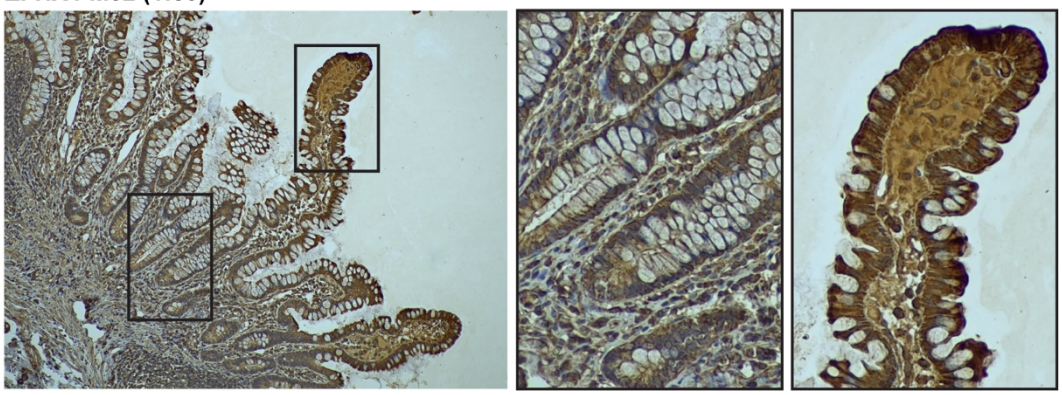
ZFHX4-abc (1:50)

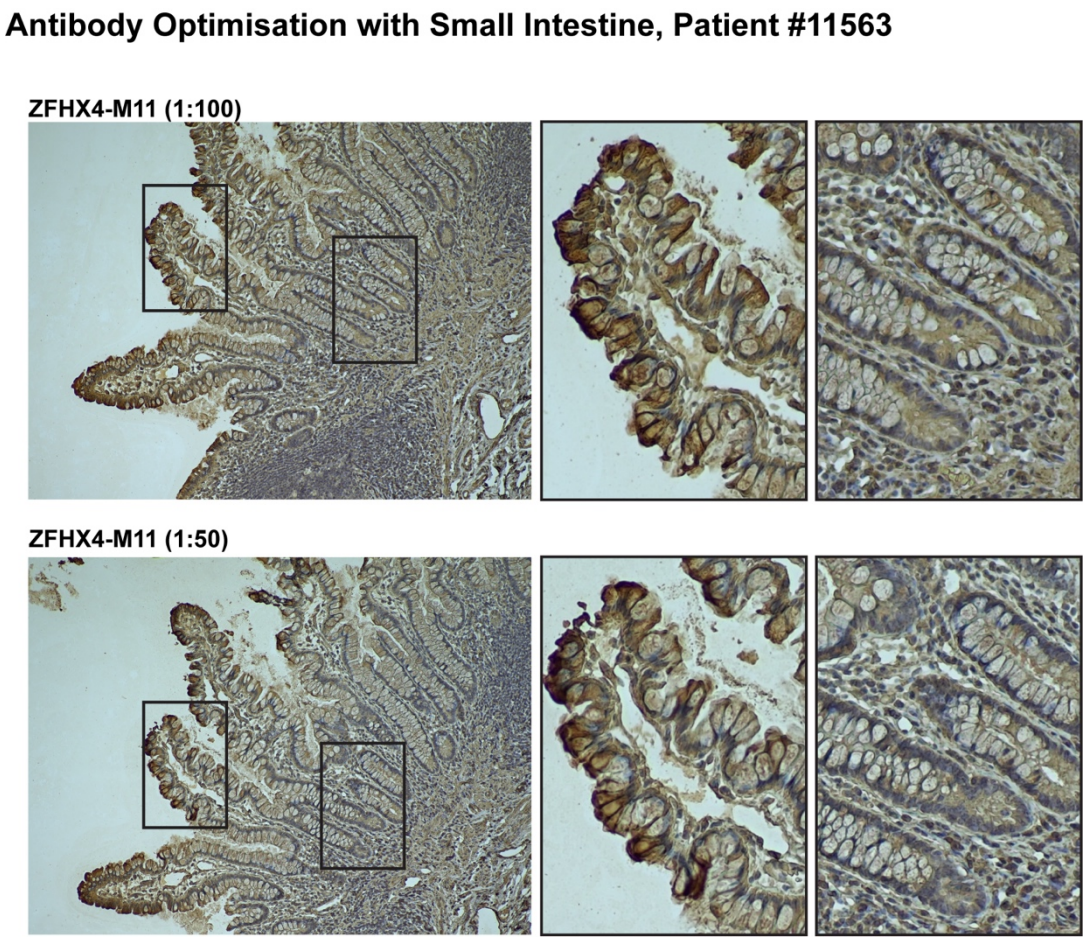


ZFHX4-M02 (1:100)



ZFHX4-M02 (1:50)





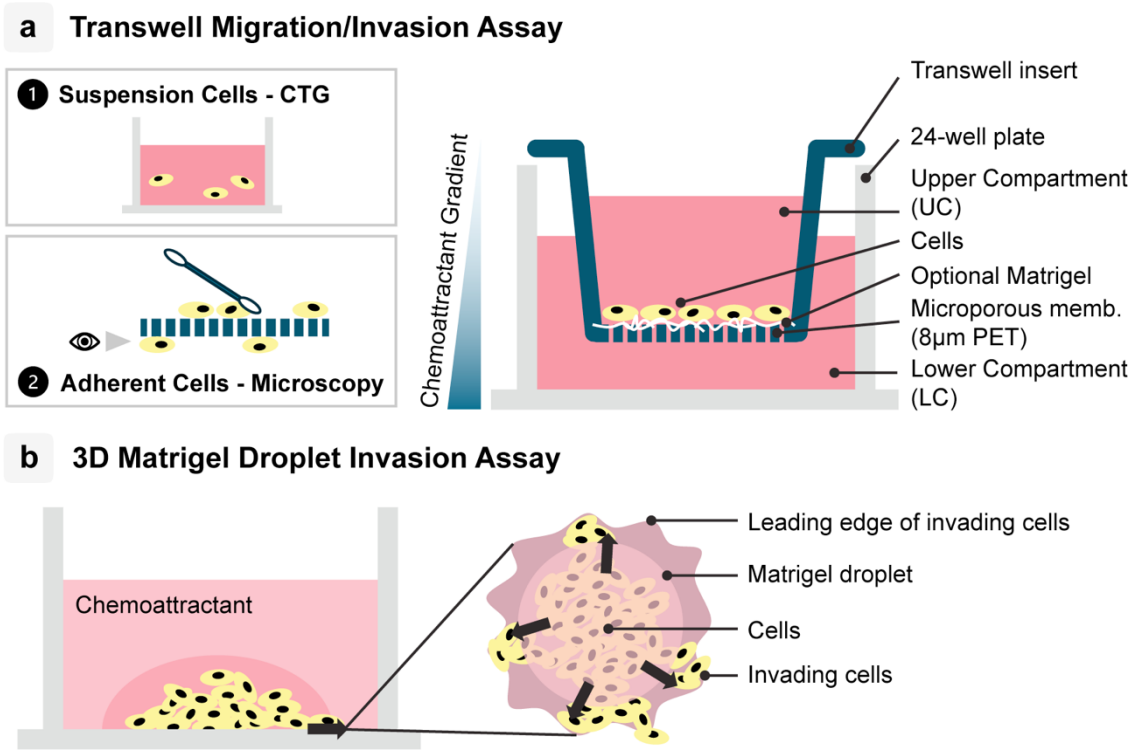
Supplementary Figure S4.1 | Optimisation of staining ZFHX4 protein expression by IHC using small intestine tissue (n=1). Slides were counterstained with Haematoxylin (blue signal) and DAB reagent produced brown signal for antibody binding. Background controls represent omission of primary antibody but use of secondary antibody. All images were taken at 10X and zoomed images highlighted with black boxes represent 30X. White triangles depict distinct dark brown/black staining.

Supplementary Note 4.2: A comment on available genome editing technologies

Traditional genome editing (CRISPR/Cas9) employs a Cas9 nuclease alongside single guide RNA (sgRNA) to induce double strand DNA breaks (DSBs) that are preferentially repaired by erroneous non-homologous end-joining (NHEJ), generating diverse indels and complete knockout (KO) and LoF at the gDNA and consequently mRNA and protein levels (Jinek et al., 2012, Cong et al., 2013). LoF phenotypes are subsequently investigated to extrapolate original gene function by means of reverse genetics. This approach was chosen to investigate ZFHX4-driven tumorigenicity in preference to RNA interference (small interfering or short hairpin RNA; si/shRNA), which typically yields transient, partial knockdown and hypomorphic phenotypes, whereas CRISPR/Cas9 enables consistent, robust and permanent knockout suitable for long-term studies (Boettcher and McManus, 2015). This approach has been further modified to model specific point mutations, for example through HDR-mediated point mutation knock-in using single- or double-stranded oligonucleotide donors and HDR-promoting conditions but it still relies on DSBs, suffers from low efficiency and risks unintended indels (Anzalone et al., 2020, Hess et al., 2017). Other technologies removing the need for template addition and limiting DNA damage during editing by employing nickase-based Cas9 include base editing. This introduces specific substitutions depending on Cas9 fusions with specific enzymes such as cytidine deaminases (to induce C>T or G>A substitutions; (Komor et al., 2016)) and adenosine deaminases (to induce A>G or T>C substitutions, (Gaudelli et al., 2017)). The most novel and precise method of specific genome editing, prime editing further expands editing scope beyond base-editing windows and transition types (Anzalone et al., 2019). Whilst base and prime editing offer greater precision than Cas9-mediated HDR and are efficient in cell line models, they are relatively recent technologies which are constantly evolving, and currently have shown limited success in primary colon organoid models with base editing efficiencies as low as 2% (Nicosia et al., 2025), with no prime editing published before and only a couple of studies recently performing base editing (Boretto et al., 2024, Geurts et al., 2023).

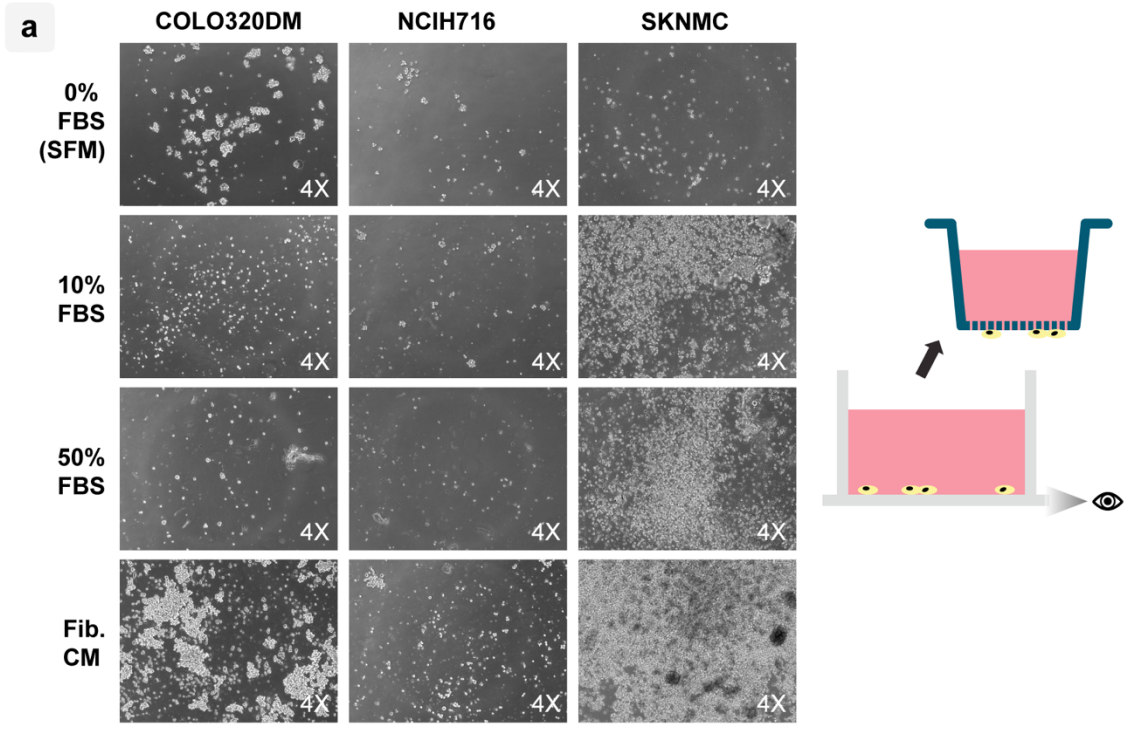
Supplementary Note 4.3: Migration and invasion assays

