Colorectal cancer: The EMT Transcription Factor ZEB1, The Interaction between Myofibroblast and Colorectal Cancer Cells and GMM Analysis of Cancer Gene Expression

Ta-Chun (Jeff) Liu

Queen's College

A thesis submitted to the Division of Medical Sciences, University of Oxford, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Oncology

Hilary Term, 2021

Weatherall Institute of Molecular Medicine
University of Oxford





I dedicate this thesis to my amazing parents, sister, and family.

Thank you for always being there. Thank you for your unconditional love and unwavering support through ups and downs, late nights and early mornings.

I want to especially dedicate this thesis to my dear grandfather who is fortunately still with us and my other loving grandparents who had passed away. Receiving an education at the most prestigious institution had always been a dream we collectively shared. How I wish you were here with me today to share this moment. I know all of you would be so proud to know my completion of my Ph.D. at the University of Oxford.

Abstract

Colorectal cancer is one of the most prevalent types of cancer. Many factors are associated to the poor prognosis of colorectal cancer including genetic mutations, epigenetic changes, myofibroblast recruitment, and metastasis. Metastasis is the major cause of death in colorectal cancer patients. The most common sites are liver and lymph node metastasis with approximately 20-30% of patients with liver metastasis. Most literature discuss epithelial to mesenchymal transition (EMT) as the molecular mechanism tumor cells undergo to increase motility and invasiveness, ultimately leading to cancer metastasis.

This thesis is consisted of two major sections: computational and experimental work. The first aim of the thesis is to demonstrate our Python-implemented pipeline ***GMMchi*** developed to identify bimodal patterns in a mixture of normal and non-normal components (**Chapter 3**). ***GMMchi*** is further applied onto multiple large datasets to demonstrate its capability in pattern-based analysis (**Chapter 4**). In addition, we built a ***GMMchi***-based scRNA-seq postprocessing pipeline efficient in separating clusters in a population of similar cell types (**Chapter 7**).

In the second aim of this thesis, we studied colorectal cancer metastasis in the context of EMT-like colorectal cancer cell lines and myofibroblast recruitment via co-cultured system consisting of colorectal cancer and myofibroblast cell lines. We showed a novel epigenetic regulation of ZEB1, the gene encoding an EMT transcription factor highly associated with our CRC cell lines exhibiting EMT markers, via CpG island methylation (**Chapter 5**). The second part of this section focused on understanding PDGF-C as an important ligand responsible for myofibroblast and CRC cell line interaction (**Chapter 6**).

In this thesis, we identified a novel mechanism for ZEB1 promoter methylation, revalidated myofibroblast recruitment phenotype previously observed in co-culture assays as well as developed and demonstrated our suite of tools for researchers to carry out pattern-based analysis. (292 words)

Acknowledgements

I would like to first thank all those that have helped me throughout my DPhil. I would especially like to thank the one and only Professor Sir Walter Bodmer FRS. None of the work would be possible without the supervision, guidance, and support from Walter throughout my DPhil journey. Over the past 3 years, it was more than the science that Walter has taught me, it was the enthusiasm, discipline, and relentless motivation to solve difficult problems that have and will continue to inspire me to become a true scientist at heart.

I would like to also thank all the members of the Cancer and Immunogenetics (CIL) Laboratory. For the first half of my DPhil, I want to particularly thank Laura Colling, Tom Barnes, and Marahaini Musa for creating such a supportive and motivated environment. I want to also thank Jenny Wilding, Djamila Ouaret, and especially Viorica Liebe Lastun for your supervision and support. Also, huge thanks to Norah Alrishedan, Gulnar Abdullayeva, Luyao Wang, Stefania Monterisi, and Johanna Michl for keeping the lab life before and during the COVID-19 pandemic not only stimulating but also enjoyable. All of you are amazing scientists and, most importantly, extraordinary people that I am extremely blessed to have had the opportunity to work with and to have known.

Besides, I want to sincerely thank Professor Valentine Macaulay and Professor David Wedge for advices and support given during my transfer of status and confirmation. Thank you, Kevin Clark and Sally Clark, for your expertise in FACS. I would also like to thank Dr. Neil Ashley and Dr. Agne Antanaviciute for all of your help with sequencing and post-analysis. I want to especially thank Dr. Philip Hublitz for your guidance and support on Crispr-related work.

Additionally, I would like to thank the Ministry of Education for providing and supporting me with the scholarship for my DPhil. Lastly, there is no way to end this without mentioning all of the friends that I have met here in Oxford and Queen's College. I would love to list everyone out, but you know who you are, and I want to thank you for all the laughter, encouragements, and friendship. Especially Cindy Tsai, my girlfriend, for always being there and believing in me.

For the past couple years, it has been one hell of a ride. To the years ahead, YOLO and do it all.

Author Declaration

I, Ta-Chun (Jeff) Liu, hereby declare that the whole of this thesis is based on my original work at the Weatherall Institute of Molecular Medicine, University of Oxford Department of Oncology, and that neither the whole work nor any part of it has been, is being, or will be submitted for any other degree at this or any other university. The work is wholly original except as indicated by reference in the text.

Signed: ..... Date: .....

Table of Contents

Chapter 1: General Introduction 1

1.1 An overview of epithelial-to-mesenchymal transition (EMT) 2

1.1.1 EMT pathways 2

1.1.2 The role of EMT in development 3

1.1.3 The role of EMT in cancer metastasis 5

1.1.4 ZEB1-mediated EMT 6

1.1.5 Upstream regulation of EMT Transcription Factors 7

1.1.6 EMT in clinical application 9

1.2 The interaction between colorectal cancer cells (CRC) and myofibroblasts 10

1.2.1 Myofibroblasts in the crypt 10

1.2.2 Stromal fibroblast classification 12

1.2.3 Myofibroblast in Crypt Maintenance 13

1.2.4 Myofibroblasts in Wound healing and Fibrosis 15

1.2.5 Interaction between colorectal cancer cells and myofibroblasts 16

1.2.6 Myofibroblasts in Adenocarcinomas and Primary Tumor 17

1.2.7 Myofibroblast Roles in Tumor Metastasis 19

1.3 Genome-wide Expression analysis in cell lines and tumors 20

1.3.1 Analysis of continuous data 21

1.3.2 Discrete distributions in evolution 25

1.4 Single-cell sequencing analysis 26

1.4.1 Single-cell sequencing postprocessing challenges 26

1.4.2 Single-cell sequencing analysis in cell line data 29

Chapter 2: Methods and Materials..... 30

2.1 Cell culture methods 31

2.1.1 Cell lines 31

2.1.2 Cell culture 32

2.1.3 Cell Line Storage 33

2.1.4 Cell counting 34

2.2 DNA methods 34

2.2.1 DNA extraction from cultured cells 34

2.3 RNA methods 35

2.3.1 RNA extraction 35

2.3.2 cDNA library preparation 35

2.4 Quantitative real-time Polymerase Chain Reaction (qRT-PCR) 36

2.5 Polymerase Chain Reaction (PCR) 37

2.6 DNA and RNA quantification 38

2.7 Gel Electrophoresis of PCR results 38

2.8 siRNA Transient Transfection 38

2.9 Myofibroblast Recruitment Assay 39

2.10 Protein Methods 40

2.10.1 Whole-cell lysate extraction 40

2.10.2 Bradford Protein Assay (BCA) protein quantification 40

2.10.3 Western blot (WB) 41

2.10.4 Membrane stripping 42
2.11 Fixation and Immunofluorescence (IF) 42
2.12 Immunofluorescence imaging 44
2.13 Image Analysis 44
2.14 CRISPR-Cas9 stable cell line 44
2.14.1 Design CRISPR guides 44
2.14.2 Prepare sgRNA expression construct 45
2.14.3 Bacterial Transformation 47
2.14.4 Validating the Insertion into plasmid 48
2.14.5 Plasmid transfection 48
2.14.6 Fluorescence-activated cell sorting (FACS) 48
2.14.7 Cloning 48
2.14.8 Primer Design 49
2.15 Surveyor Assay 49
2.16 CRISPR knockout verification 50
2.17 Statistical methods 50
2.17.1 Microarray data analysis 50
2.17.2 Bimodal gene analysis 50
Chapter 3: A novel Gaussian Mixture Modeling (GMM) Method on Expression Pattern Recognition 52
3.1 Introduction 53
3.2 Results 56

3.2.1 Gaussian Mixture Modelling 56

3.2.2 Current Challenges of GMM on Expression Data Analysis 60

3.2.3 A Novel GMM with Chi-square Pipeline 64

3.2.4 GMMchi validation 77

3.2.5 GMMchi on 78 Colorectal Cancer Cell Line Microarray Data 81

3.2.6 GMMchi Characterization on Microarray Data of a Panel of 78 Colorectal Cancer Cell Lines 85

3.2.7 *GMMchisquare* Python Library 87

Chapter 4: Applications of the GMMchi Pipeline 89

4.1 Introduction 90

4.2 Results 92

4.2.1 2 by 2 Chi-Square Table 92

4.2.2 Probe Combining Pipeline 95

4.2.3 Phenotype Association 99

4.2.4 Well-defined Bimodal Distribution Criterion 131

4.2.5 GMMchi-based Clustering 146

4.2.6 GMMchi-based TCGA Purification 159

Chapter 5: Methylation Controls EMT Transcription Factor ZEB1 166

5.1 Introduction 167

5.2 Results 169

5.2.1 Identifying EMT cell lines through GMMchi 169

5.2.2 Identifying candidate genes 174

5.2.3 mRNA and protein expression level screen in EMT cell lines 177

5.2.4 siZEB1 knockdown in selected EMT cell lines 180

5.2.5 ZEB1-KO CRISPR-Cas9 Design 182

5.2.6 Establishing ZEB1-KO EMT cell lines 184

5.3 ZEB1-KO cell line characterization 192

5.4 Exploring ZEB1 activation in CRC cell lines 220

5.5 5-aza treatment in non-EMT cell lines 224

5.6 Methylation of ZEB1 in large datasets 226

Chapter 6 Myofibroblast CRC Interaction and PDGFC 231

6.1 Introduction 232

6.2 Results 234

6.2.1 Myofibroblast Recruiting CRC Cell Lines 234

6.2.2 Validation on MF Recruitment Phenotype using Different Myofibroblast Cell Lines 238

6.2.3 Establishing PDGFC-KO clones 249

Chapter 7: Single Cell Sequencing Pipeline for Rare Genes 250

7.1 Introduction 251

7.2 Results 254

7.2.1 GMMchi-based Single Cell Sequencing Pipelines 254

7.2.2 Overview of GMMchi_scs_pipeline 266

7.2.3 Dealing with Dropout Rates in scRNA-seq 282

7.2.4 Application of GMMchi-based Single Cell Sequencing Pipeline 285

Chapter 8: General Discussion 294

8.1 GMMchi Efficiently Identifies Mixtures of Normal and Non-normal Components in Gene Expression Analysis..... 295

8.2 GMMchi Provides a Powerful Approach to Many of Aspects in Gene Expression Analysis 298

8.3 Methylation is a Novel Mechanism Regulating ZEB1 Expression in CRC Cell Lines 302

8.4 GMMchi-based scRNA-seq Postprocessing Pipeline Efficiently Removes Low-Quality Barcodes in a Mix of Similar Cell Type 305

References 309

Appendix……326

List of Figures

Figure 1.1 Schematic of the 3 Types of EMT 3

Figure 1.2 Schematic of Wound Healing 4

Figure 1.3 Schematic of the Three Major Pathways for EMT Induction 8

Figure 1.4 Illustration of The Organization of The Colonic Crypt 12

Figure 1.5 Schematic Showing The Gradient of BMP, Wnt, and GREM in The Colonic Crypt 14

Figure 1.6 Schematic of CRC Cancer Progression 17

Figure 3.1 Common Distributions in Gene Expression Analysis 55

Figure 3.2 GMM on Simulated Mixture of Normal Distributions. ... **Error! Bookmark not defined.**

Figure 3.3 Dynamic Binning. 66

Figure 3.4 Determining the Chi-Square Threshold Value for Identifying a Poorly Fitted Tail 69

Figure 3.5 Examples of Iterative Tail Trimming on CDH1 and MUC2 73

Figure 3.6 Examples of Different Distributions Found in Gene Expression Analysis 74

Figure 3.7 Example Output of GMMchi Python Package 76

Figure 3.8 GMMchi's Accuracy of Tail Identification 80

Figure 3.9 GMMchi's Accurarcy on the Correct Categorization of Simulated Distributions 81

Figure 3.10 Determining Background Threshold in Microarray Data Using GMMchi 83

Figure 3.11 Examples of Genes Exhibiting Categories in Unimodal Distributions 85

Figure 3.12 GMMchi on Gene Expression Data 87

Figure 4.1 Formula for Calculating Chi-Square and R-value 94

Figure 4.2 Schematic of the Probe Combining Pipeline. 98

Figure 4.3 Examples of GMMchi Analysis on Known Differentiation Markers 102

Figure 4.4 GMMchi Analysis on Genes Associated with Goblet Cells. 111

Figure 4.5 Volcano Plot of RER-related Differentially Expressed Genes. 123

Figure 4.6 GMMchi Analysis on the 22 Overlapping RER-related Genes 130

Figure 4.7 GMMchi Analysis on the 7 Non-Overlapping RER-related Genes Reveal a General Trend in Association..... 131

Figure 4.8 Genes Used to Determine the Criteria for Well-Defined Bimodal Distributions 135

Figure 4.9 Examples of GMMchi Analysis on Well-Defined Bimodal Genes 136

Figure 4.10 10 Overlapping Well-Defined Bimodal Genes in Both Datasets..... 137

Figure 4.11 Correlations of the Same Genes Across Both Datasets 139

Figure 4.12 GMMchi Analysis on Identifying Threshold for Linkage Distance 148

Figure 4.13 Heatmap of the Correlation Matrix Between Gene Clusters 158

Figure 4.14 Comparison of Gene Expression Distributions in Our Microarray Data vs TCGA.... 161

Figure 4.15 Determining the Background Threshold for TCGA RNA-seq Data..... 161

Figure 4.16 Example of MUC2 Gene Expression Distribution in TCGA..... 163

Figure 4.17 Examples of Before and After Purification 165

Figure 5.1 CDH1 Expression of 141 CRC Cell Lines 170

Figure 5.2 CDH1 and VIM Expression on 70 CDH1 Wildtype CRC Cell Lines 172

Figure 5.3 GMMchi and 2x2 Table Analysis on CDH1-/VIM+ CRC Cell Lines 173

Figure 5.4 Volcano Plot of EMT vs Non-EMT Cell Lines. 175

Figure 5.5 ZEB1 Microarray, WB, and qRT-PCR Characterization of EMT Cell Lines 180

Figure 5.6 siZEB1 Knockdown Experiments in OXCO1..... 181

Figure 5.7 siZEB1 Knockdown Experiments in SW480 and CCO7 182

Figure 5.8 Designed Guides for ZEB1-KO. 183

Figure 5.9 PCR and WB of ZEB1-KO EMT CRC Cell Lines Clones 191

Figure 5.10 WB Detecting ZEB and CDH1 Expression in ZEB1-KO CRC Clones 193

Figure 5.11 Volcano Plot of CDH1-upregulated vs CDH1-downregulated ZEB1-KO CRC Cell Lines 195

Figure 5.12 ZEB1, CDH1, β-catenin, and VIM Immunofluorescent Staining on ZEB1-KO Cell Lines 196

Figure 5.13 Immunofluorescence Staining of β-catenin Reveals Nuclear vs Membrane Staining 210

Figure 5.14 qRT-PCR Results of ZEB1, VIM, and CDH1 mRNA Levels in ZEB1-KO Cell Lines 213

Figure 5.15 Results of Scratch Assay and Doubling Time Assay Comparing ZEB1-KO vs Wildtype Cell Lines 215

Figure 5.16 Scratch Assays of ZEB1-KO vs Wildtype EMT Cell Lines..... 216

Figure 5.17 ZEB1 qPCR Comparing Normoxic vs Hypoxic Conditions 221

Figure 5.18 siAKT3 Knockdown Experiment Reveals No Change in ZEB1 Expression 221

Figure 5.19 5-azacytidine Treatment Upregulated ZEB1 Expression in Non-EMT Cell Lines. 223

Figure 5.20 ZEB1 qRT-PCR on 5-aza-treated Non-EMT Cell Lines 225

Figure 5.21 ZEB1 Upregulation Observed in 5-aza-treatment on a Panel of 20 CRC Non-EMT Cell Lines: 226

Figure 5.22 ZEB1 is Methylated in Low ZEB1-Expressing Samples in The TCGA Data. 228

Figure 5.23 ZEB1 Expression is Negatively Correlated with ZEB1 Methylation in the CCLE Database 230

Figure 6.1 Schematic of Coculture Assay 236

Figure 6.2 Volcano Plot Reveals PDGFC and CXCR4 as Potential Candidates that Contribute to Myofibroblast Recruitment 237

Figure 6.3 GMMchi Analysis on CXCR4 and PDGFC. 238

Figure 6.4 Recruitment Assay of CRC Cell Lines Cocultured with Other Myofibroblast Cell Lines Reconfirms Previous Observations 241

Figure 6.5 ZEB1-KO Blocks Myofibroblast Recruitment in OXCO1 248

Figure 7.1 UMAP of scRNA-seq Data Postprocessed with scprep 255

Figure 7.2 Outputs of each Step in Postprocessing scRNA-seq Data with scprep 257

Figure 7.3 Gradient in Highly-expressed Genes Found in Postprocessed scRNA-seq Data with scprep 260

Figure 7.4 Correlation Matrix of Highly Expressed Genes 263

Figure 7.5 Library Size Filter Removes Doublets and Empty Droplets 267

Figure 7.6 Library Complexity Filter Removes Low-quality Barcodes 270

Figure 7.7 Housekeeping Gene Filter Results in a Better Seperation. 273

Figure 7.8 Removing Barcodes Expressing High Levels of Mitochondrial Genes Removes Apoptotic Cells. 274

Figure 7.9 Normalization Results in a Better Separation 275

Figure 7.10 Final Output of GMMchi_scs_pipeline 277

Figure 7.11 GMMchi_scs_pipeline Effectively Removes Expression Gradients in Highly Expressed Genes 280

Figure 7.12 Library Complexity and Size are Evenly Distributed 281

Figure 7.13 Comparing UMAP Prior and After MAGIC 283

Figure 7.14 K-means Clustering 285

Figure 7.15 scRNA-seq Shows a Striking Difference in CDX1 and CDX2 Levels in Lumen vs Non-Lumen-Forming Cell Lines 287

Figure 7.16 REG4 and TFF3 Shows Inclusion of MUC2 on a Single Cell Level 293

Supplementary Figure S1 Graph Showing GMMchi's Accuracy of Each Category 327

Supplementary Figure S2 Accuracy of GMMchi Based on the Overlap Between Two Normal Components 328

Supplementary Figure S3 Doubling Time Assay of ZEB1-KO vs Wildtype 329

Supplementary Figure S4 Surveyor Assay to Determine which Guides to Use for CRISPR Knockou 330

List of Tables

Table 2.1 Summary of All CRC and MF Cell Lines 32

Table 2.2 Cycling Times for qRT-PCR 36

Table 2.3 Summary of TaqMan Gene Expression Probes. 37

Table 2.4 Cycling Times for PCR 38

Table 2.5 Summary of siRNA 39

Table 2.6 Summary of Antibodies for WB 42

Table 2.7 Summary of Antibodies for IF 44

Table 2.8 Cycling Times for Annealing sgDNA 45

Table 2.9 Cycling Times for Ligating and Inserting sgDNA into Plasmid 46

Table 2.10 Primer Design for CRISPR Guide Sequencing 48

Table 2.11 Summary of Primers Used for Sequencing KO Region for ZEB1 49

Table 4.1 Genes Associated with Lumen Formation 105

Table 4.2 Genes Associated with Goblet-Cell Formation 112

Table 4.3 Differences in ZEB1 Probes 117

Table 4.4 Genes Associated with EMT 119

Table 4.5 Genes Associated with RER Status. 124

Table 4.6 Summary of Well-Defined Bimodal Genes in Our Microarray and CCLE Data 135

Table 4.7 Mutations Associated with Well-Defined Bimodal Genes 140

Table 4.8 Methylations Associated with Well-Defined Bimodal Genes 144

Table 4.9 Enhancers Associated with Well-Defined Bimodal Genes 146

Table 4.10 Gene Clusters Associated with Cell Line Phenotype 153

Table 4.11 Clusters Associated with Cell Line Phenotype. 157

Table 5.1 Top 25 Differentially Upregulated and Downregulated Genes in EMT Cell Lines 176

Table 7.1 An Example of Input Data for the GMMchi_scs_pipeline..... 265

Table 7.2 Genes Associated with MUC2+ Single Cells 293

Supplementary Table S1 List of All Genes Significantly Associated with Methylation..... 331

Supplementary Table S2 Full List of Genes Assigned to Each Gene Cluster..... 334

Supplementary Table S3 Full List of Gene Cluster Assigned to Each Cluster 382

List of Abbreviations

AKT	Serine/threonine-protein kinase
ASMA	Alpha-smooth muscle actin
AOC3	Amine oxidase, copper containing 3
ARNT1	Hydrocarbon nuclear translocator 1
BIC	Bayesian information criterion
CDH1	E-Cadherin
C.E.	Critical exon
CAF	Cancer-associated fibroblast
CRISPR	Clustered regularly interspaced short palindromic repeats
CRC	Colorectal cancer cell
dNTP	Deoxynucleotide
EM	Expectation maximization
EMT	Epithelial-to-mesenchymal transition
ECM	Extracellular matrix
EGF	Epidermal growth factor
FBS	Fetal bovine serum
FACS	Fluorescent-activated cell sorting
FGF	Fibroblast growth factor
GFP	Green fluorescent protein
GMM	Gaussian mixture modeling
HIF1α	Hypoxia-inducible factor 1 alpha
HIF-1	Hypoxia-inducible factor-1
HGF	Hepatocyte growth factor
IL-1	Interleukin-1
IL-6	Interleukin-6
IGF1	Insulin-like growth factor 1
MF	Myofibroblast
MMP	Metalloproteinase
mL	Milliliter
mm	Millimeter
µg	Microgram
µL	Microliter

µm	Micrometer
µM	Micromolar
nM	Nanomolar
mTORC	Mammalian TOR complex
NKX2-3	NK2 homeobox 3
KO	Knockout
KD	Knockdown
PI3K	Phosphoinositide 3-kinase
PCR	Polymerase chain reaction
PDGF	Platelet-derived growth factor
PHD	Prolyl hydroxylase protein
qRT-PCR	Quantitative real-time polymerase chain reaction
Ruby	Red fluorescent protein
RT	Room temperature
ROS	Reactive oxygen species
RTK	Tyrosine kinase receptors
SMAD	Mothers against decapentaplegic homolog
SNAI	Snail family transcriptional repressor
SHOX2	Short stature homeobox 2
TF	Transcription factor
TGFBR2	Transforming growth factor beta receptor 2
TWIST	TWIST family BHLH transcription factor
TGFβ	Transforming growth factor-beta
TCGA	The Cancer Genome Atlas
TNF-α	Tumor-necrosis factor alpha
VIM	Vimentin
VHL	Von Hippel-Lindau protein
WB	Western blot
ZEB	Zinc finger E-box binding homeobox
5-aza	5-azacytidine

Chapter 1: General Introduction

1.1 An overview of epithelial-to-mesenchymal transition (EMT)

Epithelial-to-mesenchymal transition (EMT) has long been suggested to be associated with colorectal cancer tumor metastasis (Banyard and Bielenberg, 2015). Thus, the study of EMT in colorectal cancer has long been thought to provide a glimpse into the possible mechanism of pro-metastatic outgrowth. This section will focus on known mechanisms of EMT and an introduction into the role of EMT under different biological settings.

1.1.1 EMT pathways

Epithelial-to-mesenchymal transition is a biological process that occurs in multiple stages during development (Heerboth *et al.*, 2015; Brabletz *et al.*, 2018). EMT is a process in which cells lose their cellular adhesion and apical-basal polarity, activating a change in the genetic landscape that allows them to transform into mesenchymal cells that possess increased mobility, invasiveness, and cellular plasticity (Loboda *et al.*, 2011; Vu and Datta, 2017). Many pathways govern the activation and silencing of these markers. However, all mechanisms initiate EMT through the upregulation of EMT transcription factors (TFs) such as SNAI1, SNAI2, TWIST1, ZEB1, ZEB2…etc (Garg, 2013; Stemmler *et al.*, 2019). To summarize, the induction of EMT mainly includes the transforming growth factor-beta (TGFβ), tyrosine kinase receptors (RTKs), and the hypoxic signaling pathway (Zhang, Tian and Xing, 2016; Zhu *et al.*, 2018). Other observed EMT pathways also include the Wnt, Notch, and Hedgehog pathways (Wang *et al.*, 2010; Wu *et al.*, 2017).

1.1.2 The role of EMT in development

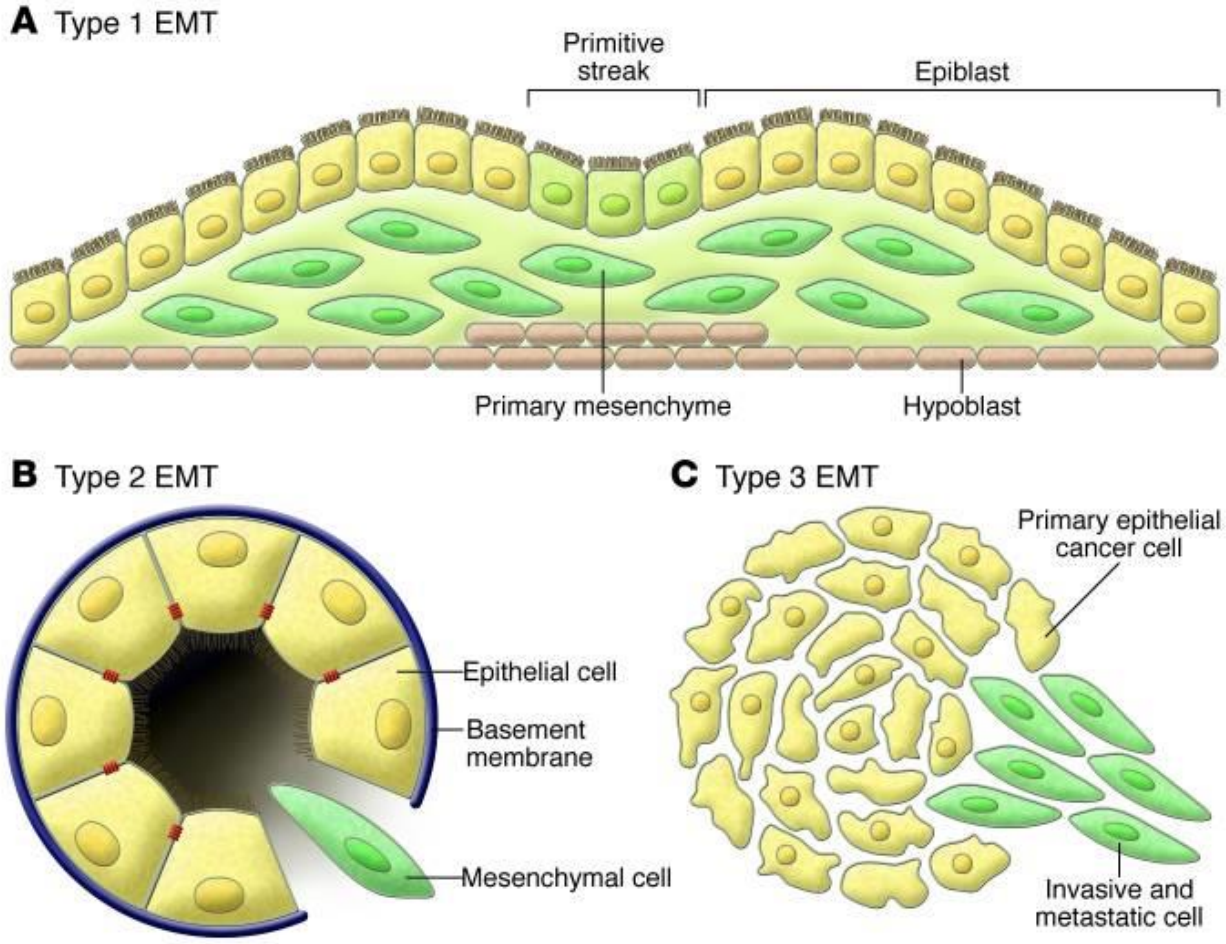


Figure 1.1 Schematic of the 3 Types of EMT: Schematic of the 3 types of EMT found during development, wound healing, and tumor metastasis (Kalluri and Weinberg, 2009)

EMT occurs in three distinct biological settings shown in Figure 1.1 thus separating EMT into 3 major types (Figure 1.1) (Kalluri and Weinberg, 2009). Type I EMT is mainly associated with embryo implantation, embryo formation, and early organ development. During the early stages of embryogenesis, the trophectoderm cells undergo EMT to invade the endometrium and ensure proper anchoring into the placenta. Furthermore, when a fertilized egg forms the three germ layers via gastrulation, the primitive streak undergoes EMT to migrate and ingress to form the mesoderm and endoderm. Another case of EMT involves neuroectoderm transforming into

migratory neural crest cells in which enables it to disperse across different regions of the embryo to form for example melanocytes, bone, and smooth muscle. Many of the Type I EMTs undergo MET (Mesenchymal to Epithelial transition) or the reverse transformation from mesenchymal to epithelial cells to form secondary epithelia after migrating to the necessary region for development and further differentiation.

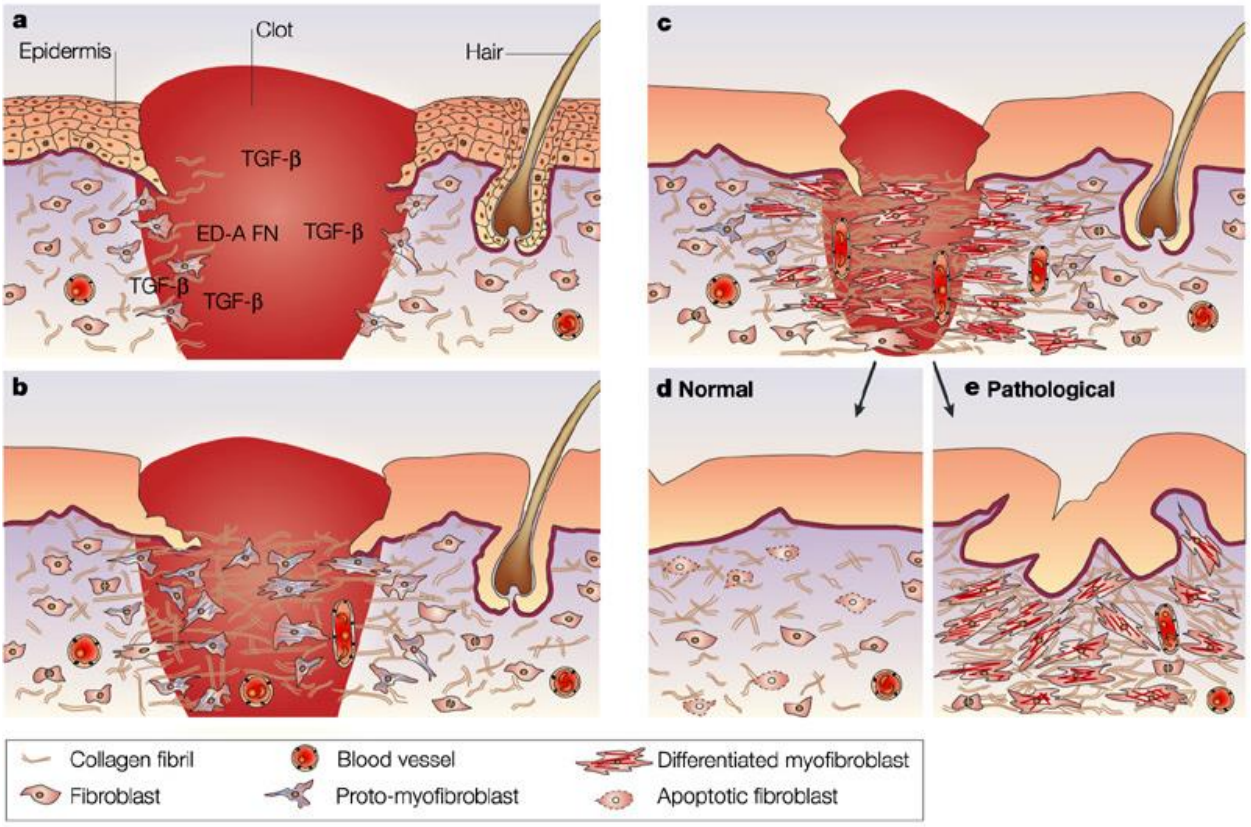


Figure 1.2 Schematic of Wound Healing: Schematic showing the process of wound healing (Brown *et al.*, 2002)

In Type II EMT, this type of EMT is mostly associated with wound healing (Figure 1.2), organ fibrosis, and tissue regeneration that takes place normally during a repair-induced event (Brown *et al.*, 2002). Different from Type I EMT whereby a careful balance of signaling factors

induces EMT to enable proper development of an embryo, Type II EMT is mainly triggered by inflammation. During a wound-healing event, inflammatory epithelial tissues and fibroblasts release inflammatory signals and growth factors such as TGF-β, PDGF, EGF, and FGF-2. This subsequently activates angiogenesis and deposition of new extracellular matrix (ECM) and the re-epithelialization of keratinocytes. During normal conditions, keratinocytes will differentiate until cell death at the outer layer of the epidermis. During an inflammatory event, they undergo EMT to move to the wound to reverse and re-epithelialization to close the wound. Nonetheless, it is observed that the overactivation and inability to remove excess fibroblasts leads to organ fibrosis which is destructive to the local structure of the organ.

Lastly, Type III EMT is observed in cancer metastasis and migration to a secondary site. Type III EMT is discussed more in 1.3. The role of EMT in cancer metastasis.

1.1.3 The role of EMT in cancer metastasis

Cancer cells undergoing epithelial-to-mesenchymal transition have long been observed in various formats and it has been suggested that EMT plays a role in cancer metastasis. EMT TFs, including especially SNAI1, SNAI2, SNAI3, TWIST1, ZEB1, ZEB2 have been shown to be upregulated and to regulate the process of EMT in many cancer models. They inhibit epithelial-like cancer cells to secrete adhesion molecules and allow them to detach from their neighboring niche and migrate to a distant organ at a secondary site (Aiello and Kang, 2019). Nonetheless, EMT in the tumor metastasis setting is highly dynamic and implicitly transient and reversible. In other words, EMT can give rise to circulating tumor cells (CTCs) that migrate to another site and reverses back to epithelial cells before growing at the new site via MET (Jie, Zhang and Xu, 2017).

However, it is important to note that whether EMT is the primary path cancer cells escape from their primary niche and metastasize requires further investigation.

1.1.4 ZEB1-mediated EMT

The hallmark EMT marker that is widely accepted in the literature is the upregulation in Vimentin (VIM) and the downregulation of E-Cadherin (CDH1) (Schliekelman *et al.*, 2015; Cheng *et al.*, 2016; An, Zheng and Dann, 2017). VIM encodes the protein Vimentin, which is a type III intermediate filament. During EMT, the transition of an epithelial cell undergoes a dramatic change in shape whereby the cell becomes elongated and flattened. At the same time, it is shown that during this transformation epithelial cells lose keratin intermediate filaments (KIFs) and gain vimentin intermediate filaments (VIFs). This change coincides with the downregulation of CDH1, the gene encoding the protein E-Cadherin, under the cadherin superfamily, which is a cell adhesion molecule present in most epithelial cells. The cadherin superfamily is a collective member of transmembrane proteins classified based on the presence of one or more cadherin repeats in the extracellular domain. The main role of cadherins is to mediate cell-cell adhesions. ZEB1, encoding the zinc finger E-box binding homeobox 1, mediates EMT through the binding on to the TATA region located in the promoter region of a wide variety of genes including VIM and CDH1. The TATA region is a region consisted of a TATA sequence flanked by two palindromic sequence motifs known as E boxes that encode the sequence CANNTG (Malik, Huang and Schmidt, 1995; Dykes *et al.*, 2017). This specific binding site enables upregulation of ZEB1 to effectively downregulate CDH1 expression and upregulate VIM expression along with other genes resulting in the loss of apical-basal cell polarity, undergo a reorganization in the cytoskeleton, and increase cellular plasticity and motility.

1.1.5 Upstream regulation of EMT Transcription Factors

TGFβ mediates through two major pathways: SMAD-mediated and non-SMAD mediated signaling (Miyazono, 2009; Xu, Lamouille and Derynck, 2009). SMAD-mediated signaling initiates with the binding TGFβ to 'type II' TGFβ family receptors (TGFBR2). This phosphorylates 'type I' transmembrane kinases that activate intracellular SMAD effectors through phosphorylation of the C-terminus. SMAD2 and/or SMAD3 forms a trimeric SMAD complex with SMAD4 that allows it to translocate into the nucleus to combine with DNA-transcription factors to initiate activation of EMT TFs. The Non-SMAD signaling pathway begins with TGFβ activating serine/threonine-protein kinase (AKT) through phosphoinositide 3-kinase (PI3K). This results in the activation of mammalian TOR complex (mTORC) I&II and subsequent activation of EMT TFs (Miyazono, 2009) (Figure 1.3).

Tyrosine kinase receptors are transmembrane RTKS that activate the PI3K-AKT pathway and the ERK, MAPK, and JNK pathways during the binding of relevant growth factors (GFs) (Uribe *et al.*, 2017) (Figure 1.3). A wide range of GFs is capable of RTK-binding. These GFs include fibroblast growth factor (FGF), hepatocyte growth factor (HGF), insulin-like growth factor I (IGFI), epidermal growth factor (EGF), and platelet-derived growth factor (PDGF).

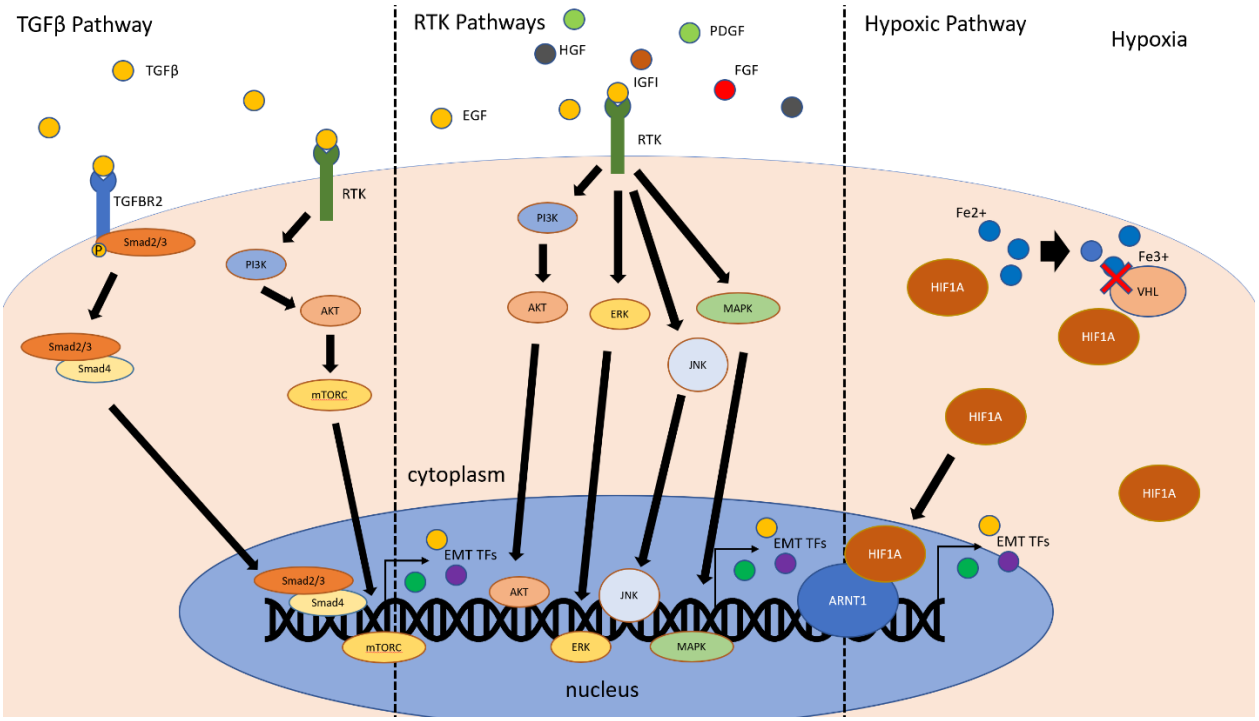


Figure 1.3 Schematic of the Three Major Pathways for EMT Induction

The hypoxic pathway is a universal pathway in response to hypoxia. The literature suggests that multiple molecular mechanisms are activating the hypoxic pathway (Zhang *et al.*, 2015; Choi *et al.*, 2017). The focus of this overview is to give a general overview so here we discuss one proposed mechanism. Under hypoxia, excessive oxidative stress caused by the accumulation of reactive oxygen species (ROS) oxidizes intracellular iron ions from a +2 to a +3 state (Qutub and Popel, 2008). This prevents the proper binding of oxidized ferric ion to prolyl hydroxylase proteins (PHD) (Smith, Robbins and Ratcliffe, 2008). Without the PHD complex, hypoxia-inducible factor 1-alpha (HIF1α) accumulates due to the lack of hydroxylation and subsequent ubiquitination by Von Hippel-Lindau protein (VHL) for proteasomal degradation. HIF1α binds to aryl hydrocarbon nuclear translocator 1 (ARNT1), enters the nucleus, and initiates transcription of various genes including EMT TFs (Figure 1.3).

1.1.6 EMT in clinical application

The characterization of EMT is the change in the expression of EMT-associated genes that have been used to establish the EMT status of a clinical sample (Santamaria *et al.*, 2017; Yang and Zhang, 2017). This information is used to correlate with the clinical significance of the tumor. Nonetheless, pathologists have argued the significance of assessing EMT in a clinical setting in tumor development. This is mainly due to the difficulty in assessing EMT properties in biopsies due to its transient status. It is particularly difficult to capture the right snapshot of tumor development because cells that are undergoing transient EMT and are in the stage of shifting between E and M states appear at various and transient time points while cells that have fully committed in the M state will morphologically appear to resemble stromal cells that line and surround the tumor (Kröger *et al.*, 2019). This makes clinical assessment through histopathological means particularly difficult.

Thus, the focus has been turned to molecular identification of protein expression levels in maximizing the value of EMT assessment. Other than E-cadherin and Vimentin, EMT TFs have been regularly incorporated in assessing cancer progression. For example, SNAI1 nuclear expression was found to be clinically significant in identifying invasive ductal breast carcinomas and in breast tumors more generally. Furthermore, ZEB1 was also found associated with prostate cancer progression and metastasis. ZEB2 expression was also found related to tumor recurrence in hepatocellular carcinomas. Recently, *Fici et al.* established and have shown a scoring metric to determine the EMT index with potential prognostic value in a sample cohort of breast carcinoma samples (Fici *et al.*, 2017). Nonetheless, assessing protein expression levels of EMT TFs such as those mentioned above including ZEB1, ZEB2, or SNAI1 is infrequently used for diagnostic

purposes but mostly used to assess patient outcome and therapy resistance under a clinical setting.

Due to the unique characteristics of EMT including increasing the cell resistance to radiotherapy, stemness, and the implied association with metastasis in which dramatically decreases the survival rate of the patient (Weidenfeld and Barkan, 2018; Zhou *et al.*, 2020), EMT is naturally an important phenomenon in cancer biology that requires further careful study.

1.2 The interaction between colorectal cancer cells (CRC) and myofibroblasts

Myofibroblasts are a mesenchymal cell type that lines the full extent of the colonic crypt axis, known to secrete a wide range of signaling factors that modulate and balance crypt homeostasis (Klingberg, Hinz and White, 2013; David *et al.*, 2020). The important role myofibroblasts play under normal conditions alludes to its importance in a tumor setting. In this section, we explore the crosstalk between colorectal cancer tumors and cancer-associated myofibroblasts.

1.2.1 Myofibroblasts in the crypt

The crypt is the functional unit of the colonic epithelium and is lined by an extracellular matrix (ECM) that maintains the three-dimensional structure with different cell types that support the colonic stem cell niche (Fre, 2015). The colonic crypt is composed mainly of three cell types originating from the same progenitor cells: goblet cells, enterocytes, and enteroendocrine cells. Myofibroblasts secrete collagen type IV, the main basement membrane that makes up the thin layer of ECM separating them from the epithelial cells (Figure 1.4) (Klingberg, Hinz and White,

2013). This adjacent structure provides neighboring crosstalk between the two cell types and their environment.

The surface of the crypt is lined by epithelial cells and has a turnover time of 4-5 days. The high turnover rate is maintained by stem cells that reside at the bases of the crypts, which extend into the lamina propria (Humphries and Wright, 2008). The homeostasis of the constant turnover is dependent on the regulation of crypt stem cells as well as transit-amplifying (TA) cells (Medema and Vermeulen, 2011). It has been shown that disruption of this homeostasis due to accumulated mutations or epigenetic changes that lead to disruptive phenotypes can likely lead to sequences of events that result in a paradigm shift towards colorectal oncogenesis (Walther *et al.*, 2009).

Within the bottom of the crypt, the stem cell niche consists of a microenvironment with molecular signals where intestinal stem cells consistently self-renew and are maintained in the undifferentiated state (Humphries and Wright, 2008). As intestinal stem cells divide, their progeny migrate upwards the walls of the crypt and partially differentiate in TA cells. The TA cells continuously migrate upwards the crypt and terminally differentiate into the 3 primary cell types of the colonic epithelium mentioned above: goblet cells secrete mucin into the lumen, enterocytes responsible for absorptive functions, and enteroendocrine cells secreting hormones that modulate intestinal epithelium and smooth muscle function (Fre, 2015). Around the crypt lie the myofibroblast below the basement membrane that extends along the entire walls of the crypt. In the literature, myofibroblasts have been implicated to play an important role in the maintenance of crypt stem cell niche and the regulation of TA cell differentiation (Marshman, Booth and Potten, 2002; Snippert *et al.*, 2010).