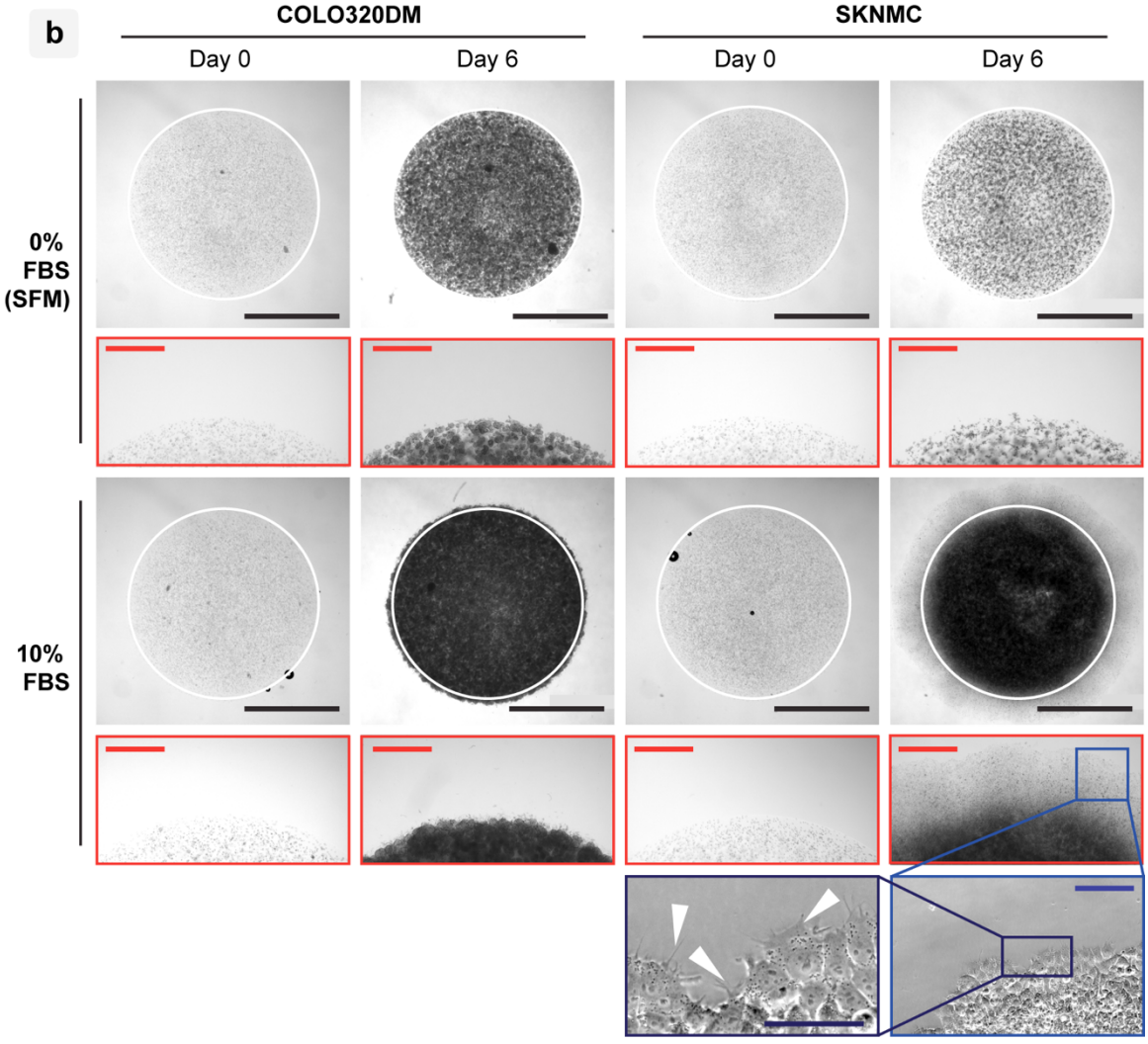
To investigate whether migration and invasion phenotypes were present in my models, Transwell assays measuring both migration and invasion (Supplementary Figure S4.2A) as well as 3D Matrigel droplet assays (Supplementary Figure S4.2B) were employed. Due to the nature of how the experimental colon lines grew (spherical-like COLO320DM and non-adherent NCIH716), scratch assays were not feasible as a complementary approach.



Supplementary Figure S4.2 | Schematics describing Transwell (A) and 3D Matrigel Droplet (B) assays to measure cell line migration and invasion. *[CTG=cell titre glo]*

Both methods exploit the principle of a chemoattractant gradient driving cell movement across a porous membrane to assess migration or through a Matrigel-coated barrier to additionally capture invasive potential. To my knowledge, migration and invasion haven't been previously characterised in COLO320DM and NCIH716, thus assays were firstly optimised in WT-nosgRNA lines at 5-6m passages (Supplementary Figure S4.3A/B).


a
COLO320DM
NCIH716
SKNMC
0% FBS (SFM)
10% FBS
50% FBS
Fib. CM
4X


b
COLO320DM
SKNMC
Day 0
Day 6
Day 0
Day 6
0% FBS (SFM)
10% FBS

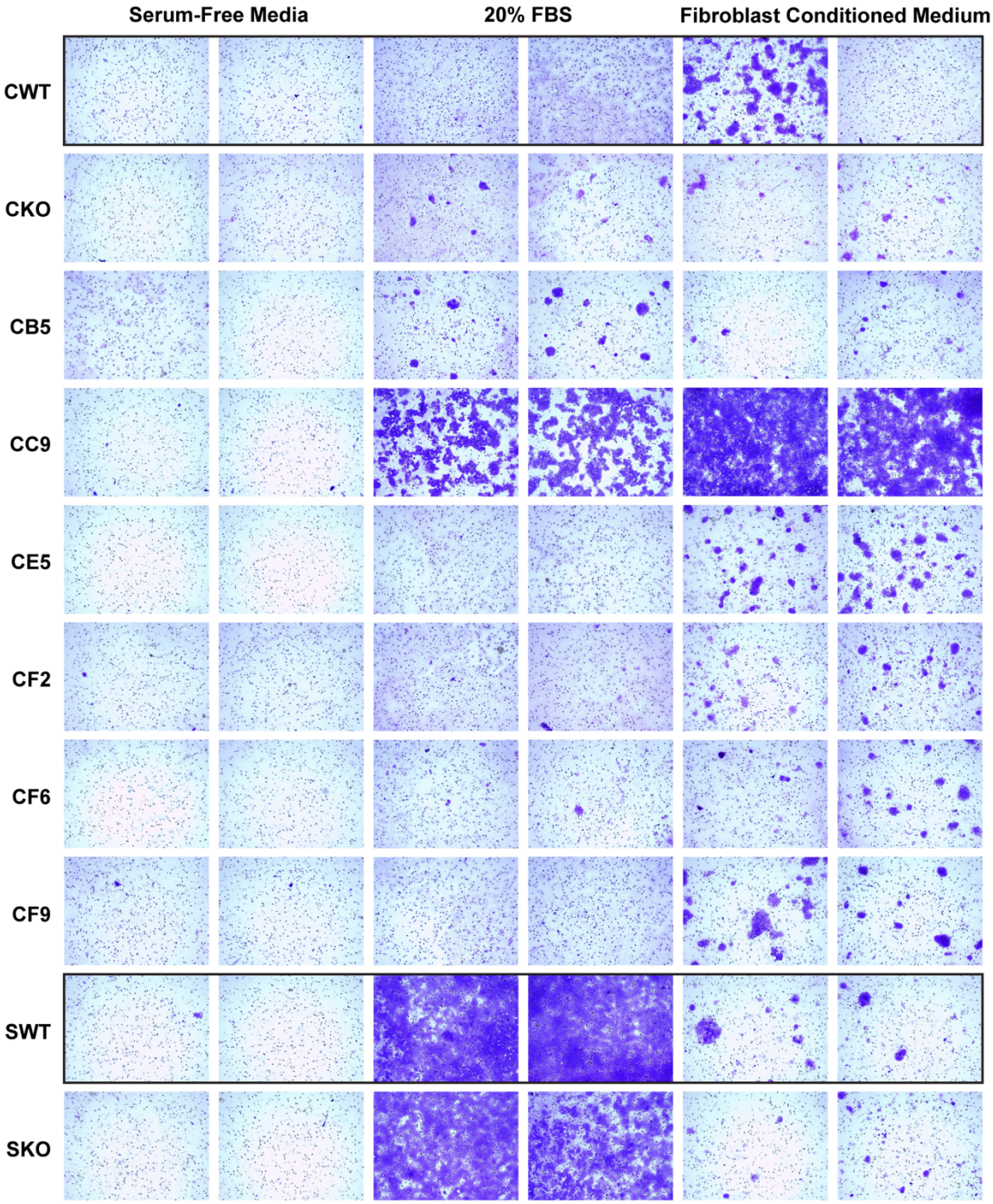
Supplementary Figure S4.3 | Initial optimisation of chemoattractants used for migration and invasion assays of COLO320DM, NCIH716 and SKNMC cells. (A) Microscope images following a 9-day Transwell migration assay. Images represent Transwell-removed wells of a 24-well plate following 9 day migration to different chemoattractant. Technical replicates were performed but only one representative image per condition is depicted. **(B)** Microscope images taken at 0 and 6 days following a 3D Matrigel droplet invasion assay. Black (2050μm), white (750μm), blue (150μm) and dark blue (75μm) scale bars represent scales respectively. White circles have been drawn to represent the Matrigel dome. White arrows indicate lamellipodia on invading cells with spreading morphology. *[FBS=foetal bovine serum, fib.CM = fibroblast conditioned medium]*

Initial optimisation included testing duration and choice of chemoattractant (FBS concentrations and fibroblast CM; FCM). To save work in staining membranes of potentially non-migrated conditions and there being difficulties in visualising cells migrated to the membrane underside by live microscopy, cells that instead detached and settled at the bottom of the 24-well plate were considered a surrogate readout of migration. Surprisingly, at 24 hours no migrated cells were observed in either COLO320DM and NCIH716 lines (data not shown), so the assay was extended to a total of 9 days with daily inspections to assess the effect of time. Images taken following Transwell-insert removal at 9 days were used as an approximate measure of chemoattractant efficiency (Supplementary Figure S4.3A).

The SKNMC line was highly migratory, with robust migration evident towards 10% FBS, 50% FBS and maximally towards FCM. Minimal migration in serum-free controls support the success of this assay and validity of chemoattractant gradient permitting migration. Unexpectedly, both COLO320DM and NCIH716 exhibited minimal migration towards 10% and even 50% FBS compared to serum-free controls, despite an uncommonly extended period of 9 days. By comparison, FCM did promote some migration in these lines and was therefore selected as the chemoattractant for subsequent experiments to enable detection of any differences between ZFHX4-WT and KO cells. Nevertheless, the extended assay duration particular to 50% FBS led to signs of cellular stress apparent on microscope images, suggesting that a shorter protocol with FCM and lower FBS concentrations would be preferable.