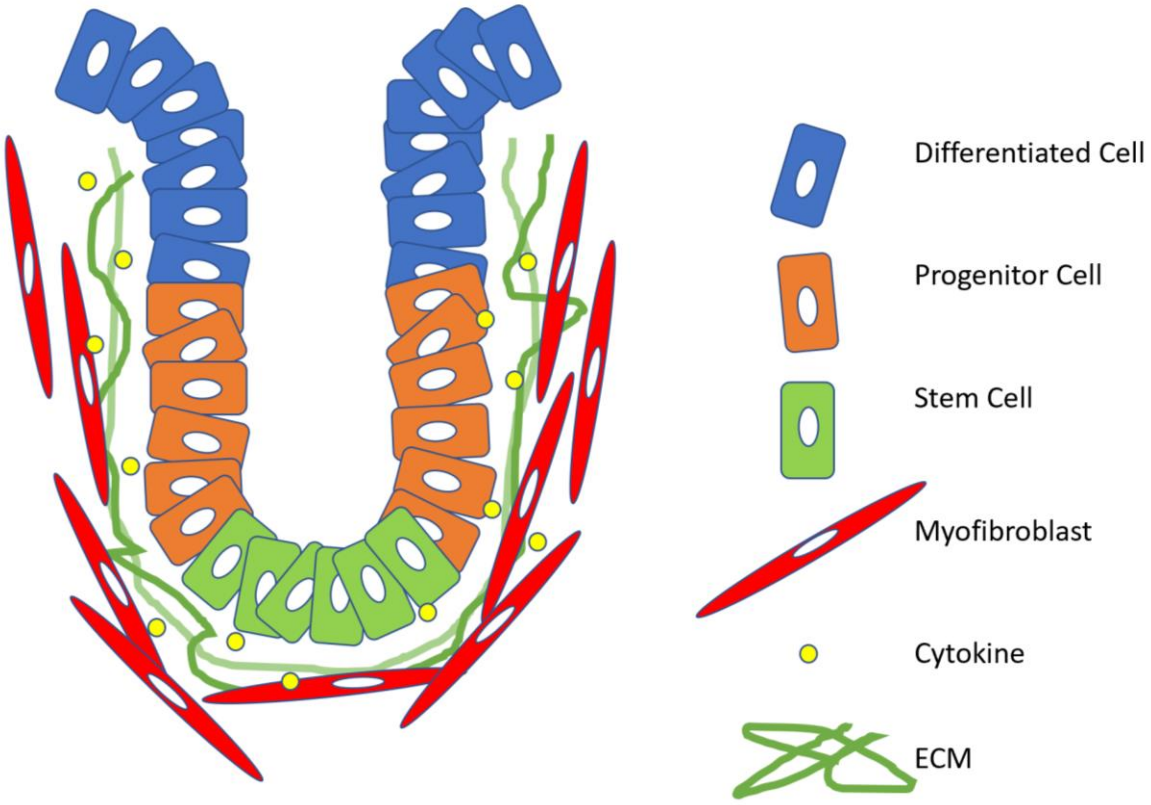


Figure 1.4 Illustration of The Organization of The Colonic Crypt

1.2.2 Stromal fibroblast classification

In the tumor environment, stromal fibroblasts associated with a cancer setting are sometimes referred to as cancer-associated fibroblasts (CAFs) or tumor-associated fibroblasts (Tsujino *et al.*, 2007; Baum and Duffy, 2011). Their recruitment to the tumor niche is mainly initiated with cytokines secreted by cancer cells and infiltrating leukocytes. A publication in 1999 by Olumi et al. first defined CAFs from conventional fibroblasts in prostate cancer using a set of specific biomarkers that revealed a distinct expression pattern (Leigh, 2000). However, the lack of definitive understanding of stromal fibroblasts gave rise to a complex mix of contradicting names in the literature. A recent study shows that CAFs are a form of activated MFs that possess

an increase in alpha-smooth muscle actin (αSMA) expression suggesting an improved ability of contractile motion and heightened mobility. Myofibroblasts are found in the normal physiological setting including lymph nodes, blood vessels, lung septa, and the stroma of the testis (Bochaton-Piallat, Gabbiani and Hinz, 2016). Nonetheless, they are also found within the tumor setting in varying amounts correlated with the stage of the tumor. In our lab, we have shown that a specific set of markers AOC3+, NKX2-3+, SHOX2- can distinguish myofibroblasts from other forms of fibroblast[11]. Furthermore, TGFβ-treated myofibroblasts (activated myofibroblasts) reveal a downregulation in *AOC3* suggesting a potential marker for an activated state (Hsia *et al.*, 2016).

1.2.3 Myofibroblast in Crypt Maintenance

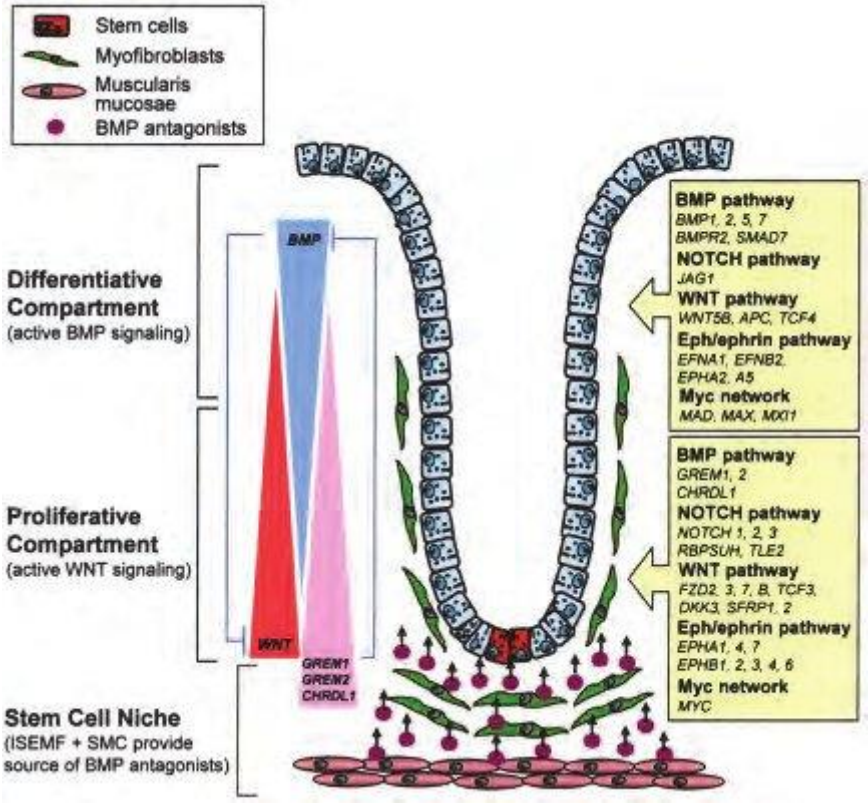


Figure 1.5 Schematic Showing The Gradient of BMP, Wnt, and GREM in The Colonic Crypt (Maric *et al.*, 2012)

Myofibroblasts that line the colonic crypt play a crucial role in the maintenance of the stem cell niche via classical morphogenetic pathways. The gradient of the WNT pathway as well as the bone morphogenetic protein (BMP) pathway has long been implicated to regulate the homeostasis in the colonic crypt (Figure 1.5) (Maric *et al.*, 2012). A majority of sporadic colorectal cancers exhibit inactivating mutations in adenomatous polyposis coli (APC) or activating mutations in β-catenin (Kinzler and Vogelstein, 1996; Reya and Clevers, 2005). It was found that nuclear β-catenin is mostly found in the crypt where the stem cell niche lies as shown in the WNT gradient from bottom to top (Batlle *et al.*, 2002). Interestingly, even though Paneth cells are responsible for secreting WNT ligands such as Wnt3, Wnt2, Wnt5a in the small intestine, the colon lack these cell types, and it is thought that pericryptal myofibroblasts are important cell types that secrete WNT stimulation to maintain the crypt stem cell niche (Pinto *et al.*, 2003; Kuhnert *et al.*, 2004).

BMPs are a group of growth factors that belong to the TGF- β superfamily. BMP signals through Smad-dependent pathway (Hardwick *et al.*, 2008), also considered as the canonical pathway, as well as other non-canonical pathways including MAPK. Extracellularly, BMP is modulated by Noggin and has been shown to play an important role in limiting crypt stem cell expansion into the upper compartment of the crypt and to prevent malignant hyperproliferation. Moreover, it plays a critical role in controlling TA cell proliferation and differentiation. In the mouse model, inactivation of BMP via overexpression of Noggin leads to an overproliferation and expansion of the mouse crypts (Haramis *et al.*, 2004; He *et al.*, 2004). Opposite to the WNT pathway gradient, the BMP gradient is the highest closest to the top of the crypt and decreases

as it approaches the bottom of the crypt. It was observed that BMP antagonists GREM1 and GREM2 have an expression pattern opposite to the BMP gradient. GREM1 and GREM2 are pericryptal myofibroblasts secreted glycosylated proteins and are shown to restrict the activity of the BMP (Auclair *et al.*, 2007; Kosinski *et al.*, 2007). Thus, it is thought that myofibroblasts that line the wall of the intestinal crypt are the main source of BMP antagonist responsible for modulating the level of BMP, which results in balanced homeostasis along the crypt vertical axis. Furthermore, it has also been shown that myofibroblasts that line the top compartment of the crypt secrete BMP ligands. Thus, it has been suggested that there exist different myofibroblasts along the vertical axis of the crypt that secrete different cellular molecules such that crypt-top myofibroblasts that exhibit high PDGFR-A levels secrete BMP ligands while the crypt-bottom myofibroblasts express low PDGR-A levels and secrete BMP antagonists that result in the gradient of BMP along the crypt (David *et al.*, 2020).

1.2.4 Myofibroblasts in Wound healing and Fibrosis

Mild wound healing in the colonic crypt is slightly different from deep wound healing that damages the basement membrane and subepithelial tissue as described in Type II EMT. Mild injury in the colonic crypt leaves the basement membrane intact, usually caused by mechanical scraping, lack of mucin that leads to mechanical abrasion, or mild epithelial exfoliation such as intestinal responses that occurs when the intestine encounters toxins (Moore, Carlson and Madara, 1989; Wallace and Granger, 1996). Damages in the epithelium are quickly replenished by mobilizing epithelial cells near the wound. Epithelial cells move along the basement membrane to join and close the wound to reform tight junctions. Different from deep wound healing, this process does not require inflammatory cytokines and is usually followed by cellular

proliferation to replenish the lost cells (Silen and Ito, 1985). In the literature, it has been reported that the restitution of the colonic epithelium is promoted by pericryptal myofibroblasts through secreting various cellular factors such as TGF- β, stromal-derived factor-1 (CXCL12), and hepatocyte growth factor (HGF) (D.W. *et al.*, 1999; McKaig *et al.*, 1999; Moyer *et al.*, 2007; Iizuka and Konno, 2011). Nonetheless, if the colonic crypt were to sustain substantial damage, the response to wound healing is analogous to what happens in deep wound healing in the skin. Pericyptal myofibroblasts similar to skin fibroblasts are thought to enter an activated state due to inflammatory cytokine. They will exhibit increased motility and migrate into the wound site while contributing to ECM synthesis. Under conditions such as serious injury or disease that damages the colonic epithelium including inflammatory bowel disease (IBS), cystic fibrosis … etc. (Bochaton-Piallat, Gabbiani and Hinz, 2016), they are known to result in fibrosis similar to deep wound healing in the skin. All in all, this indicates the important role myofibroblast plays in colonic wound healing (Rieder *et al.*, 2007).

1.2.5 Interaction between colorectal cancer cells and myofibroblasts

The tumor environment is an important makeup for the cancer cells, fostering growth and potentially contributing to tumor malignancy (Takatsuna *et al.*, 2016). The bidirectional crosstalk between myofibroblasts and cancer cells includes the secretion of a variety of cytokines and growth factors namely tumor-necrosis factor-α (TNF-α), interleukin-1 (IL-1), interleukin-6 (IL-6), TGFβ, EGF, PDGF (Rogler *et al.*, 2001). Some of these secreted factors are capable of inducing the EMT pathway thus contributing to the increased migration and invasiveness of the cancer cells. Myofibroblasts also secrete matrix metalloproteinases (MMP) 6 and 9 which aid in the restructuring of the ECM and could potentially explain the increased presence of MFs in a

metastasized tumor (Yeung *et al.*, 2013). Myofibroblasts play important roles both in normal steady-state as well as during inflammation and wound repair. Thus, carcinomas consisting of aberrantly expressed genes that lead to a complete change in expression profile containing multiple mutations adds an extra layer of complexity. The interaction between the stromal cell types and the cancer cells thus forms an extremely complex network that will require further research and analysis.

1.2.6 Myofibroblasts in Adenocarcinomas and Primary Tumor

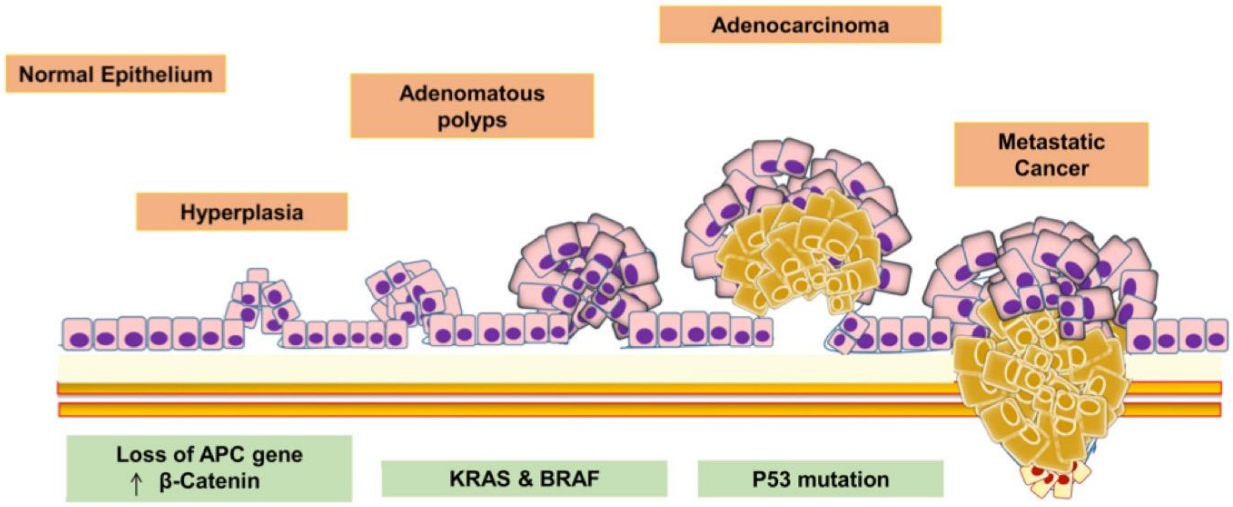


Figure 1.6 Schematic of CRC Cancer Progression: Schematic showing the progression of CRC cancer as it accumulates mutation (Figure from http://cancer-research-frontiers.org/2016-2-1/)

The most prevalent initial mutation that induces colorectal adenocarcinoma occurs within the WNT pathway, usually an inactivating mutation of the APC gene. Specifically, in the case of inherited loss-of-function (LoF) mutations of APC which leads to familial adenomatous polyposis (FAP), a syndrome where the colon exhibits hundreds, even thousands, of colonic polyps by adolescence with a very high chance of later giving rise to cancer (Bodmer *et al.*, 1988; Killeen and Killeen, 2004). Thus, nearly, if not all, patients carrying such a mutation subsequently develop

colorectal cancer. However, 65-80% of sporadic colorectal cancers acquire APC mutations and loss of function of the APC protein (Powell *et al.*, 1992; Kinzler and Vogelstein, 1996). Beyond the APC mutation, colorectal adenocarcinomas that slowly outgrow to become malignant colorectal cancers and often metastasis usually exhibit inactivating mutation of the P53 pathway as well as activating mutations in KRAS and BRAF (Palacio-Rúa *et al.*, 2014). Additionally, 15% of the sporadic colorectal tumors are found to exhibit mismatch repair deficiency, which is a form of genetic instability where the mismatch repair pathway is unable to repair DNA replication errors throughout the genome (Poulogiannis, Frayling and Arends, 2010). It is shown that mismatch repair-deficient colorectal cancer cell lines have an increased mutation rate especially in regions of repetitive DNA sequences in the 3' region of the gene (Wilding *et al.*, 2010). The combination of events along with the accumulation of mutations ultimately leads to the outgrowth of the tumor.

It has been reported that the stroma of adenocarcinomas and more advanced carcinomas accumulate myofibroblasts that surround the tumor epithelial cells (Adegboyega *et al.*, 2002). Moreover, colorectal tumors with excess myofibroblasts have a worse prognosis compared to those without. An increase in αSMA expression, a myofibroblast marker, in the stroma of a primary colorectal tumor is a strong predictor of radioresistance as well as recurrence after treatment (Tsujino *et al.*, 2007). As discussed above, pericryptal myofibroblasts are responsible for the secretion of a wide range of cellular signals including the WNT pathway and BMP that play an indispensable role in maintaining the homeostasis of the colonic crypt. For example, it has been shown that mice with mutations in both APC as well as SMAD4, which is downstream of the SMAD-dependent pathway of BMP, exhibit poorly-differentiated tumors with increased stromal

invasion (Takaku *et al.*, 1998; Satoh *et al.*, 2014). Furthermore, overexpression of noggin, a BMP inhibitor, results similarly in tumors that exhibit poor differentiation and aggressive outgrowth (Haramis *et al.*, 2004). Clinically, many groups have also shown a strong correlation between lymph node and liver metastasis of colorectal tumor and the number of myofibroblasts infiltrating the tumor (Yeung *et al.*, 2013; Takatsuna *et al.*, 2016). All the evidence suggests important crosstalk between myofibroblasts and tumor outgrowth/metastasis that needs to be further understood in order to better understand methods of treating stromal-infiltrated colorectal cancer tumors.

1.2.7 Myofibroblast Roles in Tumor Metastasis

The risk of lymph node and liver metastasis of colorectal cancer is estimated to be in the range of 10-15% (Filip *et al.*, 2020). Nonetheless, one challenge of identifying myofibroblasts in histopathology is the high degree of similarity of expressed genes shared between cancer cells undergoing EMT and myofibroblasts since both are expressing mesenchymal cell type biomarkers (Yeung *et al.*, 2013). That is EMT-related proteins may also be expressed in myofibroblasts. However, studies have shown identifying mesenchymal cell type biomarkers such as myofibroblasts and EMT-related markers are predictors of CRC metastasis (Cheng *et al.*, 2016).

Aside from the identification of myofibroblast in tumor metastasis, myofibroblasts have also been implicated to secrete signaling factors that influence tumor invasion and metastasis. For instance, it was thought that TGF-β secreted by colorectal cancer cells induces and promotes myofibroblasts, while IL-11 secreted by myofibroblasts is thought to promote invasiveness in cancer cells (Calon *et al.*, 2012). Furthermore, TGF-β is known to induce the upregulation of a

range of MMPs and ECM components such as collagen I and laminin-B1 (Hawinkels *et al.*, 2014; Rupp *et al.*, 2015). This thus leads to the second observation of myofibroblasts' ability to remodel the ECM that favors tumor metastasis. Moreover, the outgrowth of a tumor can create necrotic cores and fibrotic lesions due to over-expansion into the submucosal layer. Fibrosis can also result from excessive accumulation of ECM molecules, previously described, due to activated myofibroblasts. At the same time, myofibroblasts may migrate to fibrotic sites where pro-inflammatory cytokines are abundant or in response to signaling molecules such as PDGFs, IGF-I, or TGF-β secreted by cancer cells (Pena *et al.*, 2013). In fact, liver fibrosis has been known to greatly increase the chance of breast, melanoma, and colorectal cancer metastasis. Also, TGF- β has been found to stimulate the upregulation of fibroblast growth factor 1 (FGF1) in myofibroblasts under a cell culture setting. FGF1 is known to increase the invasiveness and motility of CRC lines (Henriksson *et al.*, 2011). Putting it all together, there is a clear range of influences myofibroblasts have over pro-metastatic colorectal cancer tumors. In addition, there exists crosstalk between colorectal cancer cells as well as myofibroblasts, co-stimulating to result in a much-amplified effect that can trigger tumor metastasis and proliferation. Thus, infiltration and participation of myofibroblasts is an important event in establishing a pro-metastatic environment consisted of the right mix of molecular signals and ECM remodeling.

1.3 Genome-wide Expression analysis in cell lines and tumors

With the rise of big data in the field of biology, the challenges of analyzing and handling evermore increasing data are quickly becoming a challenge of our decade (Marx, 2013). In recent years, many analytical tools have been developed specifically targeting large datasets and large meta-datasets. These including clustering, visualization, machine learning, and pathway analysis

that aims to provide the biological context in enormous datasets that are difficult to decipher. They have resulted in a fast-growing knowledge that uncovers previously unidentified patterns only observable through big data analysis. In this section, we will explore some of the analysis and pipelines developed for genome-wide expression analysis and the applications of the developed methods.

1.3.1 Analysis of continuous data

Gene expression is a quantification of a continuous measure representing the number of gene transcripts transcribed within the input sample. Modern gene expression quantification methods mainly include microarray gene expression platforms and RNA-seq. Microarray directly measures the intensity of fluorescence thus generating a continuous measure of intensity based on the number of transcribed RNA transcripts. RNA-sequencing, on the other hand, measures and counts discrete RNA transcripts. The integer counts are subsequently normalized resulting in non-integer continuous values. Both methods ultimately produce continuous values making gene expression analysis a study of continuous distributions (Slonim and Yanai, 2009). Thus, many of the statistical tests that measure the statistical significance in order to validate or reject the null hypothesis, which proposes no difference between certain characteristics of different groups, are commonly used in understanding gene expression data. One of the most commonly applied statistical tests in studying biological data is the Student's t-test. The Student's t-test quantifies the likelihood of the observations or input data following a t-distribution, which is inherently a normal distribution. The most common interpretation looks at the p-value or the significant value generated from the t-test and compares it with an assumed significance threshold, usually set at 0.01 or 0.05. In other words, the p-value measures the probability of obtaining the result under

the assumption the null hypothesis is correct. Thus, the lower the p-value, the lower the probability of observing the test result, and the more significant the result.

Extending the concept of Student's t-test further, a common analytical technique that aims to understand the most different and significant genes between two groups is known as differential expression analysis. The groups can be anywhere from control vs treated to groups segregated by particular phenotypes. Differential expression analysis (DEA) aims to discover quantitative changes in expression levels by calculating both the fold change using the mean expression level and the Student's t-test of each gene of each predefined group. The result $\log_2$(fold change) and the -log10(p-value) are commonly visualized on a scatter plot. The aim is to identify genes that are located on the top right, representing significantly upregulated genes, and the top left, representing significantly downregulated genes, in the first and second quadrant of the plot, respectively. Another area of interest is to understand the relationship or correlation between two genes. Pearson's correlation is commonly used to understand how well two genes or variables relate to one another.

With enormously large datasets that can generate matrices consisted of millions of data points, visualization of such complicated datasets becomes a major challenge. These datasets are too large to use standard histograms or plotting techniques to visualize and understand. Thus, either large datasets need to be broken down into smaller datasets based on *a priori* knowledge or applied with a dimensionality reduction algorithm to reduce the complexity of the datasets, retaining the most important features for visualization. Dimensionality reduction transforms the data from a high-dimensional space into a low-dimensional space while retaining meaningful features of the original data for further downstream analysis. One of the most commonly used

algorithms is Principal Component Analysis (PCA). PCA starts with standardizing the data, followed by computing the covariance matrix and extracting principal components consisted of lines with minimal distances from each independent data. In other words, PCA seeks to describe the data using linear combinations that capture the maximal amount of variance (Steffens, 1983). The first 2 principal components are often then plotted onto a standard scatter plot. Even though PCA is effective at reducing the dimension to multiple independent linear representations, most of the time, PCA is limited in 3 major ways: First, PCA requires standardization thus principal components are less interpretable. Second, if the first couple principal component fails to capture most of the variance, data can be difficult to interpret especially if each principal component captures a similar level of variance. Lastly, PCA relies on linear combinations thus datasets that cannot be described efficiently with linear combinations are represented poorly. More recently, with the rise of single-cell sequencing (scRNA-seq), the dimensionality of the data becomes especially large with matrices that can easily reach tens of millions of dimensions, making dimensionality reduction techniques especially important. A common technique called t-Distributed Stochastic Neighbor Embedding (t-SNE) is quite commonly used as a dimensionality reduction technique for data visualization. Different from PCA, t-SNE is a nonlinear dimensionality reduction algorithm that addresses the linearity issue of PCA and is much better at representing nonlinear datasets in a low dimensionality space. Nonetheless, two major shortcomings of t-SNE are the high computational load and the inability to capture global structure (Van Der Maaten and Hinton, 2008). Local structure refers to the relationship of each data point within a cluster and global structure refers to the relationship between clusters. Another, a technique called Uniform Manifold Approximation and Projection (UMAP) was

developed even more recently and is quickly overtaking t-SNE as the gold standard for dimensionality reduction as it addresses both shortcomings (McInnes, Healy and Melville, 2018).

Another important area is clustering algorithms that set out to group samples with similar characteristics into groups. Clustering algorithms are mainly separated into hard and soft clustering with the former grouping data items such that each item only belongs to 1 distinct group/cluster and the latter clustering technique allowing any data items to existing in multiple groups/clusters. K-means clustering is one of the most used hard clustering techniques with a predetermined number of cluster centroids. The goal of the algorithm is to optimize the location of the centroid such that the distances from all the data items assigned to an individual cluster or centroid are closest to the center of the centroid. Another common technique is the Hierarchical Agglomerative Clustering (HAC) that seeks to build a hierarchy of clusters using a bottom-up approach where each data item is considered its cluster initially and merges as one moves up the hierarchy (Patrik, 2005). Lastly, clustering can also be model-based meaning where the input data are expected to behave to a pre-defined model such as Gaussian or Poisson distribution. One used method is Gaussian Mixture Modeling (GMM) where the input data are assumed to be a mixture of n numbers of normal distributions (Mclachlan and Rathnayake, 2014).

1.3.2 Discrete distributions in evolution

Family history and studies on twins have strongly suggested heredity in the susceptibility of many genetic disorders (Pomerantz and Freedman, 2011). The first genes discovered associated with the inherited risk of cancer were discovered using the Mendelian approach (Fisher RA, 1918; Mayr and Provine, 1998). Modern evolutionary synthesis describes a unified theory combining Mendelian genetics with Darwinian evolution (Fisher RA, 1918). Essentially, this theory describes the evolutionary process as series of discrete changes in allele frequencies within populations that result in a range of phenotypes. The prevalence of alleles that confer greater advantages increases as alleles that confer greater disadvantages decreases due to the increase and decrease in the fitness of the host, respectively. In cancer evolution, a similar process takes place where individual cells gaining mutations or epigenetic changes that confer greater advantages survive to accumulate successive mutations and epigenetic changes over time (Nowell, 1976; Merlo *et al.*, 2006). This process of accumulating genetic or epigenetic changes is a discrete step change that is not gradual, similar to the theory of punctuated equilibrium where discrete step changes in evolution are found in fossils (Gould and Eldredge, 1993). Thus, in cancer evolution, finding these discrete step changes that substantially increase the fitness of a tumor cell can be done through observing discrete distributions of a measurable trait. Gene expression is a measurable trait of a gene that gives rise to a continuous distribution as previously described. The mutations or epigenetic changes that most likely confer a strong advantage in tumor outgrowth are expected to exhibit a gene expression distribution with a widespread across a panel of cancer cells. This is because cancer cells without the genetic or epigenetic change will exhibit a range of gene expression that clusters to form a normal

distribution, while cancer cells with different genetic or epigenetic change will exhibit a discretely different range of gene expressions that each cluster to form its own normal distribution. Thus, the expectation is that the gene expression distribution across cancer cells will exhibit a dichotomous or bimodal pattern with individual distributions representing the underlying change in the gene. By identifying these bimodal patterns (Ertel, 2010), it may be possible to explore and infer driver mutations and epigenetic drivers that plays a significant role in cancer evolution, making it an important study that may result in novel targets for cancer therapy. In this thesis, we will introduce a novel method using Gaussian Mixture Modeling (GMM) to identify distinct groups in gene expression analysis.

1.4 Single-cell sequencing analysis

Single-cell RNA-sequencing (scRNA-seq) technologies are a rising technique that performs sequencing not on bulk samples but individual cells. This provides an unprecedented level of granularity into understanding and studying the heterogeneity of a cell population inaccessible via traditional RNA-seq or microarray analysis. Nonetheless, scRNA has its shortcoming due to its high dropout rate (Qiu, 2020). In this section, we explore the current challenges in postprocessing scRNA data and some of the available pipelines established to deal with these challenges.

1.4.1 Single-cell sequencing postprocessing challenges

Currently, there are three major commercially available platforms for single-cell RNA-sequencing: Fluidigm C1, Clontech iCell8, and 10X Genomics Chromium (Wang *et al.*, 2019). These platforms are different in their methods as well as their sensitivity, specificity, and capture rate. The Fluidigm C1 platform uses an integrated microfluidic chip that enables a single cell to pass

through, capture it, and perform sequencing on the captured cell (Wang *et al.*, 2019). The microfluidic chip is a droplet-based approach that compartmentalizes droplets containing the cell, lysis buffer, and microbeads covered with barcodes for sequencing. The droplets are partitioned based on encompassing each droplet in a body of hydrophobic fluid such as oil surrounded by water to establish a high-throughput system. Similarly, the 10X Genomics Chromium platform is also a droplet-based system that uses their 'Gel-bead in Emulsion' (GEM) technology where cells, reverse transcriptase reagents, as well as barcode-embedded gel beads, are all together immersed in a droplet isolated from its surrounding via an outer layer of partitioning oil. The droplet passes through the chromium system where sequencing occurs. On the other hand, Clontech's iCell8 system is an alloy water with 5,184 nanowells with oligonucleotides containing poly d(T), unique molecular identifies (UMIs) and unique barcodes preprinted into each well. Single cells are dispensed into the wells according to a Poisson distribution in which on average 1800 single cells are captured (Wang *et al.*, 2019).

In each of the three systems, the generated scRNA-seq data is comparatively noisier and more complex to post-process compared to traditional RNA-seq data. It has been suggested that 82.2% of stochastic gene expression patterns are attributed to technical noise (Kim *et al.*, 2015). Currently, scRNA-seq relies on a one-step reaction of cell lysing, cDNA library preparation via reverse transcription of the poly-A fraction of RNAs, and PCR amplification of the cDNA library based on minimal starting material of just a single cell. In combination, each of the steps compounds to amount to a significant increase in technical noise levels compared to traditional RNA-seq on bulk samples. Thus, some of the most common challenges in postprocessing scRNA-

seq data include removing doublets, empty droplets, RNA soup as well as addressing the high dropout rate observed in scRNA-seq (Young and Behjati, 2018).

Doublets are artificial libraries generated from two cells that failed to separate during compartmentalization (Stoeckius *et al.*, 2018). They are especially prone to droplet-based approaches involving thousands of cells. Thus, by having double the library size in comparison to single cells, they can incorrectly influence downstream analysis and even suggest cell types that do not exist. Nonetheless, doublets are removed computationally by inferring that cells with a total library size much larger than the average library size of the scRNA-data are doublets. This is commonly achieved by statistically identify outliers of the normal distribution expected to consist of single cells. On the other hand, empty droplets are inferred by identifying barcodes (i.e. cells) with a total library size commonly thought to be less than 500 total counts. One of the core assumptions of droplet-based scRNA-seq is that each droplet contains the mRNA of a single cell. Nonetheless, it has been shown that cell-free mRNA can contaminate and mix within a droplet from the input solution. This non-endogenous mRNA is thought to originate from apoptotic cells that release RNA into the input solution. This concept of background RNA contamination is called the 'RNA soup'. SoupX is an algorithm developed to address RNA soup by estimating soup expression from empty droplets and use the quantified expression profile to deconvolute and remove RNA soup contamination. An even bigger challenge is addressing the high dropout rate of scRNA-seq. It is estimated that the dropout rate of scRNA-seq can range from 89-93% depending on the sequencing platform. Currently, methods such as Markov Affinity-based Graph Imputation of Cells (*MAGIC)* and Recovery of Single-Cell Under-detected Expression (*RESCUE)* have been developed to deal with the high dropout in scRNA-seq. Postprocessing scRNA is an

active field of research with the goal of extracting the maximum amount of meaningful data from noisy and complexly large scRNA-seq data (Dijk van *et al.*, 2017; Tracy, Yuan and Dries, 2019a).

1.4.2 Single-cell sequencing analysis in cell line data

Most of the single-cell sequencing is performed on tissue samples or biopsies consisting of diverse cell types. With dimensionality reduction techniques such as UMAP and t-SNE, the returned scRNA-seq output can easily be visualized with distinct clusters due to the difference in the transcriptome profile across different inter-cell types. Performing scRNA-seq on intra-cell type such as on cell lines of the same origin makes postprocessing of scRNA-seq data much more challenging due to lesser differences among the sub-cell types within a cell line(s). A time-series study on the response to glucocorticoids in breast cancer cells applied scRNA-seq to understand the changes in transcriptome profile in response to the treatment (Hoffman *et al.*, 2020). The group was able to identify slight shifts in transcriptome profiles based on treatment duration. Another group attempted to understand the heterogeneity in sub-cell type distribution within the breast cancer cell line SUM149, a triple-negative breast cancer cell line (Wu *et al.*, 2020). Intra-cell type granularity provides us a view into the heterogeneity within a single cell type such as the stages of individual cells including cell cycles and differentiation. This makes postprocessing scRNA-seq on intra-cell type an important area that needs further exploration. In this thesis, we introduce a novel pipeline answering some of these problems and comparing it with existing approaches using single-cell data on a mixture of two of our well-characterized CRC cell lines.

Chapter 2: Methods and Materials

2.1 Cell culture methods

2.1.1 Cell lines

All colorectal cancer (CRC) cell lines and myofibroblast (MF) cell lines were acquired from cryogenic storage of the Cancer and Immunogenetics Laboratory (CIL) at the Weatherall Institute of Molecular Medicine (WIMM). C10, C75, C80, C99, C106 were cell lines established in CIL while other cell lines were established and obtained from major repositories such as ATCC (American Type Culture Collection), DSMZ (Deutsche Sammlung von Mikroorganismen and Zellkulturen GmbH – German Collection of Microorganisms and Cell Cultures) and ECACC (European Collection of Cell Cultures). The full list was shown in Table 1.1.

Colorectal Cancer Cell Lines

Cell Line Name	Source
C10	Cancer Immunogenetics Laboratory (W. Bodmer)
C32	Cancer Immunogenetics Laboratory (W. Bodmer)
CACO2	ATCC
CAR1	Japan Health Foundation, Tokyo, Japan
CC20	A.R. Kinsella
CCK81	Japan Health Foundation, Tokyo, Japan
CCO7	JT Immunotech, USA
CL11	DSMZ, Germany
COLO205	ATCC
COLO206	Bridget Hill, ICRF, London
COLO320HSR	ATCC
CX1	DSMZ
DLD1	ATCC
GP2D	ECACC
HCA7	ECACC
HCT116	ATCC
HDC57	M.Schwab, DKFZ, Germany
HDC8	M.Schwab, DKFZ, Germany
HDC9	M.Schwab, DKFZ, Germany
HT29	ATCC
HT55	ECACC

LOVO	ATCC
LS174T	B.H.Tom, NW Uni Med Centre, Chicago
LS180	B.H.Tom, NW Uni Med Centre, Chicago
LS411	L.Suardet, ISCRE, Switzerland
NCIH716	ATCC
OUMS23	Health Science Research Resource Bank, Japan
OXCO1	Cerundolo, WIMM, Oxford
OXCO2	Cerundolo, WIMM, Oxford
PMFK014	RIKEN Cell Bank, Japan
RKO	ATCC
RW7213	John Mariadason, Olivia Newton-John Cancer Research Institute
SKCO1	ATCC
SNUC2B	ATCC
SW1116	ATCC
SW1222	M.Herlyn, Wistar Institute
SW1417	ATCC
SW48	ATCC
SW480	ATCC
SW620	ATCC
T84	ATCC
VACO429	Ireland Cancer Center, LWKU, Cleaveland, Ohio
VACO4A	J.McBain, VA Res.Ser. Madison, Wisconsin

Myofibroblast Cell Lines

Cell Line Name	**Source**
CCD-18Co	ATCC
Myo0164C	Cancer Immunogenetics Laboratory (W. Bodmer)
Myo1998	Cancer Immunogenetics Laboratory (W. Bodmer)

Table 2.1 Summary of All CRC and MF Cell Lines: Table listing all colorectal cancer and myofibroblast cell lines along with their source.

2.1.2 Cell culture

The cell lines were maintained in 75cm^2 sterile Falcon tissue culture flasks (Corning® Flasks) at 37 °C in a humidified incubator in 5% or 10% CO_2. All medium was supplemented with 10%. fetal bovine serum (FBS) and antibiotics PenStrep consisting of penicillin 200 IU/ml and

streptomycin 200ug/ml, Gibco/Invitrogen. Cells were also grown in 96-well, 24-well, 12-well, and 6-well tissue culture-treated plates (Corning® Tissue culture-treated plates).

In our panel of CRC cell lines, we include both adherent and suspension cell lines where the passage method was different. In an adherent cell line, the medium was changed every 2 days. When cells reach 70-80% confluency, cells were passaged by removing the medium and washed with 5-10ml of Absolute Phosphate Buffered Saline (PBSA). Cells were then treated with 1X Trypsin/Ethylenediaminetetraacetic acid (EDTA) to detach the cells from the flask. The cells were incubated in Trypsin/EDTA for 5 minutes before adding an equal amount of full medium (+FBS and PenStrep) to neutralize and stop the reaction of Trypsin/EDTA. The cells were collected and pelleted by centrifuging at 1500 RPM for 5 minutes. The Supernatant was removed, and the pellet was resuspended in 5-10 ml of the fully-supplemented medium before re-seeding into the flask.

In a suspension cell line, the medium was changed every 2 days by centrifuging the medium containing the cells at 1500 RPM for 5 minutes. The supernatant was removed, and the pellet was resuspended in a fresh medium. When cells reach 70-80% confluency, cells were passaged 1:10 by removing 90% of the medium and resuspending the result with a fresh medium, effectively diluting the original cell concentration to 1:10.

2.1.3 Cell Line Storage

Cell lines were routinely stored in liquid nitrogen after the first passage and at the end of the experiment. For cryopreservation, we use 10% DMSO in FBS in sterile cryogenic vials. When the cells reach 70-80% confluency, the cells go through the same process described in the

previous section during cell culture passage. Except when a pellet was formed after centrifuging at 1500 RPM for 5 minutes, the pellet was resuspended in 500-600 µL of freezing medium (10% DMSO + 90% FBS) and undergo slow-freezing on dry ice for 1 hour before being stored in the liquid nitrogen. Alternatively, vials can also be slow frozen in Mr. Frosty Freezing Container (Thermo Scientific, Ref 5100-0001). It was left in -80 ˚C freezer overnight to ensure slow-freeze prior to transfer to liquid nitrogen.

2.1.4 Cell counting

Cell counting was done using the Cellometer Auto T4 machine (Nexcelcom Bioscience). 20 uL of cell suspension was aliquoted into a cell counting chamber. The chamber was inserted into the cell counter machine. The user selects the approximated size of the cells prior to counting. The software uses contrast in the brightfield of eight individual areas located on the chamber to calculate cell concentration. The returned output was recorded.

2.2 DNA methods

2.2.1 DNA extraction from cultured cells

DNA extraction on cell culture cell lines was done by using the DNeasy® Kit (Qiagen, UK) according to the manufacturer's protocol.

Approximately $5x10^6$ cells were centrifuged to form a pellet at 300xg for 5 minutes and disrupted by incubating them with proteinase K for 10 minutes and the provided lysis buffer in the DNeasy® Kit. 200 ul of absolute ethanol to precipitate DNA. Following the protocol, the mixture was then passed through a provided DNA binding column. It was then washed with

washing buffers and eluted by incubating in the Elution Buffer (EB buffer) for 5 minutes before centrifugation at max speed for 1 minute stored at 4*°C.*

2.3 RNA methods

2.3.1 RNA extraction

RNA extraction was done by using the RNeasy® Mini Kit (Qiagen, UK) according to the manufacturer's protocol. Roughly $5x10^6$ cells were centrifuged to form pellets. The pellet was then mixed with the provided lysis buffer. The entire process was dealt with using an RNase-free syringe. The lysate was then passed through the RNA binding column. The column was washed with RNA wash buffer before eluting with RNase-free water.

2.3.2 cDNA library preparation

Purified RNA was transcribed to cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, UK) according to the manufacturer's protocol. The concentration of RNA samples was assessed with the NanoDrop ND 1000 Spectrophotometer (Thermo Scientific) and was subsequently diluted to 10 ng/uL in nuclease-free water. The 10 uL of diluted RNA sample was mixed with 2 uL of 10X RT buffer, 0.8 uL of 25X dNTP mix, 2 uL of 10X RT Random Primers, 1 uL of 20X MultiScribe Reverse Transcriptase, 1 ul of 20X RNAse inhibitor, and 3.2 uL of nuclease-free water to reach a final volume of 20 uL. The final mixture was incubated in the G-Storm Thermal Cycler for 10 minutes at 25*°C,* 120 minutes at 37*°C,* and 85*°C* for 5 minutes. The resulting cDNA library was used for qRT-PCR.

2.4 Quantitative real-time Polymerase Chain Reaction (qRT-PCR)

TaqMan® Gene Expression Assay 20X FAM dye-labeled kit, 2X TaqMan® Universal PCR Master Mix with the target cDNA were used according to the manufacturer's protocol for mRNA quantification. It was always done with at least a duplicate or a triplicate.

First, dilute the cDNA samples from reverse transcription to 25 ng/uL. Second, prepare the reaction master mix consisting of mixing 1 uL of 20X TaqMan Gene Expression Assay, 10 uL of 2X TaqMan Fast Universal PCR Master Mix (Applied Biosystems, Ref. 4352042), and 5 uL of nuclease-free water to reach a final volume of 16 uL. Mix 4 uL of diluted cDNA into 16 uL of master mix for a final volume of 20 uL per well of MicroAmp Fast Optical 96-well Reaction Plate (Applied Biosystems, Ref. 4346906). Seal the plate with MicroAmp Optical Adhesive Film (Applied Biosystems, Ref. 4311971). Run qRT-PCR on a 7500 Fast Real-Time PCR System (Applied Biosystems, Ref. 4351107). All samples were paired with the housekeeping gene Ubiquitin C (UBC) to normalize technical differences in cycling time across different samples. We performed qPCR on the Applied Biosystems 7500 Fast Real-Time PCR System Relative Quantification Machine for a total of 40 cycles with the program shown in the following (Table 1.2):

Cycling step		Temperature (°C)	Time	Cycles
Activation Stage		50	2 min	1
Polymerase Activation Stage		95	10 min	1
PCR cycles	Melt	95	15 sec	40
	Anneal/Extend	60	1min	
Storage		4	Indefinite	

Table 2.2 Cycling Times for qRT-PCR: Cycling times for qRT-PCR using TaqMan probes on the Applied Biosystem 7500 Fast Real-Time PCR System Relative Quantification Machine.

The output of the qRT-PCR was assessed by determining the converted ΔCt value. In summary, standard comparative ΔΔCt values were compared to a reference sample such as a housekeeping gene. In our case, UBC was our internal control. Expression of genes of interest was then quantified as a fold change compared to this reference by subtracting the ΔCt value of UBC from the ΔCt of the gene. ΔΔCt value was determined by calculating $\Delta\Delta\mathrm{Ct} = 2^{(n-\Delta\mathrm{Ct})}$, n was the largest integer that was above the largest ΔCt within all the samples of the same gene. We list all of the TaqMan probes used in this thesis in Table 1.3.

Target gene	**Reference**	**Source**
UBC	Hs00824723_m1	Applied Biosystems
ZEB1	Hs01566407_m1	Applied Biosystems
VIM	Hs00958111_m1	Applied Biosystems
CDH1	Hs01023894_m1	Applied Biosystems

Table 2.3 Summary of TaqMan Gene Expression Probes: Summary of TaqMan Gene Expression Probes used in this thesis along with their references and source.

2.5 Polymerase Chain Reaction (PCR)

The extracted DNA was amplified with REDTaq® DNA Polymerase according to the manufacturer's protocol using primers: 3' downstream F and 3' downstream R. The final mix includes 1.25 uL of 10 uM of primers each, 0.75 uL or 3% DMSO, 12.5 uL of DNA polymerase and 3 uL of DNA diluted at 15 ng/uL. PCR was performed on a GSTORM GS4 PCR platform for a total of 35 cycles with the program shown in the following (Table 1.4):

Cycling step		Temperature (°C)	Time	Cycles
Initial denaturation		98	1 min	1
PCR cycles	Denaturation	98	30 sec	35
	Annealing	56	1 min	
	Elongation	72	1min	
Final extension		72	10min	1

Storage	4	Indefinite

Table 2.4 Cycling Times for PCR: Cycling times for PCR performed on a GSTORM GS4 PCR platform.

2.6 DNA and RNA quantification

The concentrations and the quality of the RNA or DNA samples were assessed with NanoDrop ND-1000 Spectrometer readings at OD260 and OD280. Generally, we expect a ratio of ~1.8 for 260/280 to be accepted as 'pure' DNA.

2.7 Gel Electrophoresis of PCR results

All PCRs were analyzed on 2 - 4% Agarose gels depending on the expected size of the band (2g - 4g agarose and 100ml of 1XTBE), stained with ethidium bromide. The concentration of agarose depends on the size of bands needed to be separated. A higher percentage will help resolve smaller bands from each other. The gel was loaded with DNA samples mixed with 6x loading buffer as well as size ladder markers. The gel was run at 100 Volts for 30 minutes in 1X Tris/Borate/EDTA (TBE) buffer. The resulting gel was visualized under UV light (UV Transilluminator, BioRad).

2.8 siRNA Transient Transfection

For reverse transfection of short interfering RNAs (siRNAs), they were delivered by RNAiMAX lipofectamine to cell lines. Pre-designed siRNAs were purchased from GE Healthcare Dharmacon (USA). 6 µl of 10 µM siRNA and 6 µl of lipofectamine were each mixed in 100 µl of Opti-MEM transfection medium. Mix both in a 1:1 ratio and incubate at room temperature for 7 minutes. Add 200 µl of the mixed siRNA/lipofection/Opti-MEM solution to each well of a 6 well plate. Then, add 2 ml of antibiotic-free medium with 10% FBS containing 300,000 cells to each

well. Evenly distribute the cells in plate. Change the medium back to complete medium (medium+ 10%FBS + 1%PenStrep) after 24hrs of incubation.

Target	Catalog #	Assay ID	Source
ZEB1 siRNA	AM16708	109651	Thermo Fisher
AKT3 siRNA	AM51331	110901	Thermo Fisher

Table 2.5 Summary of siRNA: Summary of siRNA used in this thesis along with the catalog number, assay ID, and source.

2.9 Myofibroblast Recruitment Assay

Coculture assays were done using Matrigel diluted 1:1 with serum-free DMEM along with 1% PenStrep. CRC cell lines were prepared for coculture assay by detaching Trypsin/EDTA and washed 3 times with PBS by repeat resuspension, pelleting, and removing PBS. The CRC cell line cells were then resuspended in serum-free DMEM supplemented with 1% PenStrep. Cells were then transferred into another tube that was then centrifuged to form a pellet at 2,200 RPM for 15 minutes. The supernatant was removed, and the pellet was resuspended in 50% Matrigel and 50% DMEM before seeding. All this time, this must be kept on ice to prevent the Matrigel from solidifying. To establish the recruitment assay, 3 uL of 50% Matrigel containing 10,000 CRC cells were deposited in the middle of the well as a droplet in a 96-well plate. Let the gel harden for 1 hour at room temperature prior to covering it with 200 uL of 1% serum-supplemented DMEM containing 2,000 myofibroblasts. Make sure to add it carefully to prevent it from breaking the Matrigel droplet containing CRC cell lines. Remove and replace the medium with 1% serum-supplemented DMEM every 2 days. After 9 days, cultures were fixed and stained for downstream analysis.

2.10 Protein Methods

2.10.1 Whole-cell lysate extraction

Cells were grown to 70-80% confluency before being harvested for protein extraction. Cells were washed with PBS twice on the plate. Cells can either be collected into a tube or directly lysed on the plate with RIPA buffer (150mM NaCl, 1% NP40, 0.5% NaDeoxycholate, 0.1% SDS and 50mM Tris.HCl pH7.5 – PIERCE) with protease inhibitor tablet (Complete Mini Tables, Roche Diagnostics). After homogenizing the samples by pipetting up and down the plate, the lysis-buffer solution was collected into a tube and incubated on ice for 30 minutes. The tube was vortexed every 5-10 minutes. Then, we centrifuge the solution at 14,000 rpm and 4 ˚C for 20 minutes. The supernatant was extracted and represents the protein lysate. It was stored at -20 ˚C.

2.10.2 Bradford Protein Assay (BCA) protein quantification

The BCA™ protein assay kit (Pierce) was used for protein quantification according to the manufacturer's protocol. We premade the 13 protein standards using serial dilution of BSA ranging with a known concentration ranging from 0 to 2000 ug/ml. 25 uL of each sample and each standard was added into individual wells of a 96-well plate (Falcon, BD Biosciences). 200 uL of BCA reagent consisted of a 1:1 mix of Reagent A and Reagent B included in the BCA kit. Add 200 uL of BCA reagent to each well loaded with samples or protein standard. The samples and protein standards must be preloaded before adding BCA reagent to eliminate the need to change tips for each aliquot thus eliminating the differences in reaction time. Use aluminum foil to enclose the plate. Incubate at 37 ˚C for 30 minutes in the incubator. The plate was then examined on a plate reader (uQuant, Bio-Tek instrument) and the protein concentration was recorded by

measuring the absorbance at 562 nm. The value was then calculated by extrapolating a linear regression of the standard curve using the results of the protein standards.