Similar findings were reflected in a 3D Matrigel droplet invasion assay. After 6 days surrounded by 10% FBS, no COLO320DM cells invaded the Matrigel which remained intact, despite proliferation within the droplet compared to the serum-free control (Supplementary Figure S4.3A). However, SKNMCs displayed robust invasive behaviour, with a leading edge of invading cells spreading beyond the Matrigel dome that adopted elongated morphology and lamellipodia. Taken together, these assays can reliably test migration and invasion but reveal strong cell line-specific differences. Both experimental colon cell lines were only weakly migratory comparative to SKNMCs which were highly migratory and invasive, consistent with their mesenchymal and spreading phenotype in culture.

Following optimisation, the Transwell migration assay was subsequently performed on all *ZFHX4*-WT and KO cell lines at 9m age, employing serum-free, 20% FBS and FCM as chemoattractants and a 4-day duration. Each condition comprised two technical insert-replicates, and additional technical replicates were obtained from multiple measurements within each insert depending on downstream technique. For suspension NCIH716 cells, CellTitreGlo was performed on cells traversed into the bottom compartment to estimate viable cell number as a metric for migration. For COLO320DM and SKNMC cells, insert membranes were extracted after removal of non-migrated cells from the upper surface with cotton buds, then crystal violet-stained and imaged (Supplementary Figure S4.4). Percentage area of cell coverage was quantified by Fiji as the metric for migration. As observed previously, FBS albeit at a slightly different concentration failed to induce COLO320DM WT (CWT) migration whereas SKNMC cells exhibited robust migration with nearly 100% area coverage visible by eye. CWT cells displayed some migration towards FCM, although this was only evident in one replicate. For accurate interpretation, all data were quantified and statistical testing was performed. Background migration in serum-free media was subtracted for each condition and data were normalised to the respective WT control where positive migration was observed to ensure comparability across samples.

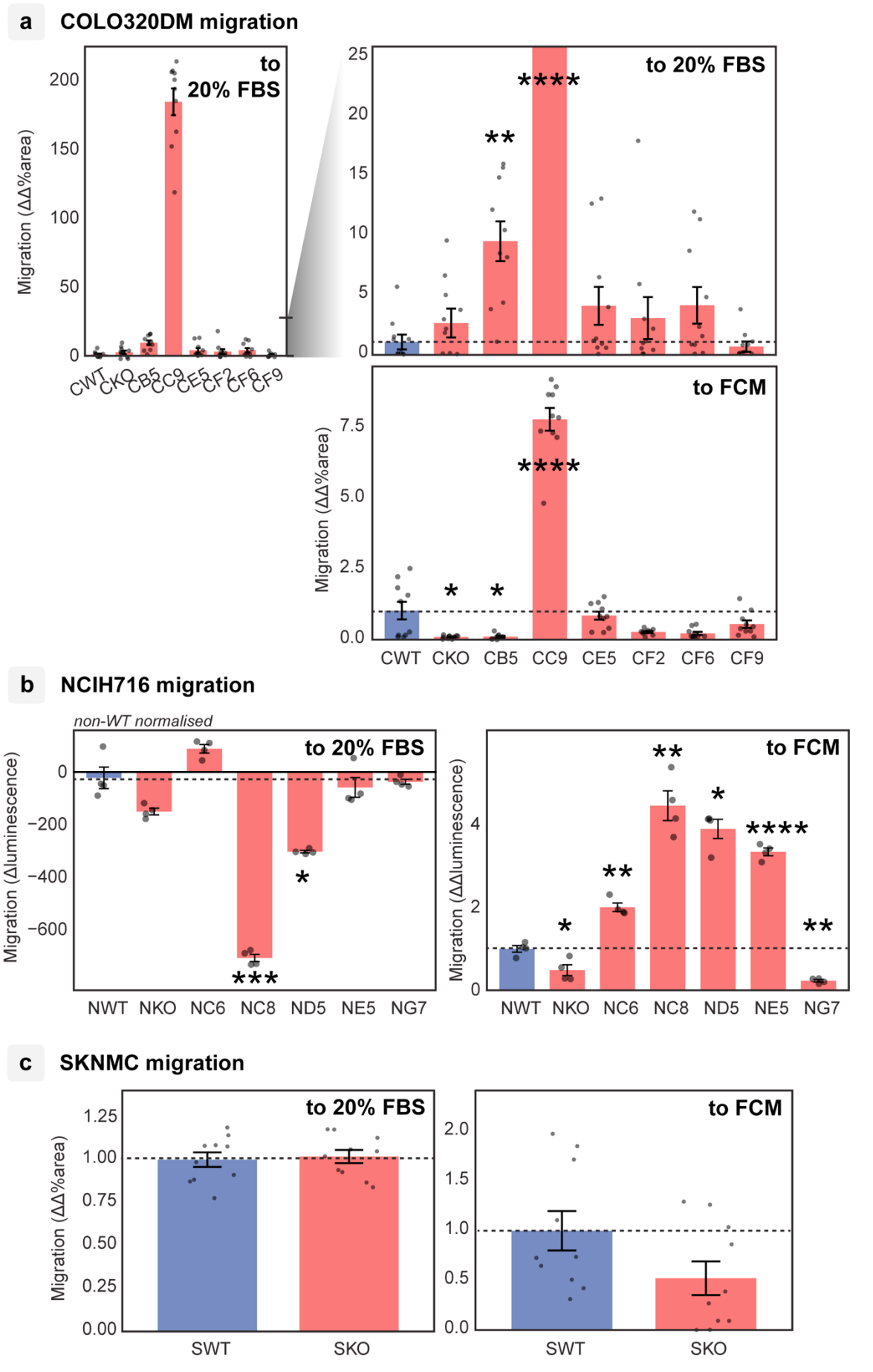


Supplementary Figure S4.4 | Microscope images of crystal-violet stained Transwell-insert membranes following a Transwell migration assay. Wild-types are highlighted with black boxes. All images are 10X. Side by side images represent one representative field of view (n=5) from two insert-technical replicates (final n=10).

In COLO320DM, CC9 demonstrated significant profound migration compared with CWT, showing a mean 184-fold increase towards 20% FBS (adj.p=9.41×10^{-8}) and 7.77-fold

increase towards FCM (adj.p=0.0000758) (Supplementary Figure S4.5A). Other ZFHX4 monoKOs also trended towards enhanced migration to FBS, with CB5 demonstrating a significant 9.46-fold increase relative to CWT (adj.p=0.00187). Conversely, other than CC9, migration towards FCM was generally reduced in *ZFHX4* KOs compared with CWT, with CKO and CB5 showing significant mean decreases of 15.3-fold (adj.p.=0.0121) and 13.7-fold (adj.p=0.0121) respectively and modest 1.22-5.18-fold mean reductions in other monoKOs.

In NCIH716, background-subtracted migration towards 20% FBS was minimal or negative across most samples, implying migration to serum-free media often exceeded that to 20% FBS (Supplementary Figure S4.5B). NWT showed no appreciable migration above serum-free controls (-20.5 luminescence) preventing meaningful normalisation to NWT as division by a negative value would distort both effect size and direction. Thus, only background-subtracted migration was reported for FBS. Several *ZFHX4* monoKOs migrated significantly less than to serum-free medium, including NC8 (−707 luminescence, adj.p=0.000952) and ND5 (−302 luminescence, adj.p=0.0178). Only NC6 (90.25 luminescence) demonstrated modest positive migration though this wasn't significant. One possible explanation for these unexpected results is that 20% FBS may have increased cell stickiness and adhesion to the underside of the membrane upon migration, causing migrated cells to remain attached and not analysed by CellTitreGlo. Supporting this, migration trends appeared consistent between FBS and FCM (for example, NC8>ND5>NE5>NG7) suggesting comparable intrinsic migratory capacity across conditions. In contrast, FCM elicited robust responses, with most *ZFHX4* monoKOs migrating significantly more than NWT. For example, NC8 exhibited a 4.46-fold mean increase (adj.p=0.00259). However, NKO and NG7 displayed reduced migration, with significant mean 2.06-fold (adj.p=0.0243) and 4.35-fold (adj.p=0.00205) decreases respectively. Collectively these findings suggest that NCIH716 cells don't reliably migrate towards FBS (at least as detected by this assay), in contrast to FCM which served as a more effective chemoattractant.

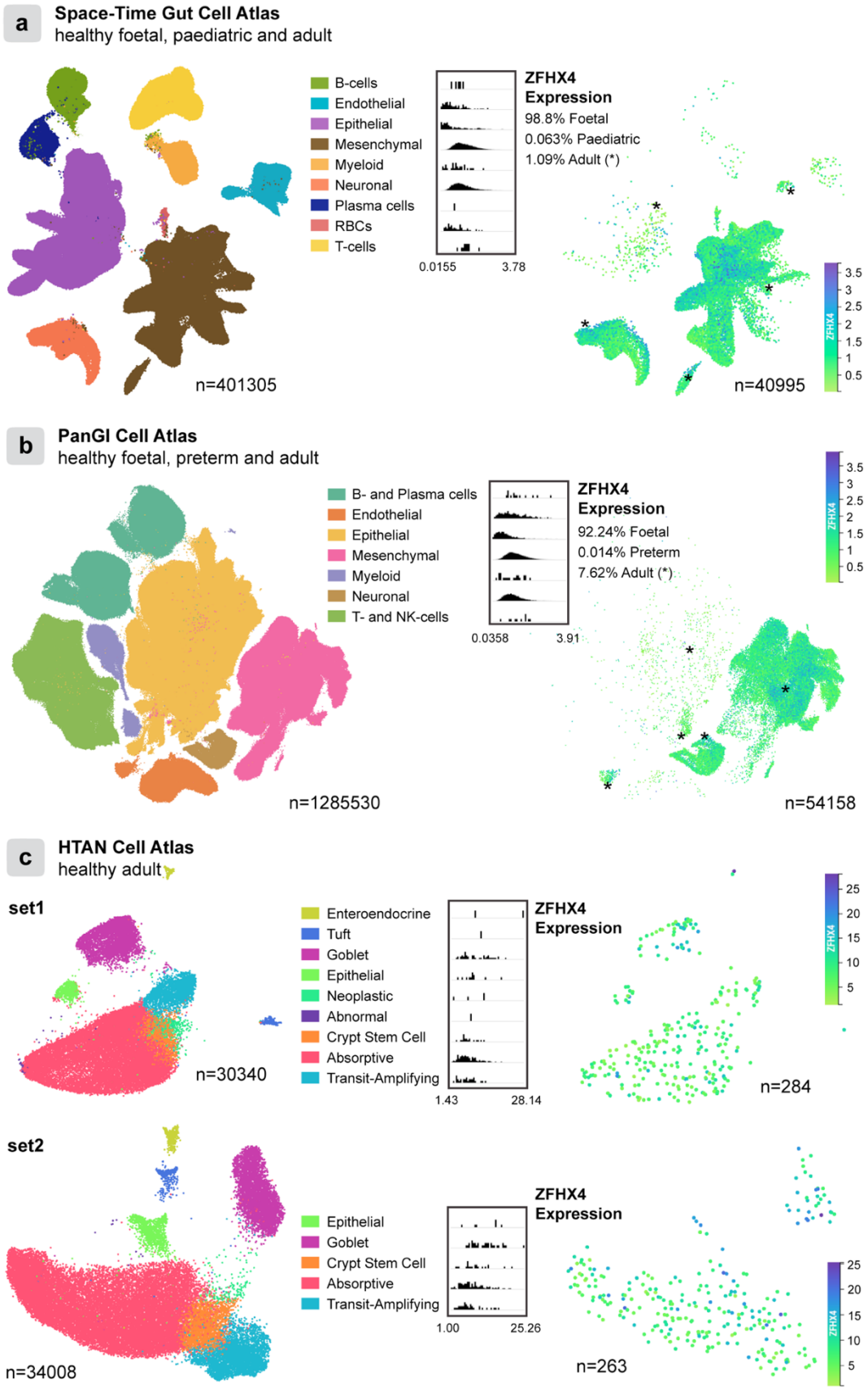

a
COLO320DM migration
to 20% FBS
Migration (ΔΔ%area)
200
150
100
50
0
CWT CKO CB5 CC9 CE5 CF2 CF6 CF9
25
20
15
10
5
0
**
****
to FCM
7.5
5.0
2.5
0.0
*
*
****
CWT CKO CB5 CC9 CE5 CF2 CF6 CF9
b
NCIH716 migration
non-WT normalised
Migration (ΔluminescenceΔ)
0
−200
−400
−600
***
*
NWT NKO NC6 NC8 ND5 NE5 NG7
Migration (ΔΔluminescence)
4
2
0
*
**
**
*
****
**
NWT NKO NC6 NC8 ND5 NE5 NG7
c
SKNMC migration
Migration (ΔΔ%area)
1.25
1.00
0.75
0.50
0.25
0.00
SWT SKO
2.0
1.5
1.0
0.5
0.0
SWT SKO