2.10.3 Western blot (WB)

Protein extracted with lysis buffer under reducing conditions was denatured at 95 ˚C for 5 minutes. Denatured protein was separated in a polyacrylamide gel with a 4% upper gel and resolved in an 8 or 10% separating gel at 110V for approximately 2 hours in 1X SDS running buffer. Proteins were transferred onto a Polyvinylidene difluoride (PVDF) blotting membrane activated with 100% methanol using the semi-dry transfer system for 1 hour at 20V according to the manufacturer's instructions (GE Healthcare Life Science). A successful transfer can be assessed with Ponceau red stain followed by a washing step with Tris-buffered saline with Tween20 (TBS-T) for 10 minutes to remove the stain. The membrane was blocked in 5% skimmed milk in Tris-buffered saline with 0.1% Tween 20 Detergent (TBST) buffer or one hour. The membrane was incubated in a monoclonal antibody against the protein of interest and 5% skimmed milk in TBST overnight at 4 ˚C. The membrane was washed 3 times 10 minutes each with TBS-T before incubating in secondary antibody conjugated with Horseradish peroxidase (HRP) at RT for one hour. The membrane was washed again 3 times 10 minutes each before developing the membrane with Clarity™ Western ECL Detection Reagent (Bio-Rad) according to the manufacturer's instructions.

Antibody	Source	Concentration
EpCAM AUA1 mAb	In-house	1:100
Vimentin 3B4 mAb	Dako, Ref. M7020	1:100
E-Cadherin 24E10 mAb	Cell Signaling Technology, Ref. 3195S	1:200
B-tubulin mouse clone TUB 2.1	Sigma-Aldrich, T4026	1:100

Table 2.6 Summary of Antibodies for WB: Summary of all the antibodies used in western blot along with the source and concentration used.

2.10.4 Membrane stripping

Membranes can be re-probed with different antibodies repeatedly after membrane stripping during western blot. To strip the bound antibody off of the PVDF membrane, we performed membrane stripping with mild stripping buffer consisted of 15 g of glycine, 1 g of SDS, 10 mL of Tween 20, and topped up with 800 ml of ultrapure water. The membrane was incubated in the solution for 20-30 minutes on a belly dancer orbital shaker. The membrane was then washed twice with PBS and twice with PBS + 0.1% Tween 20 Detergent (PBST) each for 5-10 minutes.

2.11 Fixation and Immunofluorescence (IF)

Fixation and methods of IF can be very different depending on the sample that was being processed. Here we describe two scenarios that were applicable for this thesis: 2D cell culture and Matrigel culture.

In a 2D cell culture, the medium was carefully removed from the cell culture and the cells were washed twice with PBS. Cells were fixed with 4% Paraformaldehyde (PFA) (VWR Chemicals, Ref. FOR0020AF59001) for 10 minutes. Wash cells with PBS twice and permeabilize the cells with 0.1% Triton X100 (Sigma-Aldrich, CAS No. 9002-93-1) PBS for 15 minutes. Wash cells with PBS

twice. Block cells with 2% BSA blocking solution to block nonspecific binding for 1 hour at room temperature. Remove medium and add 2% BSA blocking solution with primary antibody at the appropriate concentration. Incubate for an hour at room temperature or 4 ˚C overnight. Wash cells three times with PBS for 5 minutes each. Incubate cells in 2% BSA blocking solution plus secondary antibody at the appropriate concentration (1:1000) for 30 – 60 minutes in room temperature. Wash cells three times with PBS for 5 minutes each and incubate cells in PBS + DAPI at 1:10000 dilution for 10 minutes at room temperature. Wash cells twice and leave cells in PBS for imaging.

In a Matrigel culture, the medium was carefully removed from the Matrigel culture without disturbing the gels. Carefully wash well once with PBS while avoiding the Matrigel blob. Next, fix the wells with 4% PFA for 20 minutes at room temperature. Remove the PFA and carefully wash twice with PBS and incubated with 0.2% Triton X100 PBS permeabilization solution for 5 minutes. The Matrigel blob should be intact throughout the protocol. Carefully wash cells four times with PBS and block with 2% BSA in PBS solution and left overnight at 4 ˚C. Wash wells 3 times with PBS, 5 minutes each. Incubate cells in primary antibody in 2% BSA PBS blocking solution at the appropriate concentration. Leave it overnight at 4 ˚C. Wash wells 3 times with PBS, 5 minutes each. Incubate cells in secondary antibody in 2% BSA PBS blocking solution at the appropriate concentration over night at 4 ˚C. Wash wells 3 times with PBS, 5 minutes each. Incubate cells in DAPI solution at 1:10,000 before imaging.

Antibody	Reference	Concentration
Laminin pAb	Abcam, Ref. ab11575	1:800
EpCAM AUA1 mAb	In-house	1:100
Vimentin 3B4 mAb	Dako, Ref. M7020	1:100
E-Cadherin 24E10 mAb	Cell Signaling Technology, Ref. 3195S	1:200
ZEB1 mAb	ATLAS, AMAb 90510	1:500

Table 2.7 Summary of Antibodies for IF: Summary of all the antibodies used in IF along with the source and concentration used.

2.12 Immunofluorescence imaging

IF imaging was done using a Zeiss AxioObserver Z1 microscope controlled by the ZEN digital imaging software (Carl Zeiss AG, Germany) installed on the connected computer. We can preset the wavelength to capture multi-color imaging which allows us to visualize staining with up to 3 colors simultaneously along with DAPI nuclear staining.

2.13 Image Analysis

Images were exported from the ZEN software for further processing using ZEN software. The software has tools that enable users to remove background spots by eroding and dilating. Refer to this documentation for more details: https://www.gu.se/sites/default/files/2021-03/Beginners_Guide_to_Image_Analysis_in_ZEN_Blue_67815830.pdf. The post-processed images were further analyzed and interpreted.

2.14 CRISPR-Cas9 stable cell line

2.14.1 Design CRISPR guides

Browse the USCS Genome browser to view all isoforms of your gene of interest. Cross-reference each isoform to choose the correct critical exon (C.E.) that was shared among all isoforms. Make sure your C.E. was not the first or last exon, also the length of the exon cannot

be divisible by 3 to avoid a one amino acid frameshift. Choose two cut sites for the chosen C.E. for the CRISPR-Cas9 to recognize, one being 200 bp upstream and being 200 bp downstream. Enter both sequences into MIT CRISPR Zhang lab and Harvard Chop Chop to compare the computed results. The designed guides were ordered as oligonucleotides through Thermo Fisher.

2.14.2 Prepare sgRNA expression construct

To prepare the sgRNA, start by resuspending and diluting primers to a final concentration of 100 uM in ddH_2O. Store them at -20 ˚C. The designed upstream and downstream CRISPR guides were inserted into pSpCas9 (BB)-2A-GFP PX458 and pSpCas9 (BB)-2A-Ruby PX458 plasmids, respectively, using BbsI (5' … GAAGAC $(N)_2$ … 3') restriction enzymes. Generate the sgRNA expression construct by inserting the gDNA into pSpCas9 plasmids. Start by annealing the forward and reverse primer to form a full gDNA. Add 1 uL of forward sgDNA (100 uM), 1 uL of reverse sgDNA (100 uM), 1 uL of T4 10X ligation buffer (New England Biolabs), 1 uL of T4 Polynucleotide Kinase (PNK) (New England Biolabs), and 6 uL of ddH_2O to reach a final volume of 10 uL. Set up PCR cycle on thermocycler with the following parameters (Table 1.8):

Cycling step	Temperature (°C)	Time	Cycles
Annealing Step	37	30 min	1
Annealing Step	95	5 min	1
Annealing Step	25	5 min	1
Storage	4	Indefinite	

Table 2.8 Cycling Times for Annealing sgDNA

Dilute the annealed gDNA 1:200 by adding 1 uL of the annealed oligo to 199 uL of ddH_2O. Measure concentration of gDNA with NanoDrop. Next, ligate this annealed gDNA into the plasmid by adding 100 ng of pSpCas9 plasmid, 2 uL of diluted oligo from the previous step, 2 uL of Tango

buffer 10X (New England Biolabs), 1 uL of 10 mM dithiothreitol (DTT), 1 uL of 10 mM adenosine triphosphate (ATP), 1 uL of FastDigest BbsI, 0.5 uL of T7 ligase and top it up with ddH_2O to reach a final volume of 20 uL. Incubate the ligation reaction mix in a thermocycler with the following parameters (Table 1.9):

Cycling step		Temperature (°C)	Time	Cycles
	Annealing Step 1	37	5 min	
Ligation Cycle	Annealing Step 2	21	5 min	6
Storage		4	Indefinite	

Table 2.9 Cycling Times for Ligating and Inserting sgDNA into Plasmid: Cycling times for ligating and inserting the sgDNA into PX458 Crispr plasmid construct.

2.14.3 Bacterial Transformation

On Day 1, thaw 20-50 ul DH5alpha competent cells on ice and add 0.5 ul DNA (plasmid) to pre-chilled 1 mL tube. Add cells to the tube and incubate on ice for about 20 minutes. Heat shock at 42 °C in a water bath for 1 min and then incubate on ice for 2 min. Add 290 ul SOC medium and incubate at 37 °C, 200 -290 rpm for 1 hour. Plate 50, 100, and 150 ul on LB agar containing chloramphenicol (or your antibiotic) overnight at 37 °C.

On Day 2, add antibiotics into 4mL antibiotic LB broth (antibiotic dilution: 1:1000). Pick 1-2 colonies from chloramphenicol plates, use pipette tips to pick the colonies, and directly shoot the tips into the medium. Incubate for 14-16 hours at 37 °C, 290 rpm in a shaker. The next day, we performed MiniPrep/MaxiPrep to extract the plasmid. Lastly, use NanoDrop to quantify the concentration of the extracted plasmid.

2.14.4 Validating the Insertion into plasmid

Individual px458 plasmid with guides inserted was sequenced at an in-house facility to confirm the insertion of designed guides.

Name	Forward primer sequence (F)/Reverse primer sequence (R)	Product
CRISPRKO_seq	5'-GCCTGGTATCTTTATAGTCC-3'	Insertion of designed guides into Plasmid

Table 2.10 Primer Design for CRISPR Guide Sequencing: Primer used for sequencing the plasmid to figure out whether the guides were inserted correctly into the insertion site.

2.14.5 Plasmid transfection

Plasmid transfection into the cell line was done as described in 3.7. Transient transfection but instead of siRNA, replace that with the CRISPR plasmid.

2.14.6 Fluorescence-activated cell sorting (FACS)

FACS was used to double select GFP+ Ruby+ positive single cells for cloning CRISPR-KOs. Cells were stained with DAPI before FACS to isolate live cells.

2.14.7 Cloning

Single cells were FACS sorted into a 96 well tissue-culture (TC) treated plate and left at the back of the incubator to prevent any disturbances at 37 ˚C for a week. Cells from growing colonies in single wells were monitored and each such colony was then expanded into a single well of a 24 well plate to reach 80% confluency. Each of the original colonies from individual wells of the 96 well plate was now plated into individual wells of a 6 well plate. After cells reach 80% confluency, split half for cryopreservation and half for subsequent experiments.

2.14.8 Primer Design

The region of interest in the extracted DNA was amplified with REDTaq® DNA Polymerase according to the manufacturer's protocol. The process of PCR was described in 3.4. PCR Polymerase Chain Reaction.

Name	Forward primer sequence (F)/Reverse primer sequence (R)	Product	Product size (bp)
5' upstream_F	5'-GCCGTTAAAGGCTGGGTAGA-3'	5' upstream CRISPR target region	971
5' upstream_R	5'-TTACACCCAGACTGCGTCAC-3'	5' upstream CRISPR target region	971
3' downstream_F	5'-GTTGAGTGAGCTTTAGACAGGT-3'	3' downstream CRISPR target region	991
3' downstream_R	5'-AGATGGTGGGTGAGCACATTT-3'	3' downstream CRISPR target region	991

Table 2.11 Summary of Primers Used for Sequencing KO Region for ZEB1

2.15 Surveyor Assay

The efficiency of genetic deletion by different guides was assessed with the Surveyor Assay Kit according to the manufacturer's protocol. The surveyor assay was an assay that recognizes incomplete complementarity between two strands of DNA to identify knockout efficiency judging by the ratio of band thickness i.e. WT band: KO band. Surveyor assay requires the extraction of DNA using blood and tissue DNAeasy kit, followed by PCR to amplify the KO site with a pre-designed primer asymmetrical to the KO site to ensure a visible distinction of band sizes. PCR was carried out with a HotStar high fidelity polymerase by Qiagen with the conditions of 10ul Qiagen Buffer, 5ul Forward Primer (10uM), 5ul Backward Primer (10uM), 100ng DNA, 1ul polymerase. After the PCR, perform gel electrophoresis to check for the band size of interest. If the guides were successful in cutting the target sequence, multiple bands should appear at

known sizes. The goal was to pick the pair of guides that cut the target sequences as well as having similar cutting efficiency.

2.16 CRISPR knockout verification

CRISPR knockout clones were individually assessed for the ablation of expression both by a genetic-KO and protein expression. Clones were screened with a high-throughput method first by performing PCR on the critical exon to identify a genetic deletion by visualizing a shorter band. Those that have a genetic deletion enter the second round of screening by western blot to ensure there was an ablation at the functional level namely no expression of the protein. Clones that have both genetic deletion and ablated protein expression were confirmed as a genetic knockout model.

2.17 Statistical methods

2.17.1 Microarray data analysis

Microarray data analysis was done through Python. Student t-test with nonequal variance corrected with Bonferroni correction and fold change across different user-defined groups were used to construct volcano plots to provide an easier visualization of potential genes of interest. For statistical comparisons, a p-value of less than 0.05 ($p<0.05$) was considered insignificant.

2.17.2 Bimodal gene analysis

We presented a detailed description of our method GMMchi in Chapter 4. In brief, genetic expression bimodality was mainly assessed through Gaussian Mixture Modeling (GMM) with iterative tail-trimming. Gaussian Mixture modeling assigns a posterior probability to each data point under the most fitted mix of Gaussian mixtures through expectation-maximization (EM).

The number of mixtures that fit the given data was decided through a Bayesian information criterion (BIC) that penalizes on a log scale for excessive overfitting.

Chapter 3: A novel Gaussian Mixture Modeling (GMM) Method on Expression Pattern Recognition

3.1 Introduction

Gene expression data is one of the most widely used tools in multiple aspects of understanding cell biology. There is a range of tools developed to analyze expression levels including the microarray and RNA-seq technologies. Moreover, gene expression data produced across all platforms inherently result in a continuous-valued dataset representing the levels of expression of individual genes. The field of computational gene expression analysis is thus an analysis of continuous distributions and the correlative understanding between genes or sets of genes within a biological context.

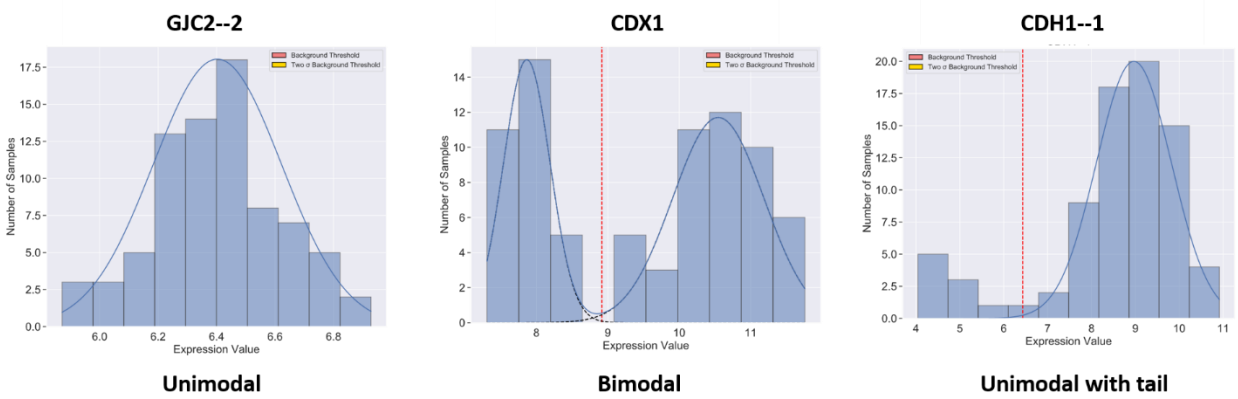
In cancer biology, tumorigenesis is driven by permanent or semi-permanent alterations on the genetic or epigenetic levels (Chatterjee, Rodger and Eccles, 2018). Many of these changes lead to subtle changes in the phenotype of a cell. However, there exists a limited number of mutations or epigenetic changes common across different types of cancers that lead to selected phenotypes that confer huge advantages such as uncontrolled proliferation, anti-apoptotic, and immune evasion characteristic of a tumor (Bailey *et al.*, 2018). These striking phenotypic changes are known as driver mutations or epigenetic drivers. The successive accumulation of these advantageous genetic or epigenetic changes will ultimately lead to tumor outgrowth.

Mutations or epigenetic changes that lead to a shift in the expression level of a gene or set of genes will naturally tend to cluster with that mutation. However, expression can be affected by more than one mutation or epigenetic change presented in the cell or through changes in the microenvironment that can act as inhibiting or activating factors on the tumor expression profile. Nevertheless, given the discrete nature of driver mutations and epigenetic drivers (Wang *et al.*, 2009), the expectation is that the distribution of the expression level of such

genes will cluster in a bimodal distribution whereby one component consists of cells with a genetic or epigenetic change and another component consisting of wildtype cells.

Thus, the closer the change in gene expression level with a given mutation or epigenetic change, the more likely a bimodal distribution differentiating wild-type and mutation will be observed. Moreover, the more striking the change is in the phenotype, the more likely it is to have a measurable effect on the biology of the cell. Naturally, the expectation is the observation of an association between a strong phenotype, a driver mutation, or an epigenetic driver, and a clear separation of the distributed gene expression across the two classes (wildtype vs mutation), ultimately leading to the strongest advantageous selection of cancer outgrowth. The interest is then to identify striking bimodal distributions that lead back to novel mutations or epigenetic drivers. Here this is approached by fitting mixtures of normal distributions, called Gaussian Mixture Models or GMM, to the expression data.

Figure 3.1 Common Distributions in Gene Expression Analysis: Examples of common distributions seen in gene expression data. GJC2, a gap junction protein, exhibits a unimodal distribution representing a single Gaussian distribution while CDX1, a homeobox protein, exhibits a bimodal distribution representing two distinct Gaussian distributions. CDH1, gene encoding E-cadherin, represents a unimodal distribution with a tail in which is one of the main challenges of GMM. Expression levels based on microarray analysis are given $\log_2$ on the x-axis, and numbers of samples with given expression levels on the y-axis. The continuous curves are fitted normal distributions and the vertical dotted red line marks the best estimate for separating low and high levels of expression.



As shown in Figure 3.1, the most common distribution patterns found in gene expression data are unimodal distributions and bimodal distributions consisted of 1 or 2 normal (Gaussian) distributions, respectively. One of the major challenges of Gaussian Mixture Modeling (GMM) is the assumption of fitting mixtures of normal components. In Figure 3.1, we have shown CDH1, the gene encoding E-cadherin, exhibiting a unimodal distribution with a 'tail', which is a non-normal distribution that lies outside of the normal component. In this chapter, we will be demonstrating a systemic approach of our novel GMM method capable of identifying normal and non-normal components across examples of any given expression data and validations of the presented method.

3.2 Results

3.2.1 Gaussian Mixture Modelling

The Gaussian Mixture Modeling (GMM) with chi-square pipeline was written and developed in Python. GMM was performed using the **sklearn.mixture** package in Python (Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas *et al.*, 2011). **sklearn** is a free Python package focused on providing pre-built analytical and statistical tools for easy access and implementation. The GMM function in the **sklearn.mixture** package is an implementation of the GMM. The documentation is found in the link:

https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html. The Gaussian Mixture Modeling is a method used to fit *n* numbers of Gaussian or normal distributions to any given data input. The expression data for each gene or probe set, which identifies a gene seen in microarray datasets, was examined as a univariate distribution across the set of samples obtained from the input data.

The matrix of the input data consists of samples across the columns and gene expression data along the row. The input data can be generalized to be sample (columns) and features (row) thus any data represented in this format can leverage this pipeline for downstream analysis.

Once the matrix is loaded onto the pipeline. One or two Gaussian distribution(s) are fitted to each gene/probe set represented in the form of a histogram. This allows unequal sub-distribution and variances using the expectation-maximization (EM) algorithm, already built into **sklearn.mixture** seen in Figure 3.1. For bimodal gene expression patterns, each sample will be labeled a 1 or 2 according to the modeled distribution it is more probabilistically likely to belong

to. This is in essence the categorization of continuous data into categories of absent, low, or high expressing sub-groups of the gene used for further analysis.

GMM is broken down into two major parts: model-fitting using EM and model selection using the Bayesian information criterion (BIC).

Modeling fitting

Briefly, during model-fitting, the EM algorithm proceeds as follows (A.P.Dempster, N.M.Laird and D.B.Rubin, 1997):

x_i = the i^{th} sample data

$\mu_{1,2}$ = the mean of the 1 or 2 distribution

$\sigma_{1,2}$ = the standard deviation of the 1 or 2 distribution

$p_{1,2}$ = the probability of x_i under the 1 or 2 distribution

1) **Initialization**: The GMM algorithm runs K-means clustering based on the number of expected Gaussian distributions to initialize $\mu_{1,2}$ and $\sigma_{1,2.}$ Then, it evaluates the log-likelihood $p_{1,2}$ with these parameters of each data point x_i.
2) **Expectation step (E-step)**: Evaluate the posterior probability using the current values of the mean ($\mu_{1,2}$) and variance ($\sigma_{1,2}^2$). In other words, how likely is each data point x_i under distribution k.
3) **Maximization step (M-step):** Estimate the new parameters using the posterior probability.
4) **Exit condition**: Iterate step 2) and 3) and stop when the lower bound average gain reaches below a convergence threshold of $1e^{-8}$ as the default of GMM in ***sklearn***.

Model selection

By fitting the data with either one or two Gaussian distributions using the EM algorithm we get 2 different fitness representing how well 1 or 2 normal components fit the data. To select the better fitting condition, model selection (1 vs 2 distributions) was based on the maximization of the Bayesian Information Criterion (BIC) defined as the following in Equation (1) (Bhat and Kumar, 2010):

$$\boldsymbol{BIC_{M,n} = -2\,log(L_{M,n}(x|\hat{\theta})) + k\,log(N)} \quad \textbf{(1)}$$

BIC can also be rewritten in terms of the χ^2 as Equation (2). The use χ^2 is to measure the goodness of fit of the estimated distribution. In other words, χ^2 is an approximation of the maximum likelihood component presented in Equation (1) of the fitted distribution. The incorporation of χ^2 in GMMchi will be discussed in depth in section 3.2.3.3 Iterative Tail Trimming.

$$\boldsymbol{BIC = \chi^2 + k\,ln(n)} \quad \textbf{(2)}$$

In BIC, $L_{M,n}(x|\hat{\theta})$ is the maximized likelihood function of model *M* with *n* components or the number of normal distributions, with maximizing parameter $\hat{\theta}$, determined as a result of the EM algorithm. *N* is the sample size and *k* is the number of estimated parameters. It is important to note that BIC is a penalty term that increases as a log-likelihood function with the number of parameters to prevent overfitting. A penalty term is a function that penalizes by increasing the returned value. In other words, the better the fitness, the lower the score, and vice versa. BIC has been shown to work well across a range of applications seen in the literature. For our use, BIC is calculated for deciding between fitting our data with one or two Gaussian components.

The model with the lower BIC value is thus selected. Essentially, BIC ensures that if we are fitting the data with 2 normal distributions, the fit must be sufficiently better than fitting it with 1 normal distribution.

3.2.2 Current Challenges of GMM on Expression Data Analysis

The challenges faced with directly applying GMM to expression data analysis and the identification of bimodal distributions lie within the nature of gene expression datasets. Although expression data like many other datasets mostly behave as Gaussian distributions after $\log_2$ transformation (Peter Kalugin, 2017), a large proportion of gene expression data exhibits a pattern where a long tail of outliers trail outside of a normal distribution where most of the data lie (Figure 3.2). We have coined this as the 'tail problem'.

Essentially, the tail problem is a pattern where a long tail of non-normally distributed outliers falls outside of 1 or 2 normal distribution(s). An example is shown in the distribution pattern of CDH1 in Figure 3.1. This non-normally distributed tail is thus inadequately dealt with under the GMM where input data is expected to be a mixture of normal distributions.

Here, we demonstrated an example of the tail problem in the returned output of GMM in Figure 3.2. In Figure 3.2a, the simulated bimodal distribution consists of two distributions with means of 5 and 20 and variances of 3 and 10, respectively. Each distribution is simulated with 1000 random samples. While in Figure 3.2c, the simulated unimodal distribution with a tail problem has a mean of 10 and a variance of 1 and a uniform distribution, which forms the tail, with a minimum and maximum of 12 and 20, respectively. After fitting both distributions with GMM, we recreate the fitted normal distribution in the dashed line on the graph. The green line that is vertically drawn is the cutoff that separates the two distributions, which is determined by identifying the crossover of the sample distributions from the left to right. The mean of the expression level of the samples on either side of the crossover is then calculated and deemed as the cutoff that separates the two distributions. Furthermore, in Figures 3.2b and 3.2d, a Q-Q plot

is shown to present whether the quantile of the fitted distribution is roughly similar to a given distribution of assumption, which in this case is assumed to be a normal distribution. If the quantile of the fitted sample and the quantile of the theoretical normal distribution match, we will observe the data lie along the 45-degree diagonal line. In Figure 3.2b, we see that the data perfectly lie on the 45 diagonal line when plotted against a theoretical normal distribution, indicating normality in the sample distribution. On the other hand, in Figure 3.2d, we see a skewed line on the diagonal at the tail of the distribution indicating non-normality.

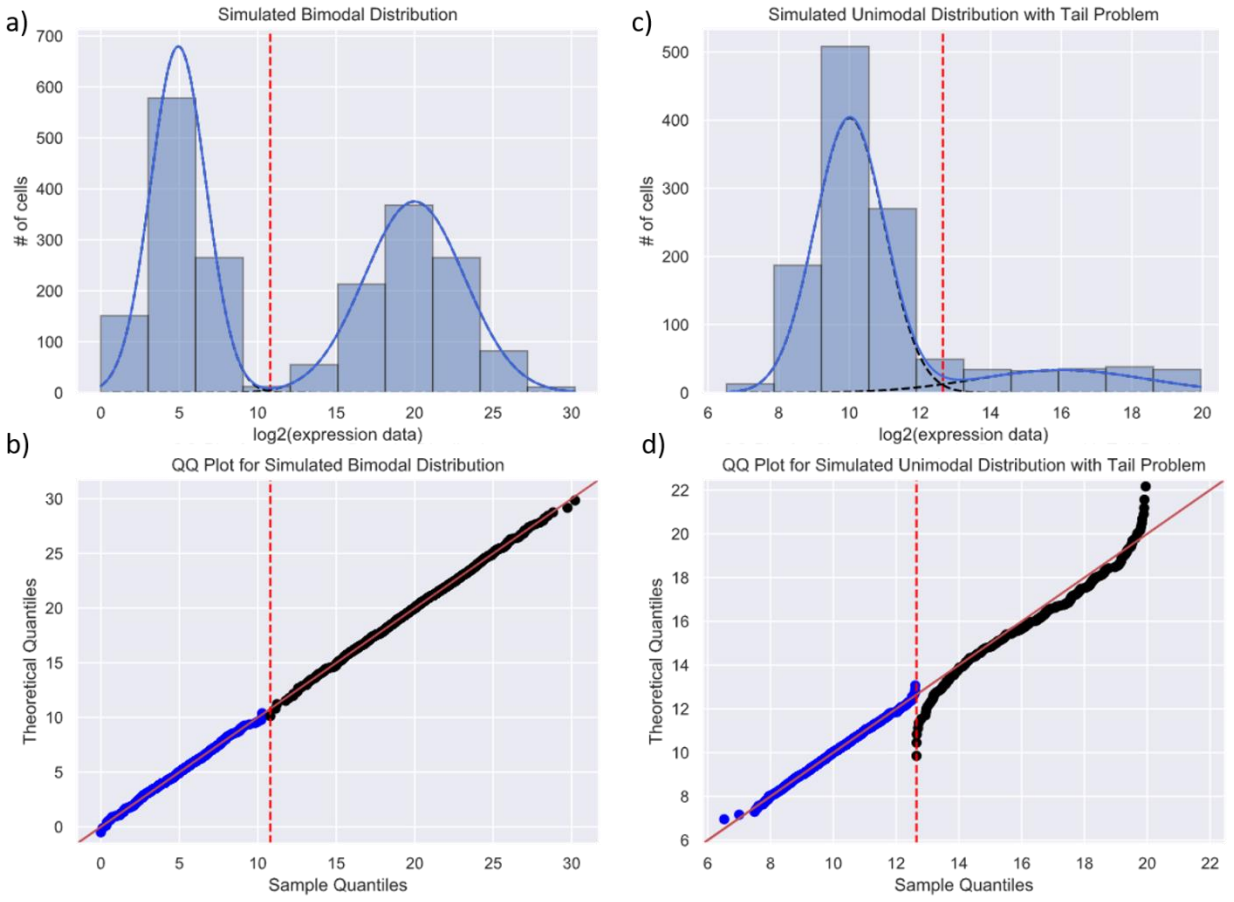


Figure 3.2 GMM on Simulated Mixture of Normal Distributions: (a) shows a mixture of two Gaussian components fitted onto a simulated bimodal distribution using sklearn.mixture. The red dotted line represents the separation between the two fitted normal components. In the case of two normal distributions, the red dotted line is the intersection between the two distributions where the probability of a data point categorized under either of the two distributions crosses over from to another. (b) is a Q-Q plot that measures the normality of the fitted data by plotting the theoretical quantile of a normal distribution with the computed mean and variance against the sample quantile of the input data. Normality is visualized by the scatterplot lying on the 45-degree line suggesting the quantiles of the theoretical and sample data behave similarly. (a) and (b) are examples of bimodal distributions adequately identified by GMM. (c) shows a simulated bimodal distribution with a 'tail problem', representing a non-normal spread of outliers, inadequately fitted by the GMM. Compared to (b) where the dots lie on the 45-degree line, (d) shows how a non-normal tail can skew the alignment of dots on the 45-degree line of the Q-Q plot. The slight uptick of blue dots near the red line represents the small overlap between the tail and the normal distribution on the left where the assignment of the distribution each datapoint belonged to was relatively more uncertain.

To deal with this challenge, we have developed a pipeline that can identify cutoffs between normal and non-normally distributed patterns. The primary objectives of the novel GMM method are thus defined as:

1) Develop a pipeline capable of discerning non-normally distributed tails and normally distributed subgroups of the gene expression data.
2) Develop an objective method that makes no assumptions and can be used to deal with a large body of data. The goal is for this method to determine a unimodal or bimodal pattern across a set of samples. In the case of a bimodal distribution, make the appropriate cutoff separating different samples into subgroups for downstream analysis.

The application of the novel GMM method will be demonstrated in Chapter 4 of this thesis. This chapter will solely focus on the explanation and validation of the novel method developed.

To prevent confusion of which algorithm the name 'GMM' refers to, we will hereon refer to the standard GMM as 'GMM' while referring to the novel GMM method as 'GMM with Chi-square Pipeline' or 'GMMchi'.

To deal with this challenge, we have developed a pipeline that can identify cutoffs between normal and non-normally distributed patterns. The primary objectives of the novel GMM method are thus defined as:

3) Develop a pipeline capable of discerning non-normally distributed tails and normally distributed subgroups of the gene expression data.

4) Develop an objective method that makes no assumptions and can be used to deal with a large body of data. The goal is for this method to determine a unimodal or bimodal pattern across a set of samples. In the case of a bimodal distribution, make the appropriate cutoff separating different samples into subgroups for downstream analysis.

The application of the novel GMM method will be demonstrated in Chapter 4 of this thesis. This chapter will solely focus on the explanation and validation of the novel method developed.

To prevent confusion of which algorithm the name 'GMM' refers to, we will hereon refer to the standard GMM as 'GMM' while referring to the novel GMM method as 'GMM with Chi-square Pipeline' or 'GMMchi'.

3.2.3 A Novel GMM with Chi-square Pipeline

In the following section, we will address each step of the GMMchi pipeline in depth. A common problem when fitting gene expression data exhibiting a tail problem with GMM, which assumes a mixture of Gaussian distributions of varying variances, is that the best-scoring model may result in either of the two scenarios:

1. Model adequately fits: GMM performs a good fit on a mixture of normal distributions.
2. Model overfits: GMM identifies a mixture of normal distributions where the data exhibit a normal distribution + tail problem.

In one of the two scenarios, our pipeline deals with the tail problem using a different method named the Chi-square iterative fit protocol, part of the GMMchi. This process is an iterative optimization consisting of four steps: dynamic binning, tail identification, iterative tail trimming, and the final condition criterion.

In all the simulated data below, we simulate normal distributions with ***multivariate.normal*** available in Python. The mean(s) and variance(s) of the normal distribution are inputted into the ***multivariate.normal*** function. The function is documented in the link:

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.multivariate_normal.html.

The function generates random samples that are plotted onto a histogram where we will further perform downstream testing analysis with our GMMchi.

3.2.3.1 Dynamic binning

One of the most important factors that determine how a series of data are presented in a histogram is the number of bins and the width of bins used. Too few bins are detrimental as that provides too little granularity and misrepresents the data, while too many bins result in many unoccupied bins that give a broken distribution.

To create a fully automated pipeline, we have incorporated the Mann and Wald bin criterion (Bowman, D'Agostino and Stephens, 1988) defined in Equation (3):

$$\boldsymbol{b = 1.88 \times n^{2/5}} \qquad \textbf{(3)}$$

The result *b* is the number of bins while *n* represents the number of samples. This creates a consistent and unbiased method that has been shown to perform stably in the literature.

Next, since the Chi-square iterative fit protocol within the GMMchi is based on the chi-square goodness of fit test as the main criterion for measuring the fitness of the model, an important assumption for a properly functioning chi-square goodness of fit test is having a minimum of 5 samples in each of the independent observation bins. Thus, dynamic binning is an

algorithm developed to achieve this goal: automatically combine bins while applying the least manipulation to the histogram for ensuring optimal results of the underlying chi-square test within GMMchi.

Dynamic binning is a three-step algorithm shown in the following:

1) Mann and Wald bin criterion to determine the number of bins b. Using b and the range of our data we could arrive at the width of our bins by calculating $\frac{range}{b}, range = max - min$.
2) Quantify the number of observations per bin.
3) Combine the bins that have observations < 5.

The results of dynamic binning can be seen in Figure 3.3. As shown in the result, this removes the worry of having too few observations within any given bins that may undermine the chi-square goodness of fit test used in the subsequent steps. Dynamic binning will thus be applied to all data before GMM fitting.

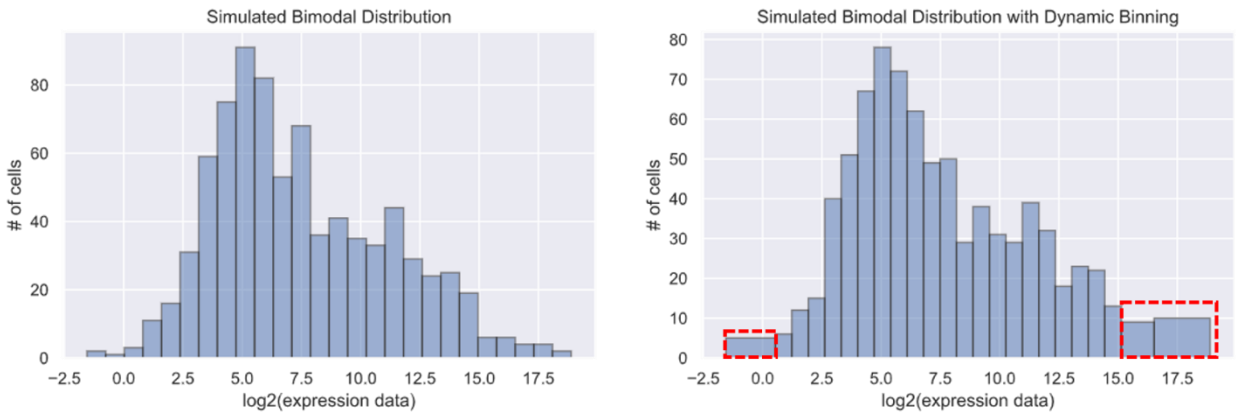


Figure 3.3 Dynamic Binning: Both graphs are based on simulated data that creates a bimodal distribution. The mean and variance of the simulated 400 random samples on each graph are 5, 10, and 3, 10, respectively. On the left is the histogram before applying dynamic binning. Samples

within each bin in the far left and right of the left histogram are less than 5 thus falling below the prerequisite of applying the chi-square goodness of fit test. On the right, dynamic binning solves this by dynamically combining bins without over-manipulating the overall distribution. The bins combined via dynamic binning are boxed in red dashed boxes.

3.2.3.2 Tail Criterion

The next step deals with the tail problem. Our pipeline first identifies whether the data is a mixture of normal distribution(s) or a mix of the non-normal tail and normal distribution(s). If the data fits the tail criterion and is deemed a tail problem, GMMchi will iteratively remove one data point at a time while fitting the remaining with GMM. At the same time, calculating both the χ^2 value as well as the BIC to determine the best fitting outcome. Below, we show how we arrived at our tail criterion of identifying the tail problem.

Initially, fit the input data with GMM after applying dynamic binning. The returned output of the GMM consists of the number of normal components fitted to the data based on BIC, the posterior probability of each data point (x), and the characteristics of each component including the mean (μ) and the variance (σ^2). To determine whether the resulting model is adequately fitting or overfitting the model with a non-normal tail problem, we have designed a method of identifying the tail problem.

Tails have a characteristic flat and spread-out distribution that usually represents the outliers of the input data. Under GMM, tails will be fitted as normal distributions thus will always have a lesser fit compared to a normal distribution. In other words, calculating the $\chi 2$ value of fitting 2 normal distributions to 2 normal distributions (bimodal distribution) will be much lower than (better fit) than that of fitting 2 normal distributions to a normal distribution plus a non-normal tail.

Based on this assumption, we can derive a criterion based on the fitness of how well our model is fitting the data. We have selected the chi-square goodness of fit test as our test for model fitness due to its simplicity and generalizability. Compared to the BIC, chi-square goodness of fit is not directly affected by the total sample size inherent in the penalizing term of the BIC thus allowing the chi-square value computed using the chi-square goodness of fit test comparable across different iterations. This is an important attribute that allows the comparison of the model fit across subsequent iterations that ultimately leads to the selection of the best fit.

To determine the acceptable χ^2 threshold value, we plotted the distribution of the χ^2 value of all of our genes shown in Figure 3.4. The degrees of freedom used to calculate the χ^2 of the estimated distribution is shown in Equation (4):

$$\boldsymbol{dof = b - 1} \qquad \textbf{(4)}$$

Since ***dof*** is calculated based on the number of bins ***b***, dof decreases very gradually as we remove 1 datapoint in each iteration. Nonetheless, since tails are in most cases outliers of the normal distribution, only a limited number of datapoints are located within the tail, resulting in a minimal effect in the decrease of ***dof*** towards the χ^2 value.

Initially, we removed any χ^2 values that are derived from a unimodal distribution because they do not lead to a tail problem. We then fitted the remaining data with GMM. We can clearly see a bimodal distribution (Figure 3.4). We assume that the two distributions in the histogram each represent well-fitted data with low χ^2 value (better fit) on the left and inadequately fitted data with non-normal tails on the right (worse fit).

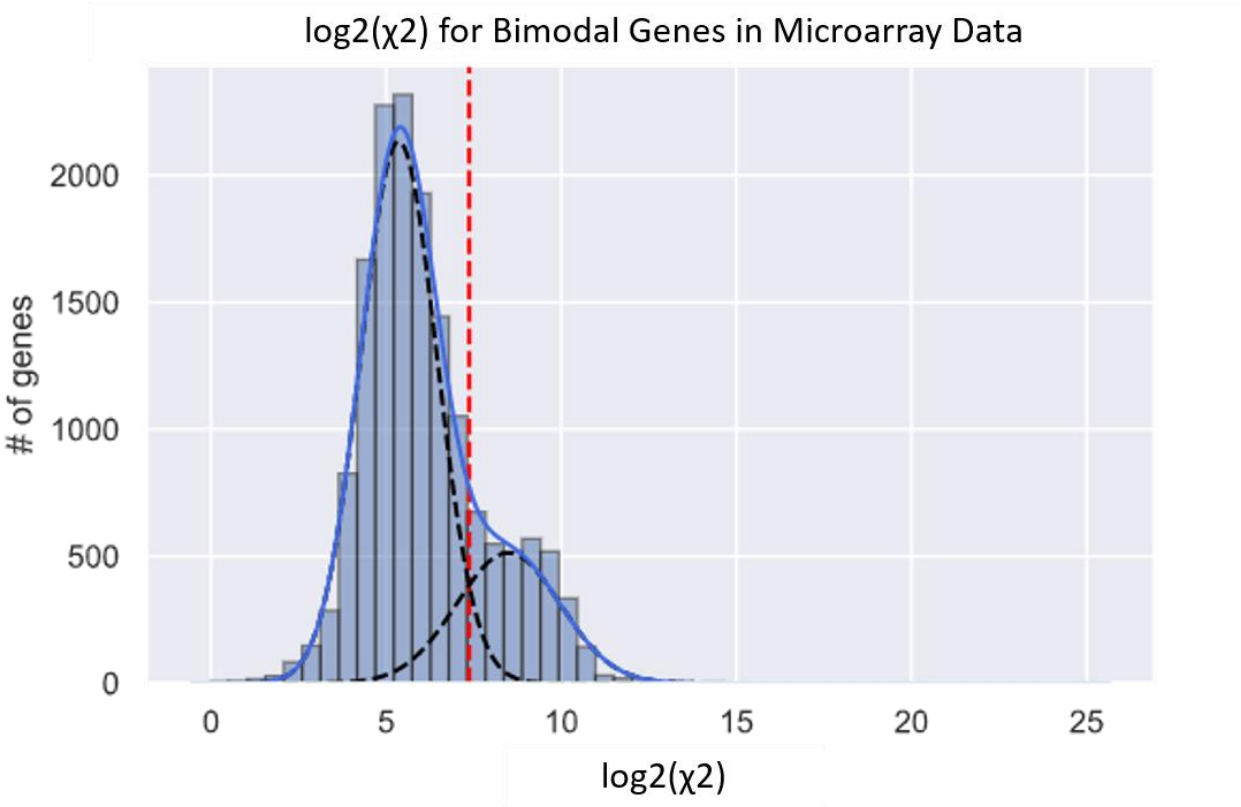


Figure 3.4 Determining the Chi-Square Threshold Value for Identifying a Poorly Fitted Tail: This is the distribution of the spread of the log2-transformed Chi-square χ2 value of bimodal genes using GMM. We do not consider unimodal distributions because they do not lead to a tail problem. In the graph, we see a clear bimodal distribution with two normal components. We assume the two distribution represents: 1) The left distribution represents the well-fitted distribution with a lower χ2 value while 2) the right represents an inadequately fitted distribution containing non-normal components such as a tail. Using the computed mean of the right distribution of 8.48 and standard deviation of 1.5, we can derive the lower threshold of 5.52 that reduces the false-negative rate to 5%.

Through multiple testing, we have concluded to use the threshold derived from calculating 2 standard deviations ($\sigma = 1.5$) away from the mean ($\mu = 8.48$) of the right distribution. The resulting χ^2 value of $2^{5.52}$ or 36.8 is set as the hard-lined threshold used for discerning a poorly fitted distribution with a tail problem vs an adequately fitted distribution(s). In other words, any χ^2 value resulted from the initial fit that is above this threshold will be considered as an inadequate model fit (tail problem) necessary for further processing.

3.2.3.3 Iterative Tail Trimming (ITT)

After determining whether a distribution is exhibiting a tail problem, our next goal is to identify the correct threshold that separates the tail from the main distribution. As discussed in the previous section, the tail criterion will label the result as 'pass' or 'tail-problem'. If the model is labeled as 'pass', the model passes the criterion and the result of the GMM will be respected and returned as the final output. If the model is labeled as a 'tail-problem', we assume the model inaccurately fitted the data and that the data consist of a mixture of normal distribution(s) with a non-normal tail and should be subjected to iterative tail trimming.

Iterative tail trimming (ITT) is an optimization process where the algorithm searches for the most probabilistic datapoint where the cutoff can be made to separate the tail from the rest of the normally distributed data. The process is as follows:

1) At each iteration, we begin by removing one data point from the tail, perform dynamic binning to adjust the appropriate number of observations per bin, and fit the resulting distribution with GMM.
2) Then, we determine the chi-square goodness of fit (χ^2) using parameters from the output to understand how well the model fits the input data.
3) Compare $\chi^2_{\ n}$ and $\chi^2_{\ (n-1)}$, the χ^2 of the current iteration (n) and the last iteration (n-1), as well as the BIC score to determine whether the model is fitting better than the previous fit. It is important to note that we utilize the BIC score to determine whether a lower χ^2 value is also the least complex model with the most optimal fit.
4) The algorithm iterates 1), 2), and 3) until no other conditions produce a lower χ^2 and BIC.

The result of ITT is thus a series of χ^2 values revealing the fitness of the model at each iterative step, as shown in Figure 3.5. In this case, the lower the χ^2 value the better the model fitness. Even though BIC is a criterion that is equivalent to the χ^2 value with a penalizing term, as shown in section 3.2.1, the reason why we are unable to solely depend on BIC is due to the penalizing term that is dependent on the sample size. Since we are removing 1 datapoint at each iteration during ITT, the sample size *k* is constantly decreasing making BIC a relative value that is only relevant when comparing between input with the same sample size. This means that the χ^2 value itself is an absolute term we use to compare across all data points to determine how well the new model fits and the BIC values are relative terms used to compare, within the same sample data, what is the suitable number of fitted distributions.

In Figure 3.5b and 3.4c, we have shown the BIC and χ^2 value, respectively, during the process of ITT that resulted in Figure 3.5a. The line graph in Figure 3.5c shows the $\chi 2$ value that steadily decreases as we remove one data point at a time. At each lower χ^2 value that passes the BIC criterion, the graph denotes it with a red arrow. Thus, we can observe that the lowest χ^2 value was achieved after removing 9 data points, which in this case represents the non-normal tail. Interestingly, in Figure 3.5f, even though we see a lower χ^2 value after removing 6 as well as 23 data points, they are not marked with the red arrow due to the disagreement with the BIC score (Figure 3.5e). In other words, we are achieving a greater fit but with a much more complex model that is reflected and rejected by the BIC criterion. In the end, removing 16 datapoints returned the most optimal output shown in Figure 3.5d. A red dotted line is then drawn at the number of data points determined to result in the most optimal fit as described above. In the case of MUC2,

a red dotted line that separates the high and low expression groups are thus drawn at 16 data points from the right.