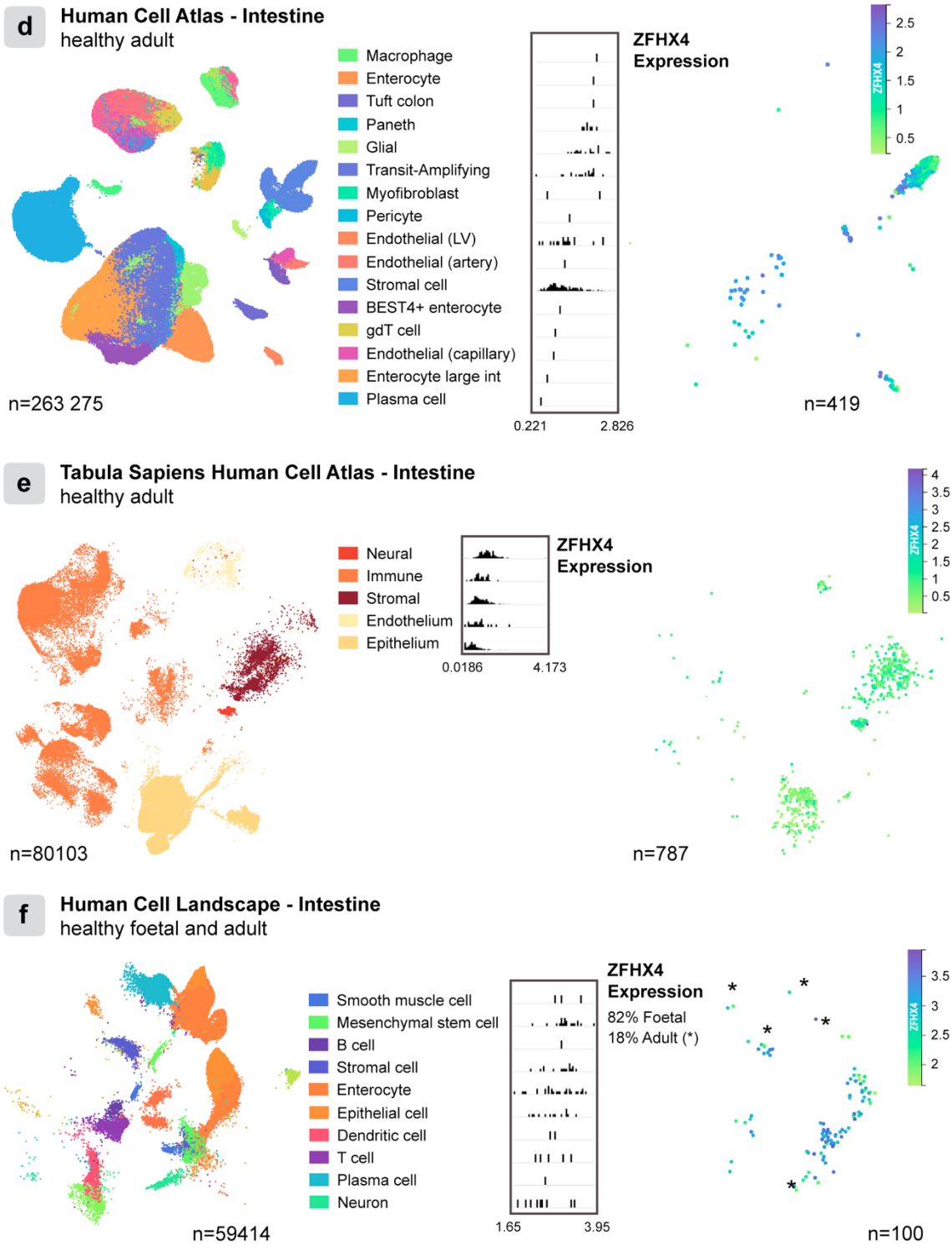
Supplementary Figure S4.5 | Quantification of the Transwell migration assay. All data pertains to one experiment conducted for 4 days on 9m cells. Plots are separated by cell line; COLO320DM **(A)**, NCIH716 **(B)** and SKNMC **(C)** as well as FBS and FCM chemoattractants. Plots in A and C refer to migration calculated from subtracting background migration in serum-free controls and dividing by respective WT. Each condition represents two technical insert-replicates, each with five fields of view quantification (total n=10). Plots in B in 20% FBS refer to migration calculated from subtracting background migration but no division of WT for normalisation due to negative migration. Each condition represents two technical insert-replicates, each with two cell titre glo replicates (total n=4). Statistical testing comparing conditions to wild-type controls employed t-testing or wilcoxon testing based on Shapiro-Wilkes normality. Asterisks denote the BH-corrected p-value. *[*=p<0.05, **=p<0.01, ***=p<0.001, ****=p<0.0001, FCM=fibroblast conditioned medium]*

In the SKNMC line, no significant differences in migration were observed between SKO and SWT towards either 20% FBS or FCM. Overall, migration patterns across all lines appeared more specific to individual KO samples rather than representing a consistent *ZFHX4*$^{-/-}$ phenotype, likely reflecting underlying genomic and transcriptomic heterogeneity between clones. Interpretation here should be approached with caution, for example CC9 and NG7 (clones with the lowest *ZFHX4* expression and considered canonical knockouts) exhibited completely opposite migration phenotypes, each with large and significant effect sizes.

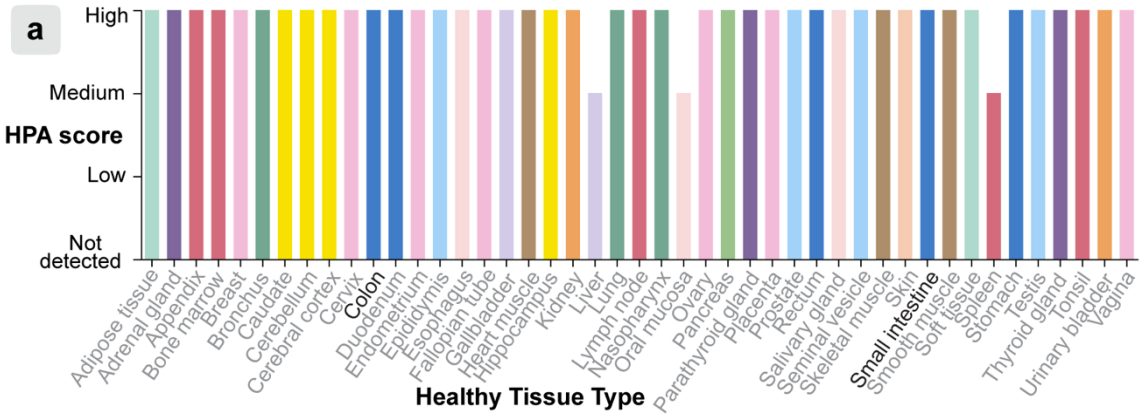
More experimental replicates are required to account for FCM batch differences, passage-related differences and spurious Transwell-related differences as well as further investigating invasion with Matrigel used in the inserts. For example, large differences between technical insert-replicates (n=2) in CWT were apparent which likely led to reduced power in statistical testing with *ZFHX4* KOs. Moreover, FCM itself is a complex stimulus containing multiple chemoattractants, complicating interpretation. While it was necessary to use because COLO320DM and NCIH716 failed to migrate well towards FBS, its heterogeneity makes it difficult to attribute observed effects to specific biology. This complexity is illustrated by CB5, which migrated significantly more than CWT towards FBS but significantly less towards FCM, highlighting context-dependent rather than inherent migratory capacity. Combining both FBS and FCM data may be a better metric for inherent migratory capacity independent of context but considering the variability within this experiment, more experimental runs would provide further insight. Taken together, these findings indicate robust conclusions about migration and invasion in COLO320DM and

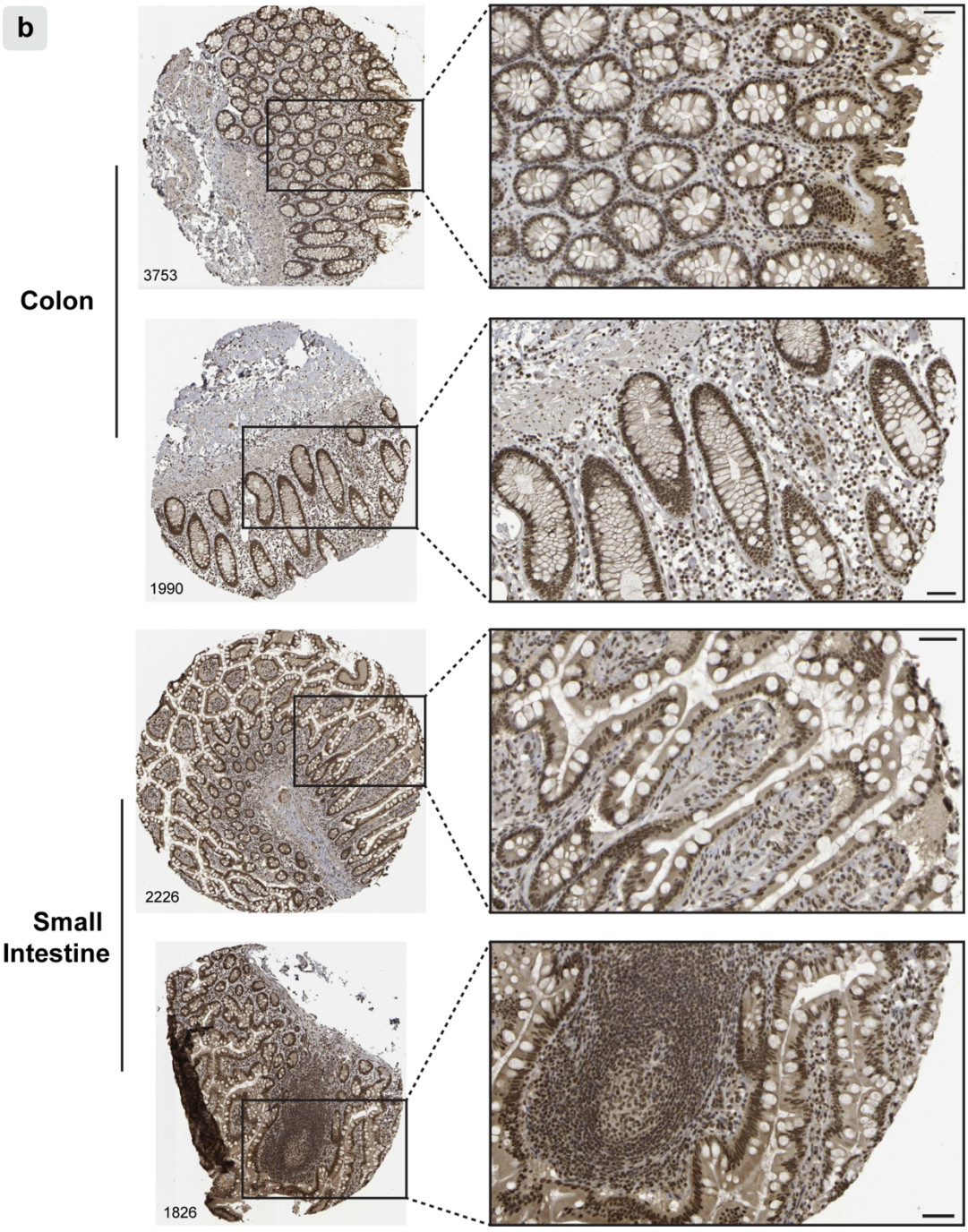
NCIH716 are difficult to draw. Their inherently poor responsiveness to FBS, combined with the confounding nature of FCM as a chemoattractant, limits the ability to define a clear biological phenotype. Additional experimental replicates including invasion assays using Matrigel-coated inserts will be required to resolve these uncertainties.