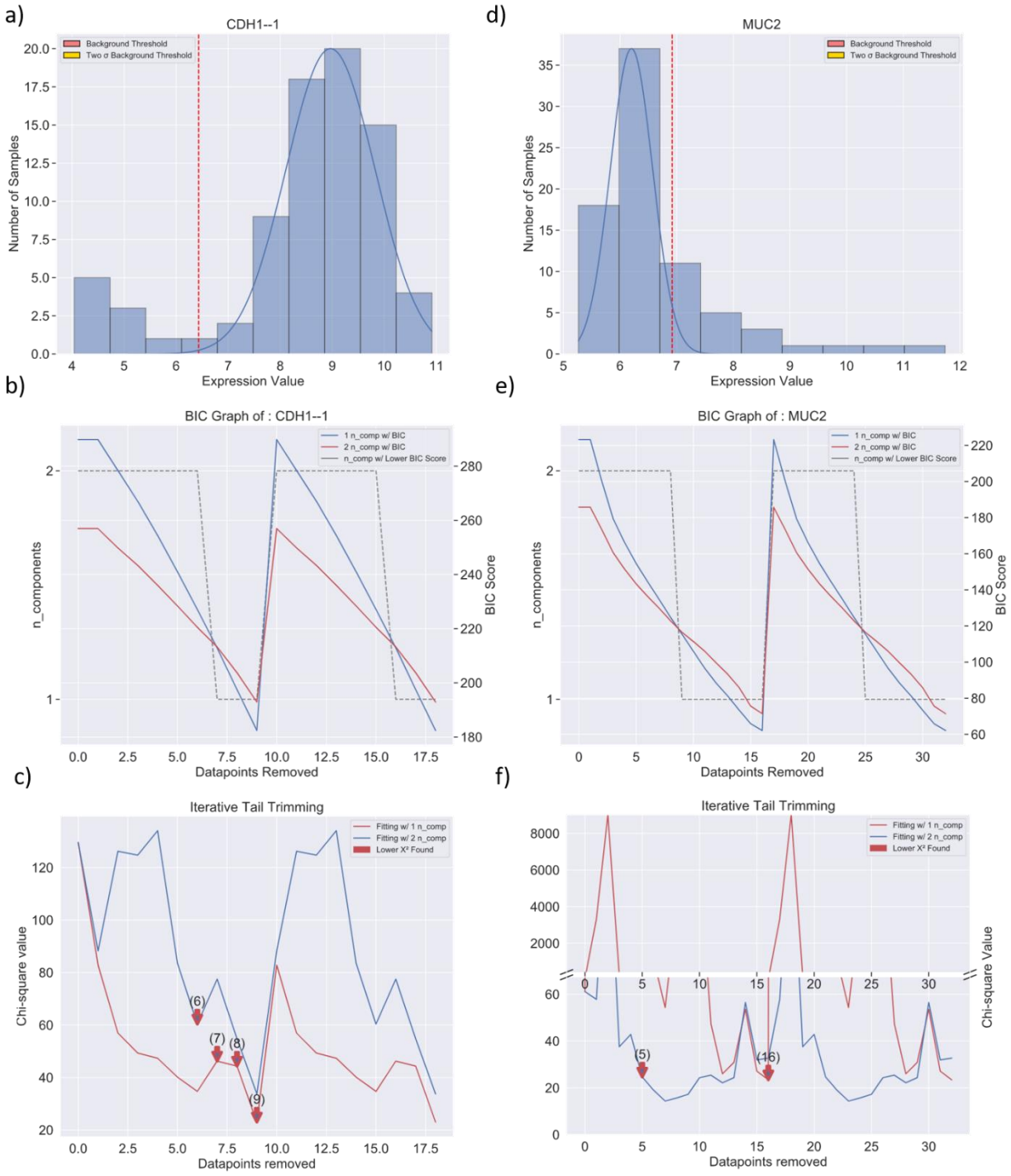


Figure 3.5 Examples of Iterative Tail Trimming on CDH1 and MUC2: (a) and (d) are both examples of a successful implementation of cutting of the tail using GMMchi. (a) is the second probe of CDH1, an epithelial biomarker, in our microarray data on the cell lines. (b) and (c) are the BIC value and χ^2 value, respectively, at each step during iterative tail trimming performed on the CDH1 expression data across our cell lines. At each lower χ^2 value (better fitness) that passes the BIC criterion, the data is marked with an arrow to indicate a better fitness found. The red and blue line in (b) and (c) represents the fit for 1 and 2 normal components, respectively. The grey dotted line in (b) and (e) represents the choice of either 1 or 2 normal components based on the BIC score. (d) is the gene MUC2, a secreted glycoprotein that marks goblet cells. (e) and (f) are the BIC value and χ^2, respectively, value at each step during iterative tail trimming performed on MUC2 expression data, revealing the best-fitted model at removing 16 data points from the right tail, whereas for CDH1 it was at 9 data points removed. A good example that demonstrates the lack of agreement between χ^2 and BIC is observed in (e) and (f) when we measured a lower χ^2 when fitted with 2 normal components after removing 7 datapoints but did not mark it with a red arrow. The reason is that BIC selected one normal component, which contradicted the 2 normal components selected by χ^2. We only achieve a lower χ^2 agreed by BIC when we removed 16 data points as shown in (f). Finally, the red dotted line in (a) and (d) is objectively drawn at the number of data points determined to be removed for the rest of the data to fit a normal component(s). In the case of MUC2, the red dotted line is drawn at 16 data points from the right.

3.2.3.4 Multiple Conditions for Iterative Tail Trimming

Having established Iterative Tail Trimming (ITT), there is another important factor to consider before applying ITT. If we look at all possible distribution patterns a tail problem can give rise to in a GMMchi analysis, it results in four possible scenarios, also shown in Figure 3.6:

1) One normal distribution plus a left tail
2) One normal distribution plus a right tail
3) Two normal distributions plus a left tail
4) Two normal distributions plus a right tail

The algorithm takes all possible scenarios listed above into account and performs ITT on the data. What it does is that it is essentially running ITT while removing data from the left or right tail and fitting the remaining data with one or two mixture of Gaussian distributions, covering all four possible scenarios mentioned above. This ensures all scenarios are well-

considered within our model and the output represents the optimal tail trimming of the right model.

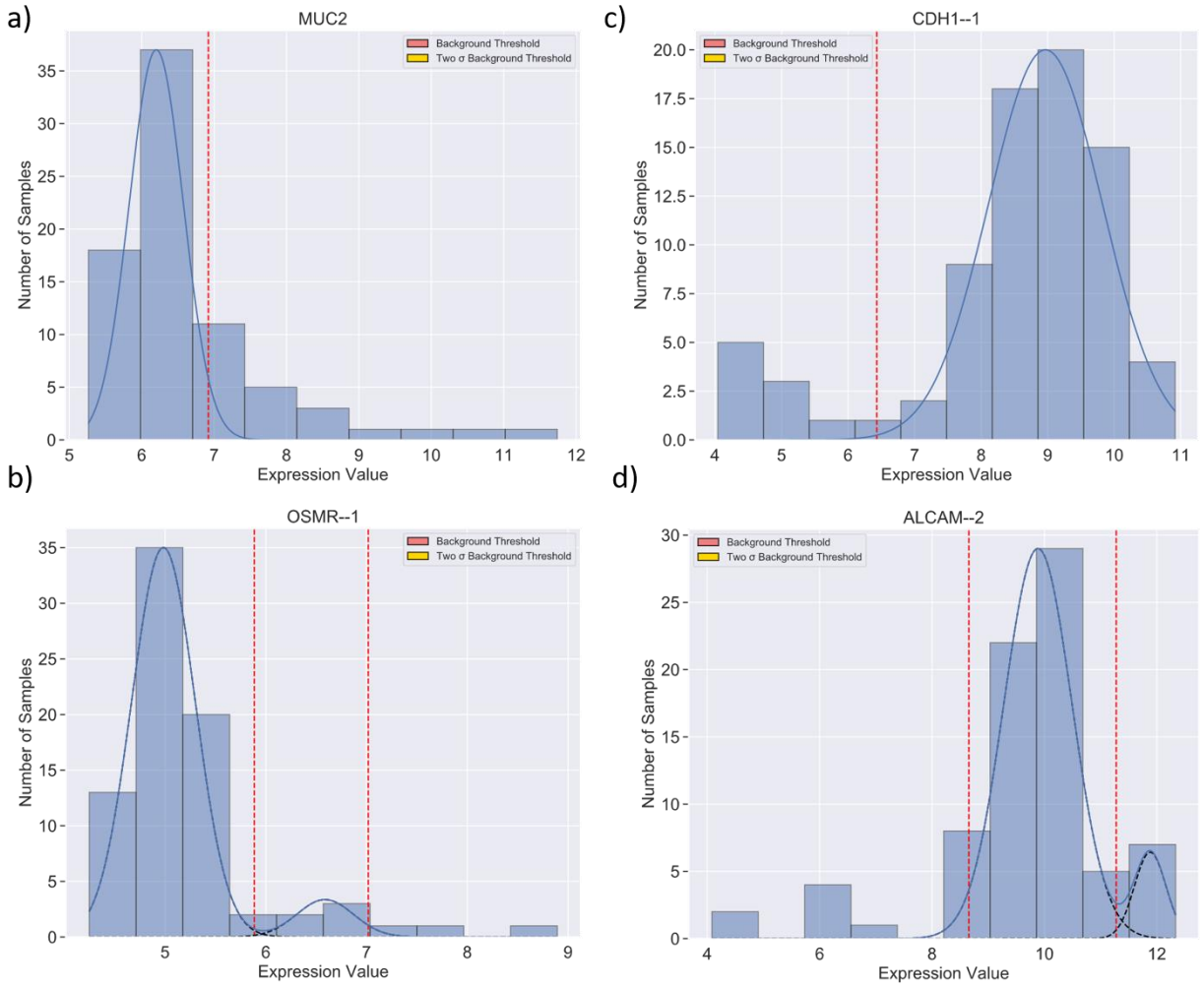


Figure 3.6 Examples of Different Distributions Found in Gene Expression Analysis: (a), (b), (c), and (d) demonstrate good examples of a unimodal distribution with a right tail, a bimodal distribution with a right tail, a unimodal distribution with a left tail, and a bimodal distribution with a left tail, respectively.

3.2.3.5 Final criterion for Determining the Cut-off that Separates the Main Distribution from the Tail

The output from the ITT will be a series of $\chi 2$ values for each condition at each iterative step. Notably, the $\chi 2$ value derived from the chi-square goodness of fit test is an absolute

quantitation rather than a relative quantitation seen in BIC. Therefore, we can directly compare the $\chi 2$ value calculated for all scenarios and conditions and select the final condition for our analysis.

The final criterion is thus determined by identifying the lowest chi-square value, which indicates the best fit, and the lowest BIC score, which suggests the least complex model, across all 4 possible scenarios. This results in the ultimate removal of the tail. The cutoff is then redrawn at the conditions that fit the final criterion.

3.2.3.6 Returned Output of GMMchi

The resulting best fit can be either the original fit, indicating that a mixture of normal distributions best represented the data, or a new fit consisting of a mixture of normal distributions and a non-normally distributed tail.

The output of GMMchi includes the categorization of input data across all samples and a set of the means and variances that describe the fitted distribution.

The visual output consists of 4 graphs seen in Figure 3.7: the distribution of the input data (shown as a histogram) with the fitted model (shown as the line graph mapped onto the histogram) (Figure 3.7a), a graph of the percentage of each categorized group (Figure 3.7b), a graph of the BIC score that decides the number of components at each iterative fit (Figure 3.7c), and a graph of the $\chi 2$ value that decides the best-fitting model at each iterative fit (Figure 3.7d).

Through this pipeline, we can begin to identify subgroup patterns across cell line and patient gene expression data using an objective and automatic pipeline applicable for large-scale analysis.

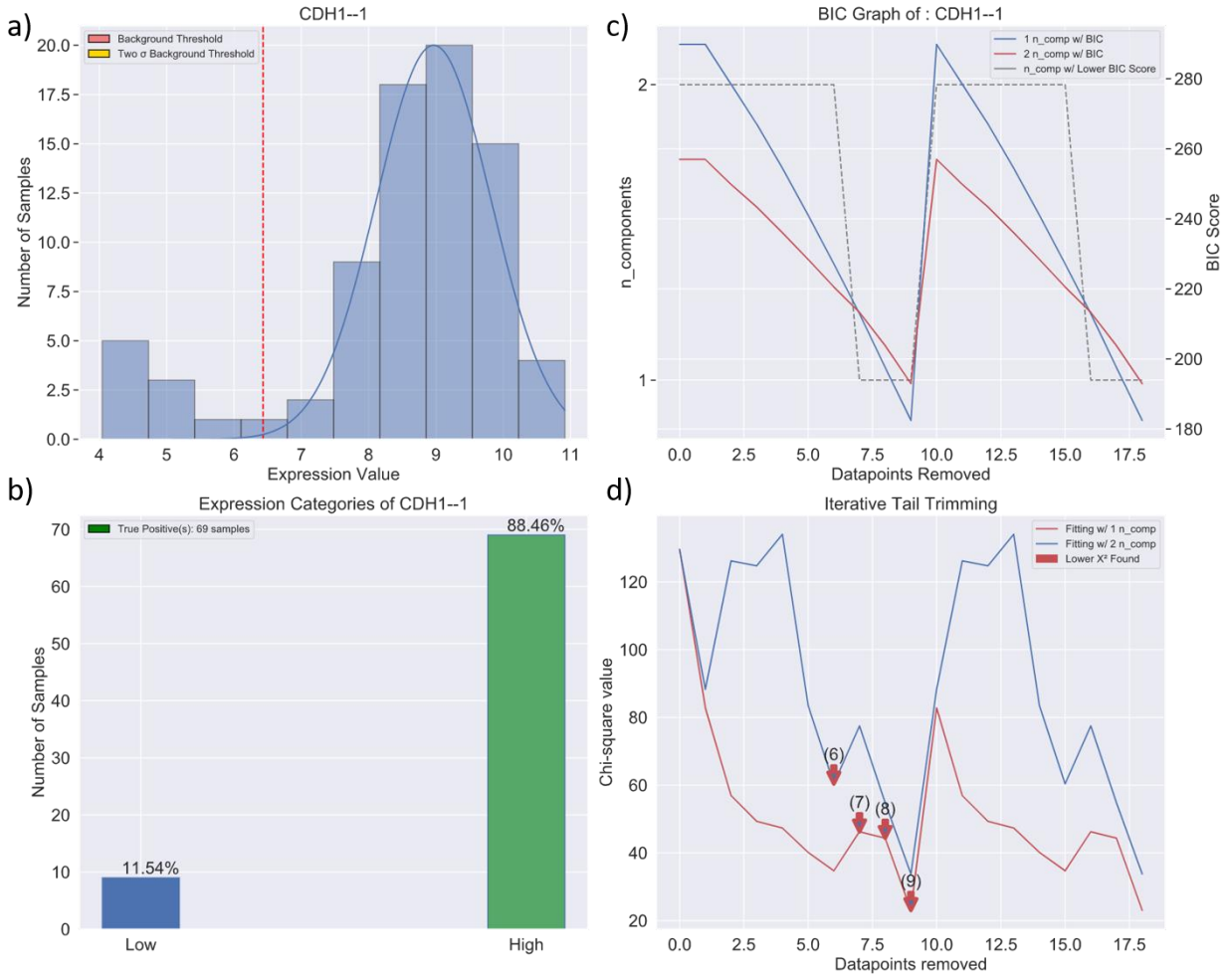


Figure 3.7 Example Output of GMMchi Python Package: The visual output of GMMchi. (a) is the log2-transformed distribution of the gene expression level of CDH1 with its fitted normal components along with a red vertical cutoff from the tail. (b) is the BIC score. The red being the BIC score of fitting 2 normal components, the blue being the BIC score of fitting 1 normal component and the grey dashed line indicating which model to pick at each iterative step. (c) is a bar graph showing the percentage of samples under each categorized group. (d) is a graph showing the χ^2 value calculated at each iterative step during iterative tail trimming. The red line represents the χ^2 value of fitting 1 normal component, the blue line represents the χ^2 value of fitting 2 normal components, and the red arrows indicate when a better fit is found.

3.2.4 GMMchi validation

To test the robustness and sensitivity of the GMMchi pipeline, we performed two separate simulations. We also included the code in Supplementary Code S1 in the Appendix.

Simulation test #1

Tail identification test: The purpose of this simulation is to assess how well GMMchi is at identifying a simulated mixture of normal and non-normal tails with the increasing spread of the tail. The spread is defined as the range the tail stretches across. The larger the spread the larger the range the tail spreads across and vice versa. We separated this simulation into two parts:

1) Unimodal case: Mixture of 1 normal distribution + 1 non-normal tail
2) Bimodal case: Mixture of 2 normal distributions + 1 non-normal tail

Method: In this simulation, we are creating a mixture of 1 or 2 normal distribution(s) and 1 non-normal tail. We randomly generate normal distributions with a mean anywhere between 5 to 10, the standard deviation anywhere between 0.5 to 5. The total number of data points is set to be anywhere between 50 to 150. The number of data points in the tail is randomly chosen to be anywhere within 5-25% of the total number of data points. To test how well GMMchi performs with an increasing spread of the tail, we determined the lower bound of the tail with $lower\ bound = m + 2 \times std$ where *m* and *std* is the mean and standard deviation of the normal component so we make sure the tail starts at the upper end of the distribution. In the case of having 2 normal distributions, we always use the *m* and *std* of the right distribution. The upper bound is determined by $upper\ bound = lower\ bound + n \times std$ where n incrementally increases from 1 to 12 thus resulting in a larger spread of the tail as *n* increases with each

simulation. Each run, we generate 5000 mixtures of a normal component(s) and a non-normal tail. After generating each simulation, we ran GMMchi on each of the mixtures to determine the pattern and compare it with the ground truth to calculate the accuracy. The accuracy is thus the percentage of correct categorization.

Simulation test #2

Accuracy of correct categorization: The purpose of this simulation is to assess how accurate GMMchi is at identifying the correct categorization at different sample sizes. The four categorizations we included in this simulation are:

1) Unimodal distribution
2) Bimodal distribution
3) Unimodal distribution + tail
4) Bimodal distribution + tail

Method: In this simulation, we are generating 500 distributions consisting of 25% or 125 distributions of each of the categorizations presented above. The mean of the normal component(s) is anywhere between 3 to 10, the standard deviation is anywhere between 0.5 to 5. The number of data points in the tail is randomly chosen to be anywhere within 5-25% of the total number of data points in the normal component(s). The upper range of the tail is anywhere below 7 standard deviations away from the normal component. After randomly generating 500 simulated mixtures, we ran GMMchi on each of the mixtures to determine the pattern and compare it with the ground truth to calculate the accuracy. The accuracy is, again, the percentage of correct categorization.

We discuss the results of our simulations below:

3.2.4.1 Simulation test #1: Tail Identification

The result in Figure 3.8a shows that once the spread of the tail exceeds 2 standard deviations, the accuracy reaches up to 90%+ with the accuracy plateauing at 97% above 3 standard deviations.

The same test was done on simulated bimodal distributions with a tail. The result in Figure 3.8a shows the accuracy reaches above 70% accuracy when the spread of the tail exceeds 2 standard deviations. However, the accuracy also quickly falls after a spread of 7 standard deviations, suggesting GMMchi is most sensitive within a tail spread range of 2 to 6 standard deviations. We suspect this is mainly due to GMMchi utilizing a static number of bins retrieved using the total number of input data. As the tail moves vastly away from the main distribution, in this case, a bimodal distribution, the bimodal distribution will be misrepresented as a unimodal distribution due to the static number of bins, leading to the drop in accuracy. In Supplementary Figure S1, we show that as the spread increases (moves away from the main distribution), the percentage of samples that are identified to exhibit a unimodal distribution with a tail sharply rises while the percentage of samples that are identified to exhibit a bimodal distribution with a tail sharply drops, suggesting a high percentage of bimodal distribution with a tail misidentified as unimodal distributions with a tail. Nonetheless, GMMchi is still able to accurately identify the non-normal tail and delineate the threshold drawn between the main normal component and the tail.

These simulations give insight into the sensitivity of the tail identification step in the pipeline. This also shows that GMMchi is consistently more sensitive with 1 normal component with a tail compared to 2 normal components with a tail.

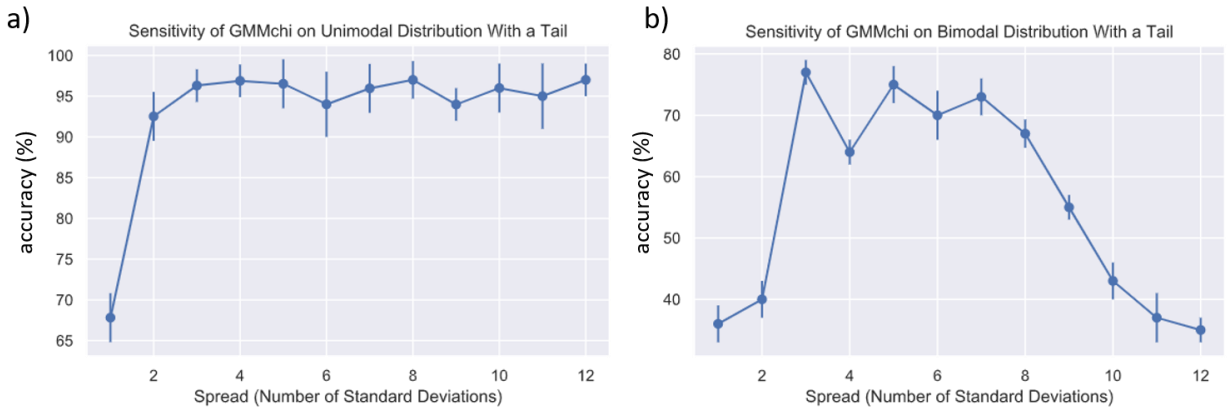


Figure 3.8 GMMchi's Accuracy of Tail Identification: (a) and (b) shows the sensitivity of the GMMchi pipeline is toward identifying tails with a unimodal distribution or a bimodal distribution, respectively.

3.2.4.2 Simulation test #2: Accuracy of Correct Categorization

The result in Figure 3.9 showed a general trend of increased accuracy as the sample size increases. GMMchi's accuracy reaches around 85% with a sample size of around n = 90. This suggests that GMMchi functions relatively well in a sample size of n > 90. In the current landscape of a lot of datasets exceeding sample sizes much greater than 100 (i.e., TCGA CRC cancer data with 637 samples or single-cell sequencing data with over thousands of individual cells), we believe GMMchi will be able to maximize its value in identifying important patterns. In our microarray data of 78 CRCs, we would expect an accuracy of around 80% while in the TCGA CRC data, which consists of 637 samples, we would expect an accuracy of around 91%.

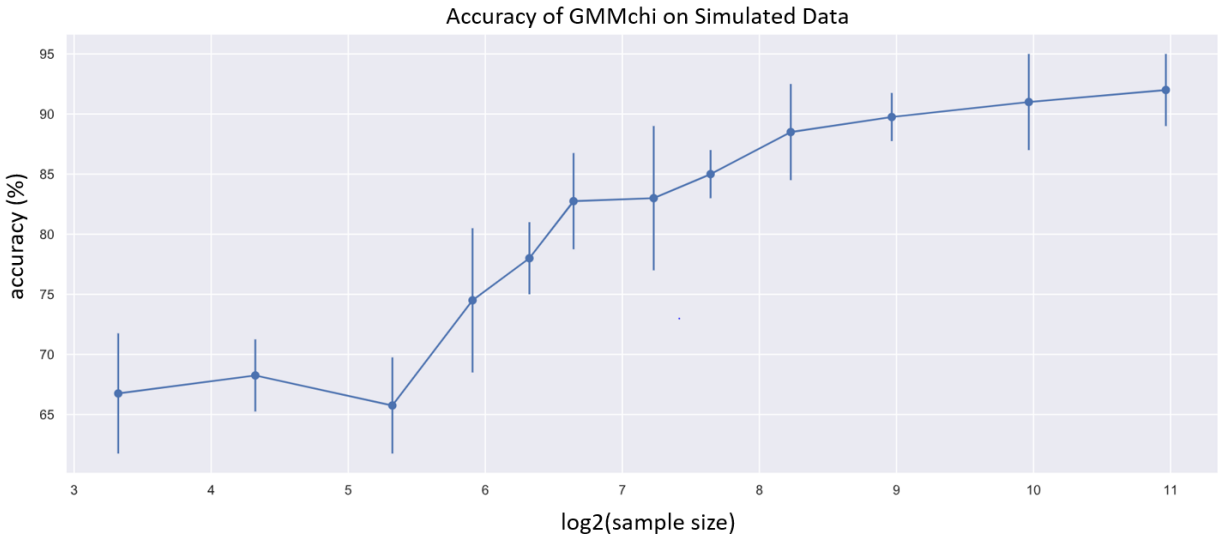


Figure 3.9 GMMchi's Accurarcy on the Correct Categorization of Simulated Distributions: The graph shows the accuracy of the GMMchi pipeline with increasing sample sizes. As expected, the accuracy of GMMchi increases as the sample size increases.