a
Space-Time Gut Cell Atlas
healthy foetal, paediatric and adult
B-cells
Endothelial
Epithelial
Mesenchymal
Myeloid
Neuronal
Plasma cells
RBCs
T-cells
0.0155
3.78
ZFHX4 Expression
98.8% Foetal
0.063% Paediatric
1.09% Adult (*)
n=401305
n=40995
ZFHX4
b
PanGI Cell Atlas
healthy foetal, preterm and adult
B- and Plasma cells
Endothelial
Epithelial
Mesenchymal
Myeloid
Neuronal
T- and NK-cells
0.0358
3.91
ZFHX4 Expression
92.24% Foetal
0.014% Preterm
7.62% Adult (*)
n=1285530
n=54158
c
HTAN Cell Atlas
healthy adult
set1
Enteroendocrine
Tuft
Goblet
Epithelial
Neoplastic
Abnormal
Crypt Stem Cell
Absorptive
Transit-Amplifying
1.43
28.14
ZFHX4 Expression
n=30340
n=284
set2
Epithelial
Goblet
Crypt Stem Cell
Absorptive
Transit-Amplifying
1.00
25.26
ZFHX4 Expression
n=34008
n=263

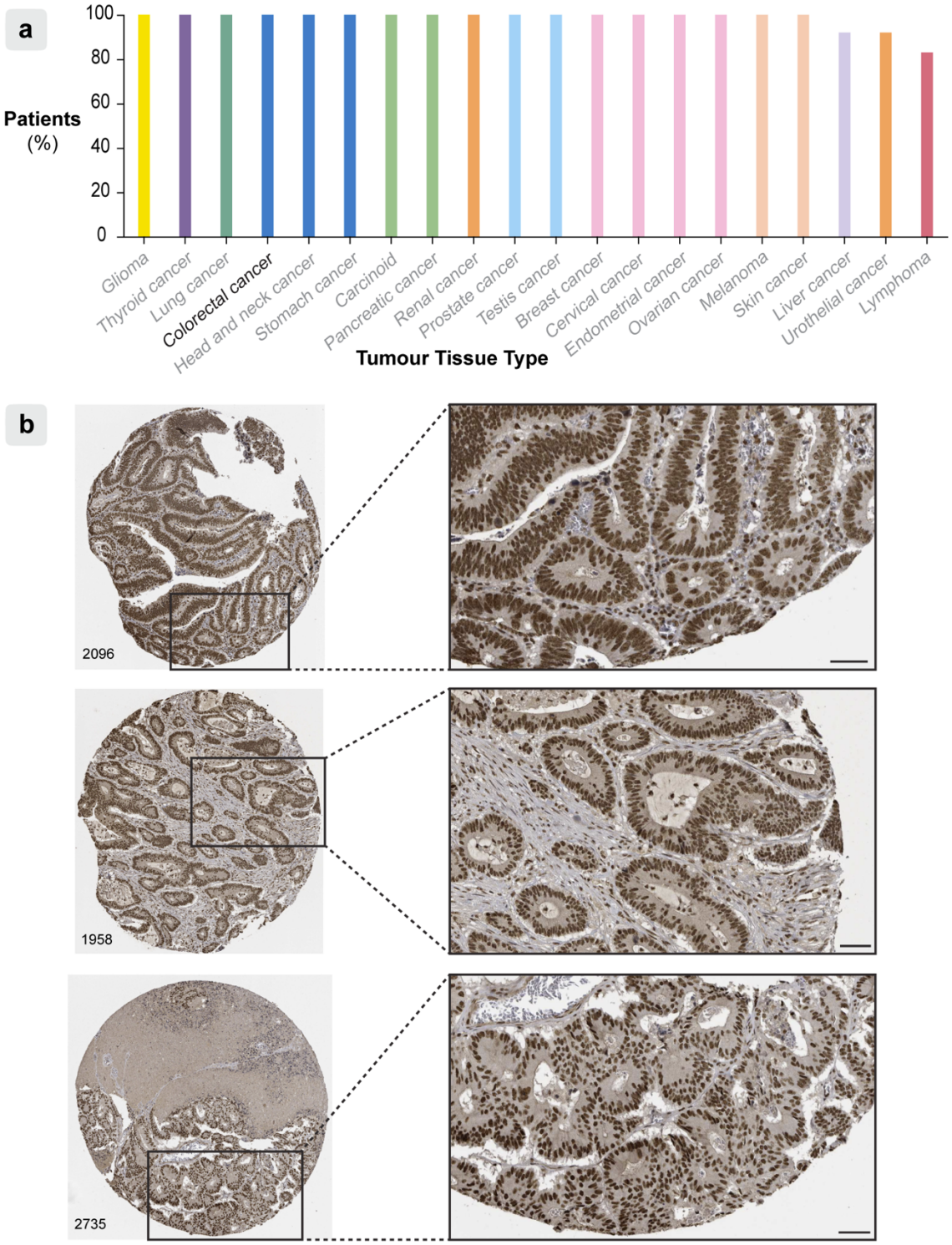


Supplementary Figure S4.6 | ZFHX4 expression in healthy intestine scRNAseq datasets. Each letter identifier represents an individual scRNAseq dataset. N numbers represent the total number of cells following filtering for intestine-specific origin. tSNE plots on the left are annotated by the lowest level of cell type annotation by the respective publication. Matching colour keys also depict mini density plots of ZFHX4 expression (x axis also denoted with expression as log-transformed normalised counts per 10k reads (ln(CP10K+1) except for the HTAN dataset that appeared to be normalised differently). tSNE plots on the right depict cells that express ZFHX4 (non-zero filtering) to degrees respective to the viridis colour scale on each plot. N numbers represent the number of ZFHX4-expressing cells. For datasets including developmental stages more than adult only, asterisks represent general clusters for approximately denote ZFHX4-expressing cells of adult origin.

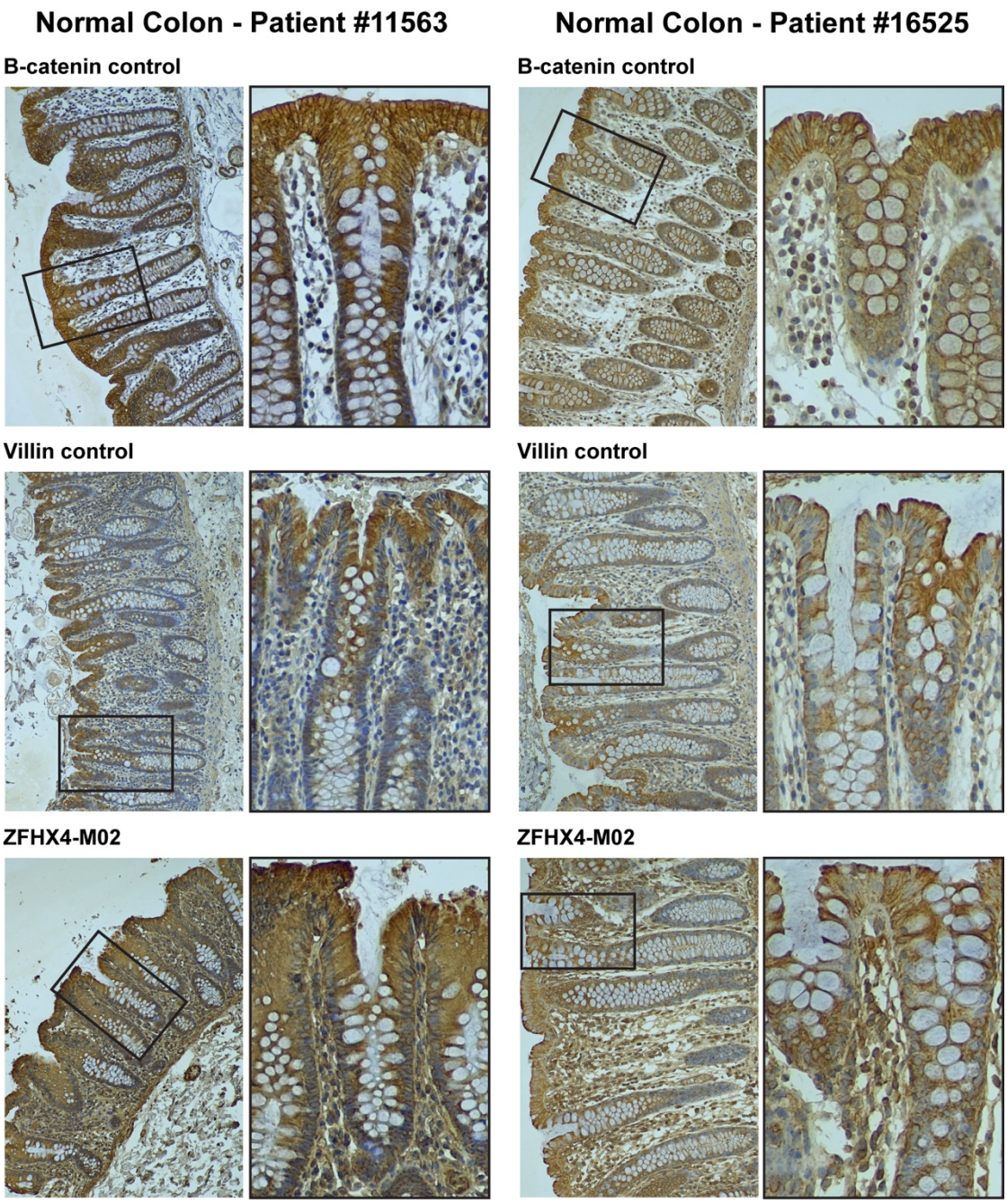

a
High
Medium
HPA score
Low
Not detected
Adipose tissue
Adrenal gland
Appendix
Bone marrow
Breast
Bronchus
Caudate
Cerebellum
Cerebral cortex
Cervix
Colon
Duodenum
Endometrium
Epididymis
Esophagus
Fallopian tube
Gallbladder
Heart muscle
Hippocampus
Kidney
Liver
Lung
Lymph node
Nasopharynx
Oral mucosa
Ovary
Pancreas
Parathyroid gland
Placenta
Prostate
Rectum
Salivary gland
Seminal vesicle
Skeletal muscle
Skin
Small intestine
Smooth muscle
Soft tissue
Spleen
Stomach
Testis
Thyroid gland
Tonsil
Urinary bladder
Vagina
Healthy Tissue Type

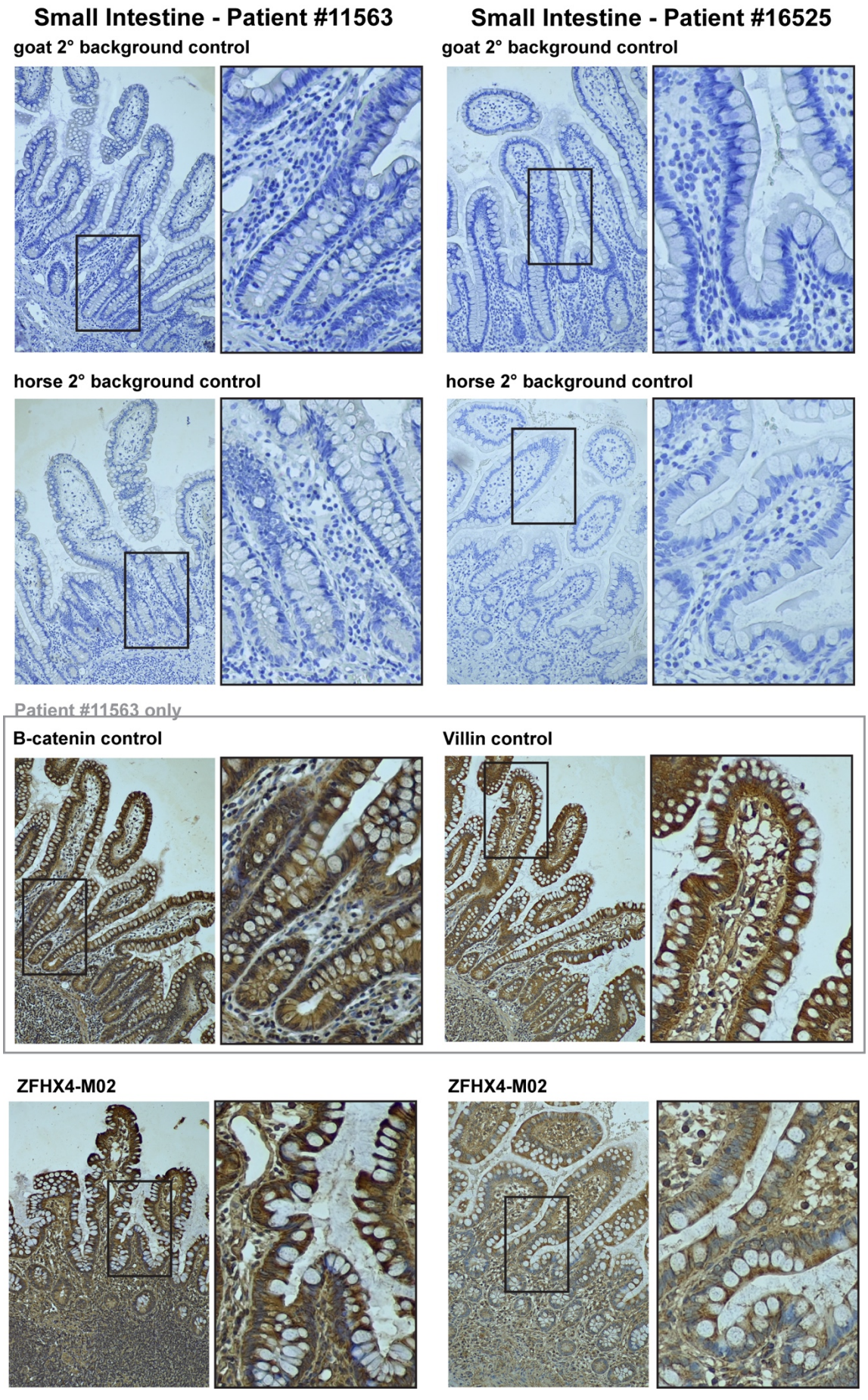

b
Colon
3753
1990
Small Intestine
2226
1826

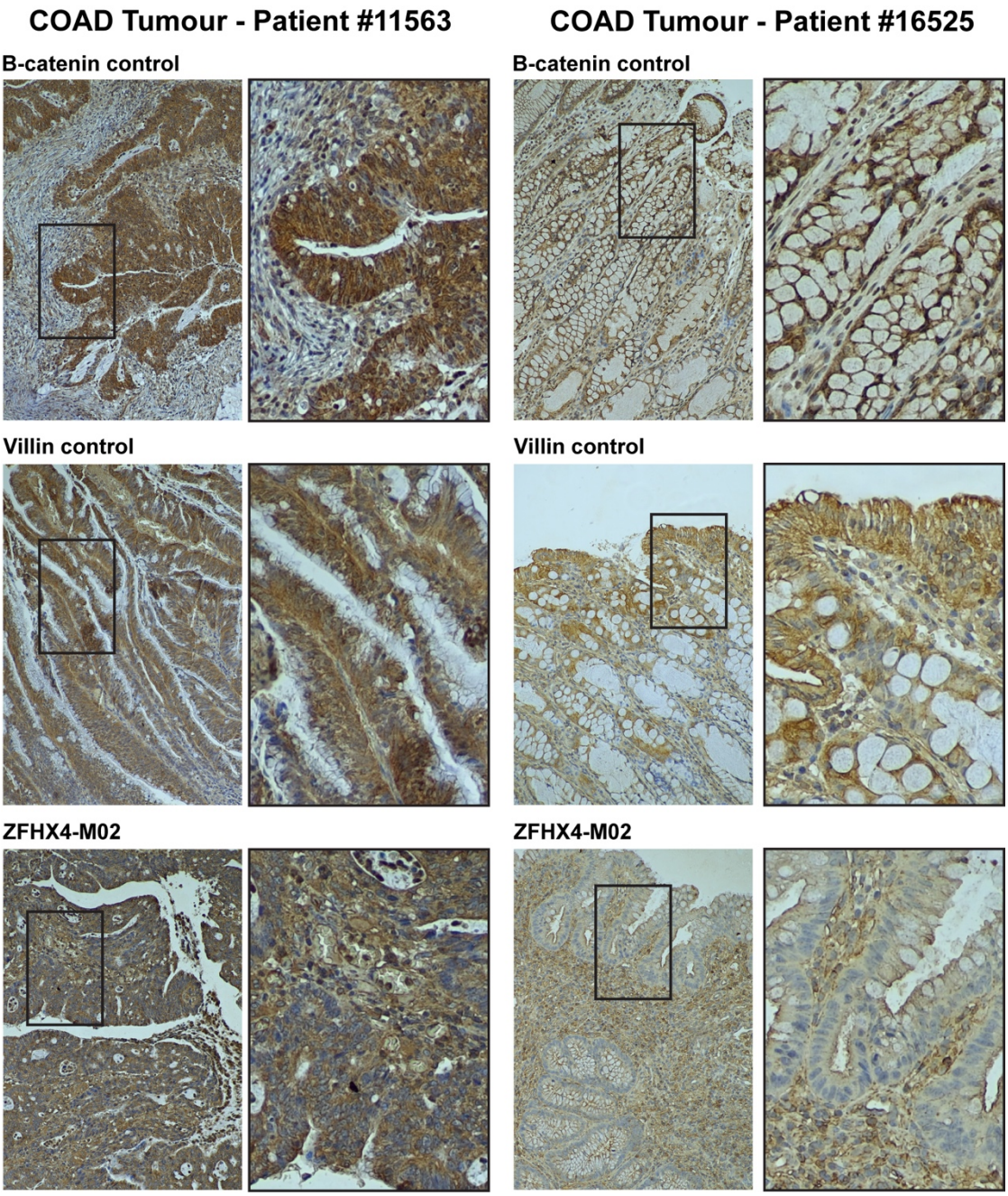
Supplementary Figure S4.7 | ZFHX4 protein expression across healthy adult human tissues, with a specific focus on the colon and small intestine. (A) Human Protein Atlas scores of ZFHX4 protein expression across multiple healthy human tissue types. Human Protein Atlas protein staining score is based on IHC annotation by a specialist and verification by a second specialist. Bar colour coding is based upon similar tissue groups. **(B)** Images of ZFHX4-stained IHC of two colon and small intestine samples. Patient IDs are written on each image. Scale bars on zoomed portions represent 50μm. Adapted from Human Protein Atlas (Uhlen et al., 2015) that uses the Sigma HPA023837 antibody for ZFHX4 staining (1:30 dilution, HIER ph6 antigen retrieval). *[HPA = human protein atlas, IHC = immunohistochemistry]*



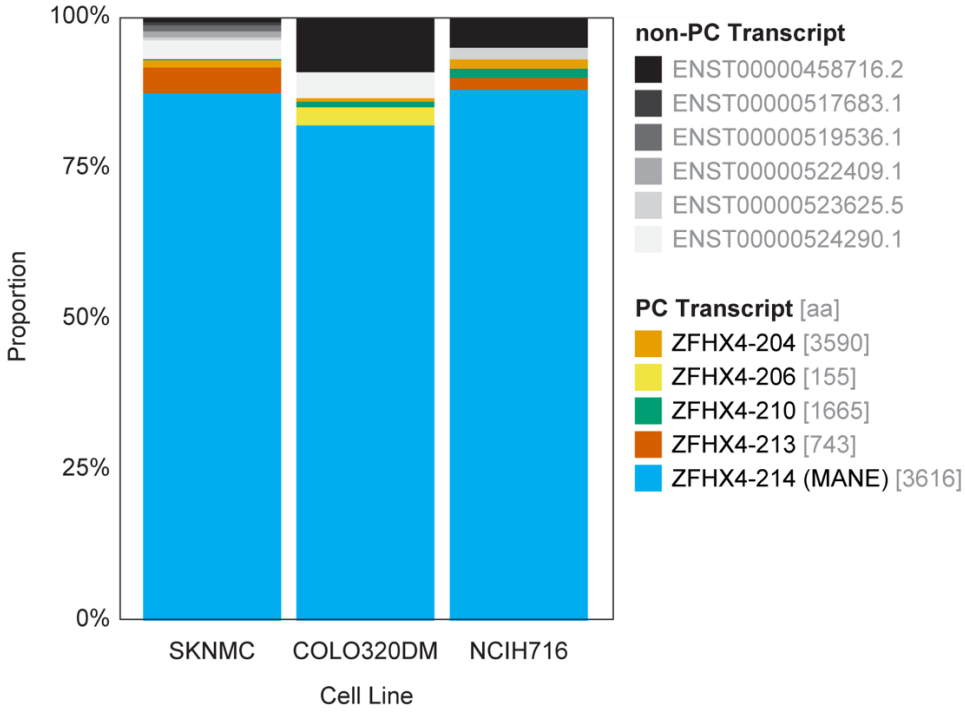
Supplementary Figure S4.8 | ZFHX4 protein expression across multiple cancer tissues, with a specific focus on COAD. (A) Percentage of patients with high and medium ZFHX4 protein expression across multiple tumour tissues. HPA protein staining score is based on IHC annotation by a specialist and verification by a second specialist. Bar colour coding is based upon tissue of origin. **(B)** Images of ZFHX4-stained IHC of three COAD samples. Patient IDs are written on each image. Scale bars on zoomed portions represent 50μm. Adapted from Human Protein Atlas (Uhlen et al., 2015) that uses the Sigma HPA023837 antibody for ZFHX4 staining (1:30 dilution, HIER ph6 antigen retrieval). *[HPA = human protein atlas, IHC = immunohistochemistry]*


Normal Colon - Patient #11563
Normal Colon - Patient #16525
B-catenin control
B-catenin control
Villin control
Villin control
ZFHX4-M02
ZFHX4-M02

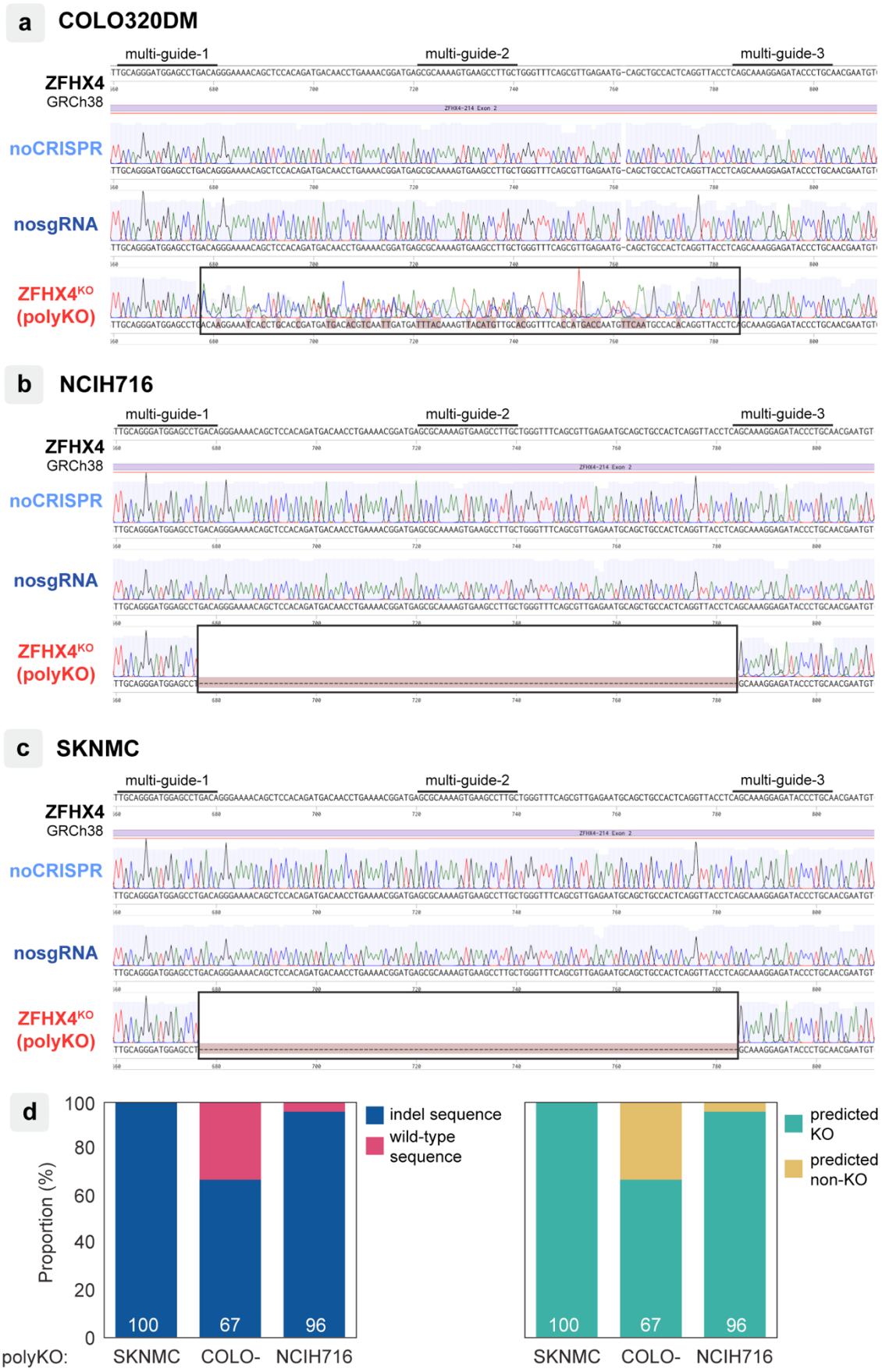

Small Intestine - Patient #11563
goat 2° background control
Small Intestine - Patient #16525
goat 2° background control
horse 2° background control
horse 2° background control
Patient #11563 only
B-catenin control
Villin control
ZFHX4-M02
ZFHX4-M02



Supplementary Figure S4.9 | IHC controls across matched normal colon, small intestine and COAD tissue from two patients. Slides were counterstained with Haematoxylin (blue signal) and DAB reagent produced brown signal for antibody binding. All images were taken at 10X and zoomed images highlighted with black boxes and annotated with respective number represent 30X.

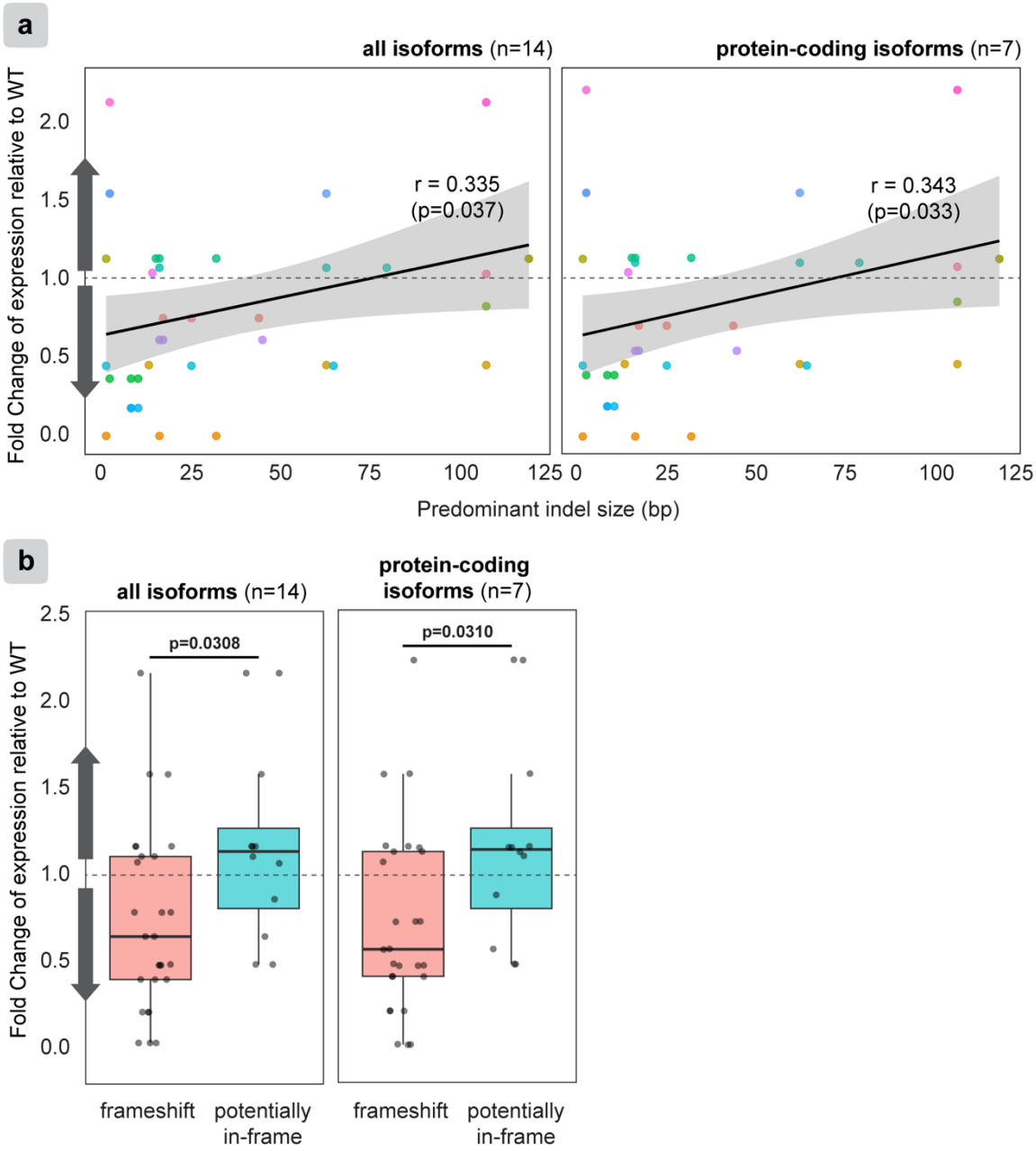


Supplementary Figure S4.10 | Proportions of protein-coding (PC) and non-protein-coding (non-PC) ZFHX4 mRNA transcripts detected in the chosen cell lines for ZFHX4 study. Data was obtained from DepMap database and transcripts are with reference to the ENSEMBL database.

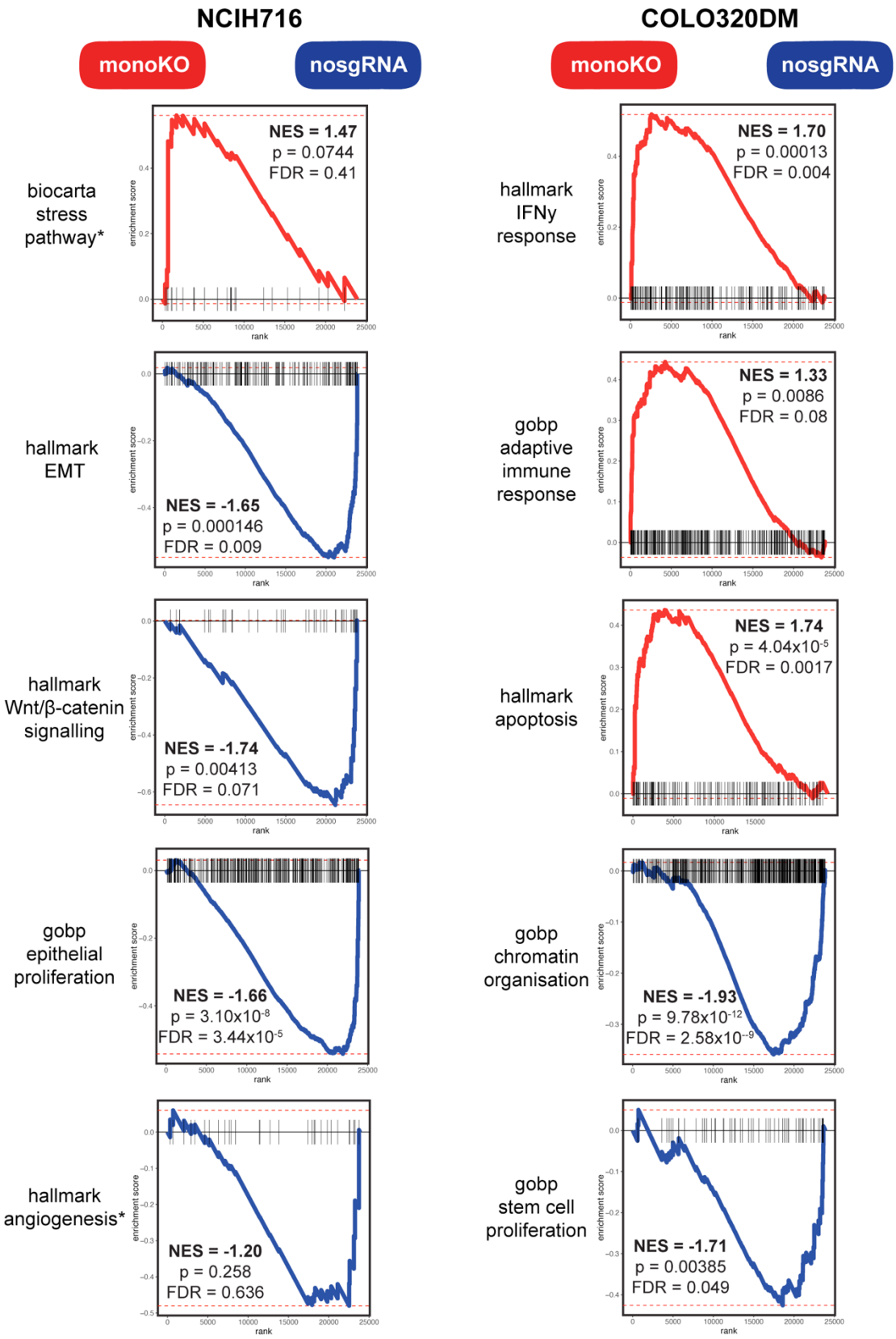


Supplementary Figure S4.11 | Genomic validation of polyclonal ZFHX4^{KO} cell lines by sanger sequencing. (A-C) Trace files of each cell line's sanger sequencing results aligned to the ZFHX4

Exon 2 region on Benchling software. **(D)** Synthego-ICE results of proportions of indel vs wild-type estimations as well as KO and non-KO predictions for polyKO samples. Values overlayed in white represent % of indel or KO according to plot.



Supplementary Figure S4.12 | The correlation of ZFHX4 expression at the transcript-level with indel size and indel type following CRISPR-mediated gDNA disruption. (A) Each point represents an RNAseq replicate. Indel sizes were the top 3 for each replicate identified in sanger sequencing during a similar time to RNAseq. For example, one replicate can have up to 3 points and these are coloured by replicate. Fold change was calculated by doing KO/WT average TPM summed of all isoforms or protein-coding isoforms only thus fold changes > 1 represent replicates where expression was greater than WT. Both polyKOs and monoKOs are represented here. Correlation coefficient r values and p-values were produced following Pearson testing. **(B)** Box plots of the same information but by indel type, designated frameshift if not a multiple of 3 and potentially in-frame if a multiple of 3. Normality testing led to parametric unpaired t-testing, indicating p-values on the plots.



Supplementary Figure S4.13 | Gene set enrichment analysis (GSEA) results of interesting pathways in ZFHX4 monoKO cell lines and respective controls. All genes sets were obtained from MSigdbr. Pathways with * highlight no significance at p-value or FDR level. *[NES = normalised enrichment score, FDR = false discovery rate]*

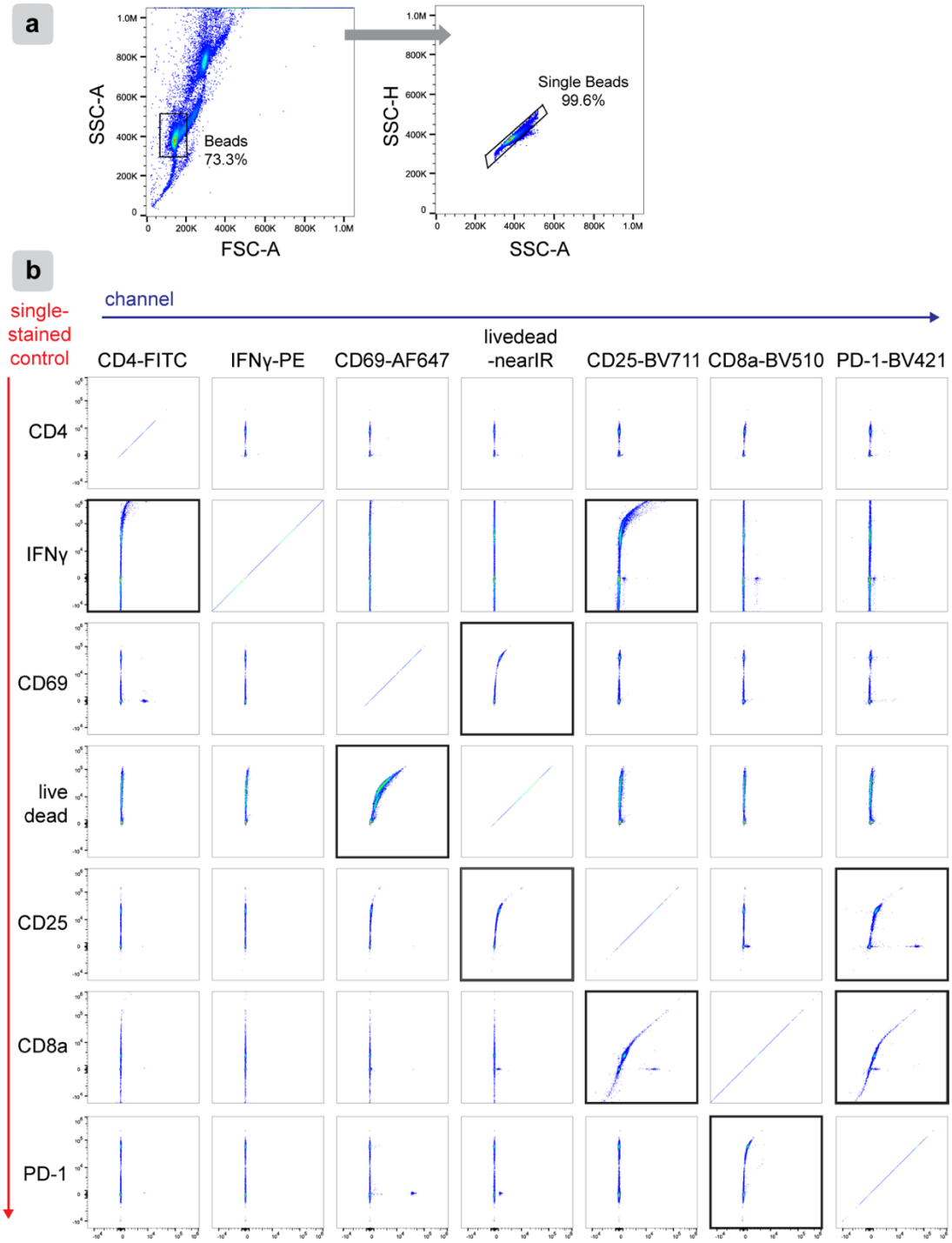
Supplementary Table 4.1 | Top 30 upregulated and downregulated genes in the metastatic organoid nosgRNA vs noCRISPR comparison, with respect to nosgRNA and ordered by t-statistic. MIR, CTD-, RP-, LINC-, AC-, NA_mapped genes were removed.

Upregulated				Downregulated			
Gene	**t-stat.**	**log_2FC**	**adj.p-val**	**Gene**	**t-stat.**	**log_2FC**	**adj.p-val**
TBC1D10C	6.86	1.18	0.211	HLA-DQB1	-8.00	-0.69	0.209
SEPT3	6.71	1.04	0.213	CA4	-7.77	-2.71	0.209
PRNCR1	6.60	0.74	0.213	FXYD3	-7.52	-2.56	0.209
TRIM58	6.55	0.85	0.213	RECK	-7.52	-0.60	0.209
SEPHS1P4	6.31	2.26	0.219	SELP	-7.47	-1.93	0.209
POLR2J3	6.25	0.53	0.219	ASNSP1	-7.28	-0.81	0.211
snoU2_19	6.25	3.52	0.219	PPP1R3C	-7.08	-1.93	0.211
NEB	6.05	0.63	0.220	PDGFC	-7.00	-0.49	0.211
DBIL5P	5.75	1.49	0.240	ITGB2	-6.95	-1.51	0.211
EGFL8	5.46	0.59	0.240	SLPI	-6.91	-1.32	0.211
MIA	5.46	1.43	0.240	CA9	-6.83	-2.67	0.211
SYPL2	5.45	0.80	0.240	APOL6	-6.57	-0.82	0.213
ADAM20P1	5.44	1.09	0.240	ENO2	-6.45	-0.49	0.214
CCDC60	5.37	1.78	0.240	BNIP3L	-6.35	-0.83	0.219
BMP5	5.29	1.34	0.240	CCK	-6.33	-1.19	0.219
CLSPN	5.26	0.42	0.240	TIE1	-6.21	-1.23	0.220
ZNF674-AS1	5.12	0.46	0.247	UPK3B	-6.13	-1.45	0.220
MCM3AP-AS1	5.04	0.41	0.247	GABRA2	-6.12	-2.37	0.220
YJEFN3	4.99	0.65	0.247	AKAP7	-6.12	-1.83	0.220
NKTR	4.93	0.43	0.247	ANKRD9	-6.09	-0.76	0.220
UCP3	4.89	0.78	0.247	ABCG1	-5.94	-1.06	0.229
GPR75	4.85	0.93	0.247	GLRB	-5.93	-1.31	0.229
SNORD3A	4.79	1.92	0.247	DCLK1	-5.90	-1.03	0.229
HIST1H4H	4.75	0.52	0.247	ARHGAP4	-5.89	-0.42	0.229
CD200R1	4.74	1.21	0.247	ACSS2	-5.69	-0.51	0.240
LRRC8E	4.73	0.50	0.248	PGM1	-5.68	-0.60	0.240
SNORD104	4.67	0.88	0.250	EGLN3	-5.62	-1.95	0.240
CD3EAP	4.65	0.39	0.250	PARP10	-5.60	-0.65	0.240
NCOA5	4.60	0.31	0.250	VLDLR	-5.57	-0.82	0.240
OFCC1	4.51	0.69	0.250	SAMD9	-5.56	-0.66	0.240

Supplementary Table 4.2 | Top 30 upregulated and downregulated genes in the metastatic organoid *ZFHX4*$^{-/-}$ vs nosgRNA comparison, with respect to *ZFHX4*$^{-/-}$. MIR, CTD-, RP-, LINC-, AC-, NA_mapped genes were removed. To note, VIM was number 33 in upreg.

Upregulated				Downregulated			
Gene	**t-stat.**	**log_2FC**	**adj.p-val**	**Gene**	**t-stat.**	**log_2FC**	**adj.p-val**
CLDN6	7.80	0.90	0.258	MAGEC2	-12.56	-1.58	0.035
ZIC1	7.25	3.73	0.270	FAM86B2	-7.19	-1.49	0.270
IL4I1	7.25	1.12	0.270	PTGDR	-6.72	-2.24	0.350
PCDHB2	6.91	0.95	0.321	NAV2	-5.95	-0.57	0.352
TNFAIP2	6.48	0.60	0.350	TEX41	-5.92	-1.18	0.352
MKRN3	6.44	1.19	0.350	FRMD4A	-5.86	-0.62	0.352
CXCL8	6.33	0.75	0.352	PTPN7	-5.84	-1.27	0.352

CASP1	6.26	1.65	0.352	COL2A1	-5.80	-1.26	0.352
KRT81	5.91	1.43	0.352	KLF15	-5.77	-0.68	0.352
SAMD9	5.89	0.70	0.352	RASGEF1A	-5.46	-0.70	0.431
OR2I1P	5.86	0.98	0.352	DMD	-5.26	-0.48	0.477
RGS4	5.79	1.11	0.352	SPON1	-5.17	-0.75	0.483
RIMBP3	5.79	1.13	0.352	BEX1	-5.16	-1.31	0.483
LCT	5.65	0.94	0.389	PLA2G4B	-5.07	-1.24	0.489
FAM110B	5.59	0.64	0.406	FAR2P2	-5.00	-1.56	0.489
GJC1	5.46	1.00	0.431	ATE1-AS1	-4.93	-0.80	0.489
YTHDF2P1	5.42	1.60	0.435	HS6ST3	-4.93	-1.22	0.489
FGF9	5.27	0.78	0.477	RAMP2	-4.93	-0.92	0.489
KCNJ9	5.25	2.31	0.477	TNFSF8	-4.87	-0.89	0.489
MAGEA1	5.24	1.43	0.477	MARC2	-4.82	-0.56	0.489
SAMD9L	5.16	1.20	0.483	SUCLG2P2	-4.81	-1.05	0.489
ADAM8	5.12	0.77	0.484	NPIPB6	-4.73	-2.16	0.502
NPR3	5.09	1.36	0.487	CLVS1	-4.68	-1.32	0.502
COLEC10	5.02	0.92	0.489	PELI3	-4.65	-0.37	0.504
ADAMTSL2	4.97	0.42	0.489	ART3	-4.39	-1.38	0.569
CA4	4.87	1.70	0.489	SLC7A11-AS1	-4.39	-0.86	0.569
P2RX2	4.83	1.56	0.489	SOSTDC1	-4.39	-0.43	0.569
FERMT2	4.81	0.63	0.489	DNAH6	-4.38	-0.40	0.570
MMP2	4.80	0.66	0.489	SEPT3	-4.37	-0.68	0.570
DNAAF1	4.74	1.70	0.502	MUCL1	-4.32	-0.43	0.590



Supplementary Figure S5.1 | Channel spillover of single-stained controls before compensation. Single-stained controls comprised UltraComp eBeads stained with single antibody to provide negative and positive signal, except for live-dead-NearIR dye which comprised a live-dead control of 50% heat-killed A cells to provide negative and positive signal (no live gating was performed for this figure, only single cells). **(A)** Using the CD4-FITC single stained control as example, flow cytometry gating of single UltraComp eBeads. **(B)** Comparison matrix of single-stained controls by channels to observe spillover. All axes are biexponential to allow easy identification of bending and diagonal patterns, inherent to spillover. Such patterns are highlighted in black boxes and necessitate the need for successful compensation using these single-stained controls. *[A=activated]*