3.2.5 GMMchi on 78 Colorectal Cancer Cell Line Microarray Data

3.2.5.1 Background Noise Threshold

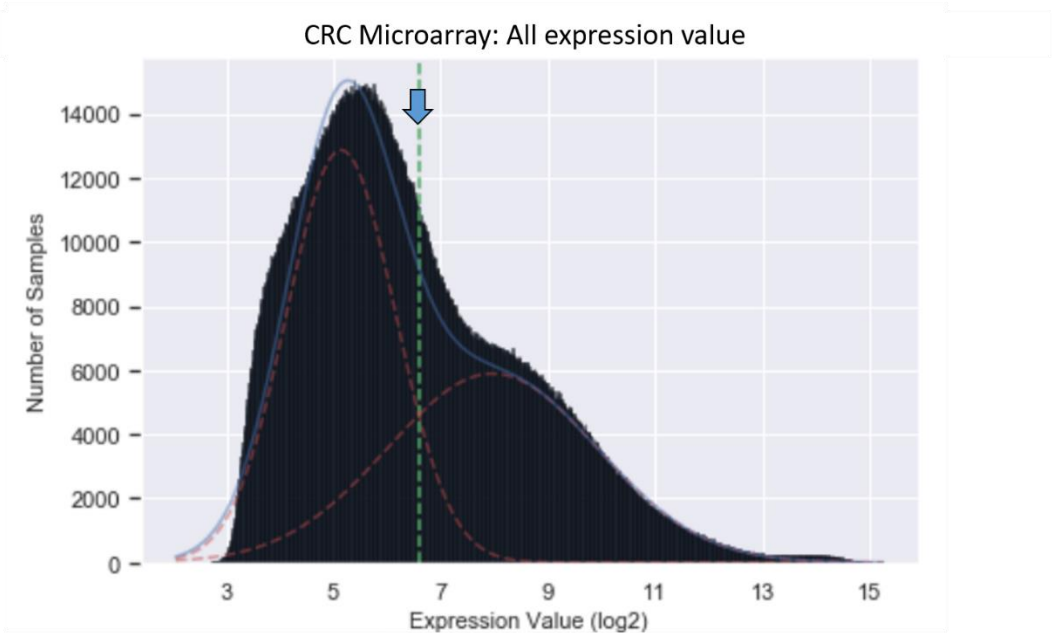
In this section, we apply GMMchi onto our CRC cell line microarray data (n = 78). Background noise threshold is an essential value needed when exploring expression data especially performed on any fluorescence-based platforms. The background threshold accounts for the within-array fluorescence noise and enables downstream analysis to statistically discern absent versus low expressing genes. In our lab, our previous analyses on the microarray data determined values within a range of 100 ($2^{6.65}$) and 128 (2^{7}) as the appropriate noise threshold. The range is drawn from many years of experience and experimental work. To validate whether previous observations can be reflected through a GMMchi analysis, our initial goal is to statistically determine the background threshold using GMMchi.

Based on *a priori* knowledge, we expect our GMMchi analysis to output a bimodal distribution encompassing a distribution representing background noise and a true positive

distribution when we map out all 54,675 probes of 78 CRC cell lines. As seen in Figure 3.10a, a histogram shows the wide range of expression values of 54,675 probe sets in a single distribution. Performing GMMchi on all the expression values showed a mixture of two normal components. The cutoff between the two normal distributions was defined as the background noise threshold, which was $2^{6.59}$ or 96.4. This result agrees with the previous observation of 2^7 as a rule of thumb for background threshold. Also, we will include a background threshold at 2 standard deviations ($\sigma = 1.01$) away from the mean ($\mu = 5.14$) of the background distribution to ensure we remove 95% of the background expression level. This will ensure more confidence in the identification of real positives for further GMMchi analysis.

Therefore, in any future analysis, the background noise threshold will be included to determine absent expression. In other words, any expression value less than the background threshold is determined to be an absent expression. This variable is a built-in function within the current GMMchi pipeline. Essentially, cell lines that undergo GMMchi are categorized into three possible groups: background, low, and high. The three groups each represent a specific biological context that is taken into consideration for downstream analysis.

a)



b)

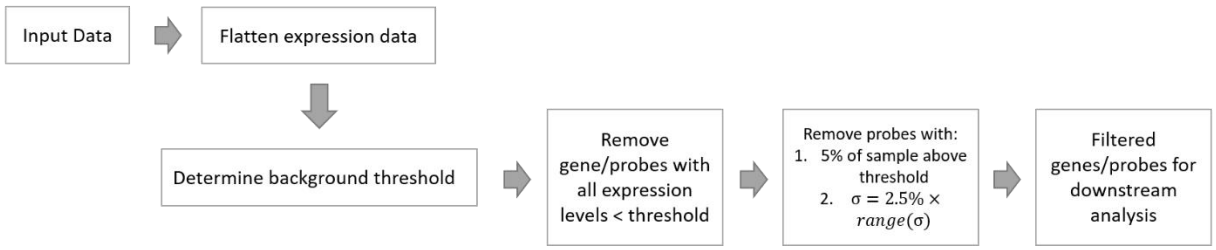


Figure 3.10 Determining Background Threshold in Microarray Data Using GMMchi: (a) is the distribution of all of the expression levels of all 78 CRCs within our microarray data. The green line, also where the blue arrow is located, delimits where the background noise threshold is that separates the background noise distribution (left) and the true positive distribution (right). (b) is the probe filter pipeline for removing probes or genes that are considered absent gene expression.

3.2.5.2 Probe filter

A probe filter is an implementation of a pipeline that removes genes that are considered as non-expressing genes seen in Figure 3.10b. The goal of this pipeline is to remove genes that are not expressing but may exhibit a pattern that can be misunderstood for a gene of interest. Besides, this process also reduces the computational load by removing genes that are unnecessary for analysis.

The function **GMM.probe_filter** initially removes genes where all cell lines are expressing levels below the background threshold. Next, the pipeline searches for genes that may be on the borderline of being a non-expressing gene. The criterion is set to filter genes that have less than 5% of the total cell lines expressing as well as ensuring the variance of the distribution is greater than 2.5% of the total range of variances across all genes. In other words, we are only keeping genes that have most cell lines expressing some level of expression with a certain variance.

3.2.5.3 Unimodal categorization criterion

Another form of pattern that revealed two distinct groups, other than bimodal distributions observed in our expression data is in fact some of the unimodal distributions (Figure 3.11). We noticed that nearly half of the unimodal distributions had parts of the expression under the background threshold indicating an absence of expression and the rest expressed. This is an interesting observation since it conveys a pattern where a group of cell lines does not express a certain gene while the rest does. Inherently, the different groups observed in unimodal distributions are of similar interest to bimodal distributions. The expectation for each situation is thus defined as follows:

1) The expectation for bimodal distribution is to identify:
    a. Low expression group vs high expression group
    b. No expression group vs high expression group
2) The expectation for unimodal distribution is to identify:
    a. No expression group vs high expression group

Therefore, either scenario leads to the possible assumption of inherent changes on the genetic or epigenetic level, making unimodal distribution with two groups also an interesting pattern to explore. Thus, we included unimodal categorization in our GMMchi pipeline for downstream analysis. This type of unimodal distribution will be hereon referred to as 'categorical unimodal distribution' for this thesis.

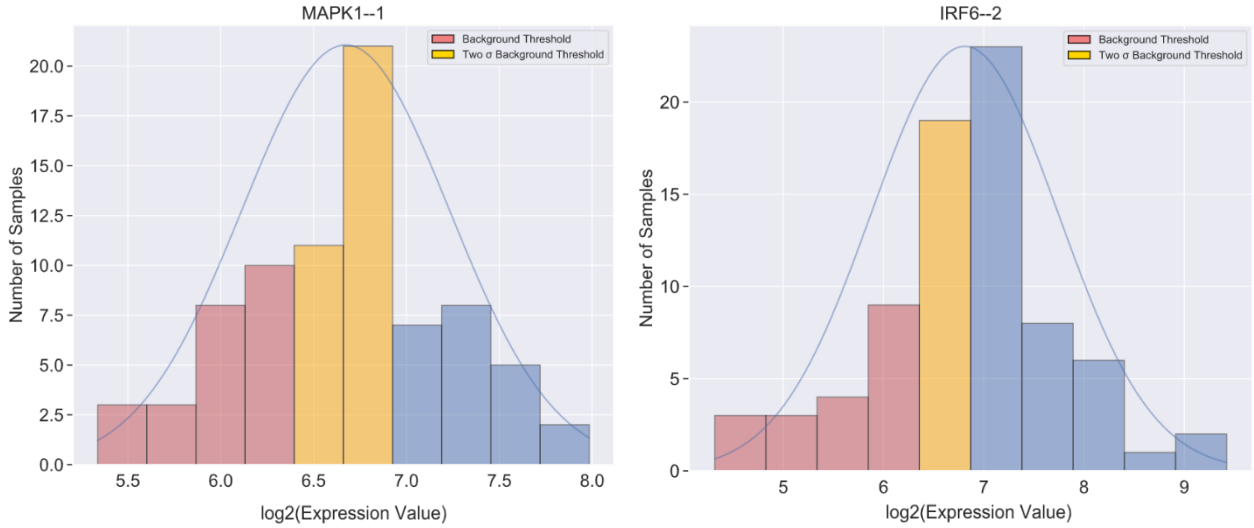


Figure 3.11 Examples of Genes Exhibiting Categories in Unimodal Distributions: Both unimodal distributions demonstrated above show a categorical pattern with parts of the distribution consisting of an absent expressing group and the rest consisting of an expressing group.

3.2.6 GMMchi Characterization on Microarray Data of a Panel of 78 Colorectal Cancer Cell Lines

GMMchi-based categorization was performed on our microarray dataset of a panel consisting of 78 CRCs. For the cell line dataset, each microarray chip produced 54,675 probe sets. Initially, only 31,917 (58.4%) probes passed the probe filter criterion. Out of those that passed the probe filter criterion, 24,214 (75.9%) were deemed unimodal, 7,703 (24.1%) were scored as bimodal. Within the bimodal genes, 3,520 (45.7%) exhibited a tail problem. Applying GMMchi, out of 3,520 probe sets identified with a tail problem, 2,624 (74.5%) were deemed unimodal with

a tail, 896 (25.5%) were deemed bimodal with a tail, and none were deemed uncategorizable. Thus, the final output resulted in 4,183 of 54,675 (7.6%) probes splitting into a bimodal distribution, 896 of 54,675 (1.6%) probes split into a bimodal distribution with a tail, 2,624 (4.8%) probes exhibited one unimodal distribution with a tail, 6,749 (12.3%) was within the categorical unimodal criterion and with the rest of the probes (31.9%) exhibiting unimodal distributions across our CRC panel. 22,758 probes (41.6%) did not pass the probe filter criterion and were deemed as non-expressing probes.

To illustrate the power of GMMchi on categorizing continuous microarray gene expression data, some examples are demonstrated in Figure 3.12. The gene *EPHB3* encodes Ephrin Receptor B3, the largest subgroup of the receptor tyrosine kinase involved in mitogenic and differentiation signaling (Figure 3.12a). A clear cutoff is made at 9.4 separating the mixture of two normal components. The Q-Q plot is also powerful proof revealing a convincing cutoff separating two normal distributions. Another gene *CDH1* encodes E-Cadherin, a key biomarker for differentiated epithelial cells. Most cell lines express high levels of E-Cadherin due to the epithelial nature of colorectal cancer cells, but a small subgroup expresses low or no expression and a correlatively low EPCAM expression. In Figure 3.12b, *CDH1* is identified to have a tail problem. GMMchi can identify the tail and perform the ITT pipeline that results in a unimodal normal distribution with a tail.

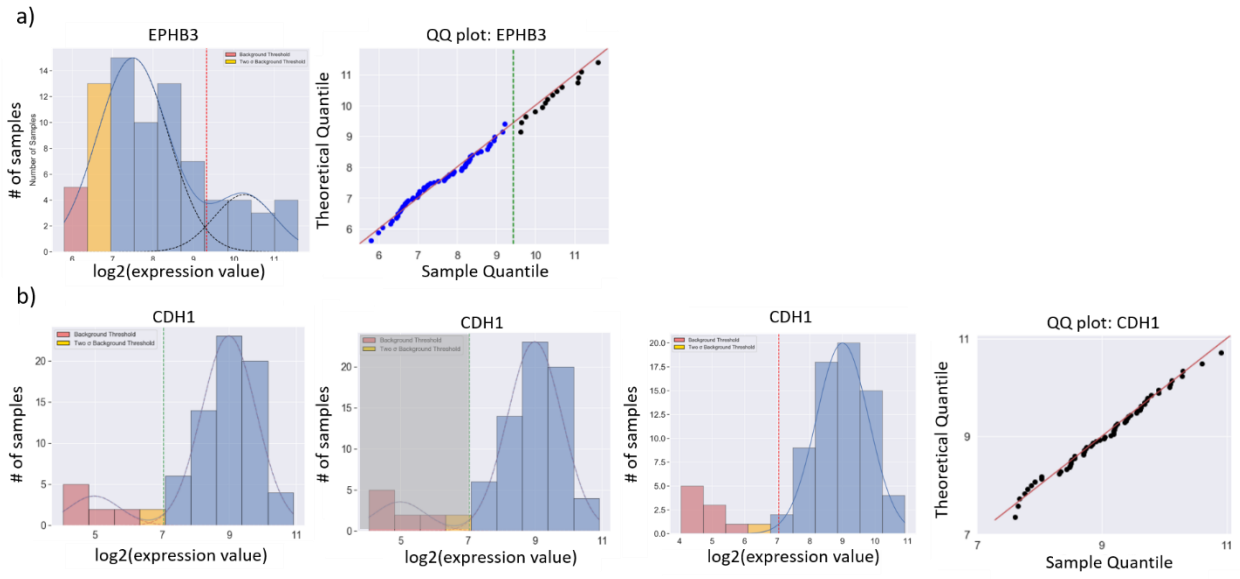


Figure 3.12 GMMchi on Gene Expression Data: (a) is the gene EPHB3, encoding the protein Ephrin Receptor B3. In this example, GMMchi is capable of drawing the cutoff between the mixture of two normal components that are adequately normal as shown in the Q-Q plot with the data aligning on the 45-degree line. (b) is a good example of a tail trimming process where the first graph shows the distribution of CDH1, an epithelial biomarker, fitted by two normal distributions. The second graph is the tail identification step where the grey shading indicates the potential identification of a non-normal tail. The third graph is the result of the step iterative tail trimming while the last graph reveals the Q-Q plot of the resulting fit, revealing an adequately fitted normal component.

3.2.7 *GMMchisquare* Python Library

To increase the utilization and ease of adoption of the GMMchi pipeline, a python library was built to facilitate easy usage. There are a couple of built-in methods (functions) within the pipeline that includes:

a) **GMM.GMM_modeling**: This method is used to perform GMMchi on an input matrix of sample x features.

b) **GMM.probe_filter**: This method is used to perform probe filtering to remove genes that are considered not to be expressed.

c) **GMM.find_hits**: This method is used to perform 2x2 contingency table analysis with the categorized data returned from the GMM.GMMmodelingt output (further discussed in Chapter 4)

d) **GMM.run_hits**: This method is used to output the full 2x2 contingency table from a pre-defined set of genes of interest. (further discussed in Chapter 4)

e) **GMM.crosstab_table**: This method is used to visualize the full 2x2 contingency table from a pre-defined set of genes of interest. (further discussed in Chapter 4)

The program was implemented in Python and can be imported as a library; the full code is shown in the appendix, the library and installation link are online on https://github.com/jeffliu6068/GMMchi

Chapter 4: Applications of the GMMchi Pipeline

4.1 Introduction

Tumorigenesis shares many of the same principles with Darwinian Evolution in the sense that cells compete and carry out to select the most advantageous phenotype that results in the greatest fitness. The overall fitness is thus measured by the average phenotype individual genetic changes contribute. Most phenotypic changes trace back to genetic mutations and epigenetic changes as discussed in Chapter 3. These changes, depending on factors including the function of the gene, the location of the mutation, and the interaction of the gene and its encoded protein with its environment, will determine the outcome on the overall phenotype of the cell. Although most changes are minimal or muted towards the cell's overall fitness, some phenotypic changes can lead to a striking gain in the fitness of the cell within its microenvironment. These changes are known as driver mutations or epigenetic drivers that are important in understanding cancer biology.

For these genetic or epigenetic changes to translate into striking changes in the phenotype, there needs to be either a shift in gene expression or a shift in the ability of the translated protein to carry out its function (Mar, 2019). The latter may lead to disturbed or loss in function due to a truncated protein or alternative splicing. The greater the effect of the genetic or epigenetic change, the larger the separation that is expected between the average level of expression in the wild-type and mutated cells. Therefore, the discontinuity between the two populations (wildtype vs mutation) is presumed to be larger. Building on Mendelian genetics, even though multiple alleles can confer an overall phenotype that has a large effect, the inherent compound probability of requiring mutations occurring on a series of highly specific genes and the difficulty in studying through downstream analysis makes it less of an interest in this thesis.

Thus, our interest is to identify a single allele change that confers a large effect identifiable through expression analysis.

In Chapter 3, the GMMchi pipeline is established to provide a systematic method of identifying bimodal distributions in gene expression data. In this chapter, we will leverage GMMchi in identifying bimodal distributions in various in-house and public datasets on colorectal cancer tumors. In this chapter, we dive into the biological context of identified bimodal distributions, their association with known phenotypes of the cell lines, as well as exploring novel genetic or epigenetic regulations. We will also demonstrate GMMchi on applications such as clustering gene clusters and modeling tumor expression purification on multi-cell type samples commonly found in patient samples.

4.2 Results

4.2.1 2 by 2 Chi-Square Table

Three main datasets that are used in the downstream analysis include the microarray data on CRC cell lines produced in-house, the Cancer Genome Atlas (TCGA) RNA-seq dataset on colorectal cancer patient samples, and the RNA-seq data from the Cancer Cell Line Encyclopedia (CCLE) project. GMMchi is applied to each body of data to categorize each gene into groups labeled 1 as low expression group and 2 as high expression group. After transforming continuous expression values into categorical groups, each gene is labeled as 'bimodal', 'unimodal', 'categorical unimodal', 'unimodal with a tail', and 'bimodal with a tail' as part of the returning output of GMMchi. The goodness of fit of each gene is also documented.

In our in-house microarray data on CRC cell lines, a previous DPhil student Peter Kalugin and Jenny Wilding put together a list of phenotypes associated with each cell line. Note that 1 is equivalent to - and 2 is equivalent to +. This includes:

1) **RER status**: Mismatch repair deficiency indicates whether a cell line has a deficiency in its mismatch repair pathway.

    *1 = mismatch repair efficiency; 2 = mismatch repair deficiency

2) **Lumen formation**: Lumen formation is a phenotype that demonstrates the differentiation ability of the cell line. This is experimentally determined by doing lumen-forming assays.

    *1 = non-lumen forming; 2 = lumen-forming

3) **Goblet cells**: Some cell lines exhibit a subset of goblet cells that express MUC2, a secreted glycoprotein, and a biomarker for mucus-secreting goblet cells. This is assessed using antibody PR5D5, specific for MUC2.

   *1 = cell lines without goblet cells; 2 = cell lines with goblet cells

4) **Epithelial to Mesenchymal Transition (EMT):** The hallmark for EMT-like cell lines is the loss of E-cadherin (CDH1) expression due to EMT transcription factors (not CDH1 mutation) and high Vimentin (VIM) expression.

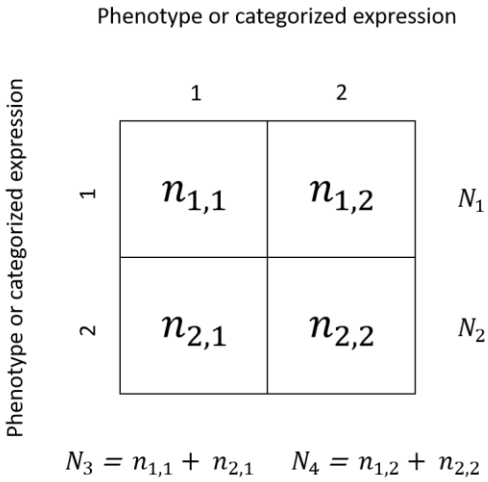
   *1 = non-EMT-like cell lines; 2 = EMT-like cell lines

5) Mutation data on known colorectal cancer driver mutations including APC, p53, CTNNB1, KRAS, BRAF, PIK3CA, FBXW7, TGFBR2, SMAD4

It is important to note that not every cell line has a phenotypic characterization associated with it. Thus, cell lines with a lacking phenotype for any given category listed above will have a value of N/A (Not available) and will thus be ignored during downstream analysis in pairwise comparison studies.

To study binary relationships between phenotypes and gene expression categories across the full panel of genes, we implemented the 2x2 table to study the statistical significance and association between two binary variables. The study will be performed by running 2x2 tables of individual genes that exhibited a bimodal distribution (discussed in Chapter 3) against chosen phenotype (EMT, lumen-forming …etc.). The 2x2 table is as shown in Figure 4.1 (Yates, 1934).

a) b)



$$r = \frac{n_{1,1} \times n_{2,2} - n_{1,2} \times n_{2,1}}{\sqrt{N_1 \times N_2 \times N_3 \times N_4}}$$

$$\chi^2 = \frac{k \times (|\, n_{1,1} \times n_{2,2} - n_{1,2} \times n_{2,1} \,| - \frac{k}{2})^2}{N_1 \times N_2 \times N_3 \times N_4}$$

$$k = N_1 + N_2 + N_3 + N_4$$

Figure 4.1 Formula for Calculating Chi-Square and R-value: (a) Shows a 2x2 table and the associated formula needed to calculate the r-value and $\boldsymbol{\chi^2}$ shown in (b). Yates correction for continuity used to account for smaller data using the 2x2 table is present in the $\boldsymbol{\chi^2}$ formula as $\frac{\boldsymbol{k}}{\boldsymbol{2}}$. If a negative value is returned following Yates correction, then a 0 is returned.

The output of the 2x2 table consists of 3 important returned outputs:

1) *r* coefficient: This value conveys the strength in the direction of association that ranges between -1 to 1. -1 = negative association; 1 = positive association.
2) *p*-value: The p-value is derived using Fisher's degree of freedom and Pearson's χ^2 test corrected for continuity by Yate's correction.
3) 2x2 table association pattern: By studying the pattern of the 2x2 table, certain association patterns can be identified such as inclusions and exclusions. Inclusion reveals cell lines that express gene A as an included subgroup of cell lines expressing gene B. This may lead to insights such as gene B being a precursor gene of gene A. We will use the formula $\frac{(a-b)^2}{a+b}$ compared to χ^2 critical value 3.84 derived with 1 degree of

freedom and a p-value of 0.05 to assess the inclusion/exclusion criterion. Essentially, we are measuring whether +/- and -/+ are skewed towards a particular direction in a positive association and whether +/+ and -/- are skewed towards a particular direction in a negative association.

The 2x2 table will be the basis of binary expression or phenotype analysis for many of this chapter's downstream analyses.

4.2.2 Probe Combining Pipeline

In microarray-based expression analysis, the expression level of a gene can be linked to multiple probe sets owing to microarray's probe-based method of detecting the level of a specific DNA or RNA fragment through groups of probe sets that identify a unique region of a gene (Liu, Bebu and Li, 2010). The probe is then mapped to a gene name using a predefined library consisting of a dictionary of probes to gene translation. Putting this together, each probe set is designed to pick up one message type that aims to identify only one gene based on the message sequence. The data generated by microarray where a gene name is associated with multiple probe sets is thus much trickier than RNA-sequencing where a gene name is only associated with a single probe set.

To address and facilitate analysis on a gene-by-gene basis, we developed a probe combining pipeline that aims to combine expression levels of multiple probe sets representing a gene into a single gene with a weighted value that encompasses all the combined probe sets. However, not all probe sets can be combined to represent a gene. When we identify probe sets that identify the same gene but do not correlate well with most of the other probe sets, we will

retain the probe set without combining it with the rest and assume it as a splice variant of the gene since we expect a splice variant to behave differently from each other.

As seen in Figure 4.2a, the pipeline starts with identifying genes with more than one probe set. The pattern of each probe set is predetermined by GMMchi. Bimodal probes consist of 1s and 2s with 1 representing low expressing cell lines and 2 representing high expressing cell lines. In unimodal probes, only probes that exhibit a categorical unimodal pattern are not excluded and consist of 1s and 2s with 1s representing lower expressing cell lines and 2 representing higher expressing cell lines. As mentioned in the last chapter, we are interested in identifying patterns that result in separation into two distinct groups.

The association between different probe sets from the same gene is determined by Pearson's r correlation. From the 2x2 table, probe sets with a correlation above 0.69 are grouped and weighted according to their chi-square goodness of fit value returned from the GMMchi output. The r cutoff value of 0.69 is determined by determining the cutoff of the bimodal distribution identified by plotting out all the r-values using GMMchi. The distribution is shown in Figure 4.2b. The formula used to arrive at the final weighted output of any given unique gene using the expression of multiple probe sets is shown below:

$Y_i = \frac{1}{{\chi_i}^2}$ is the inverse of the chi-square goodness of fit value of probe *i*

$Y_{total} = \sum_{i=0}^{k} Y_i$ is the sum of all inverse chi-square goodness of fit values for probe *i* for a given gene. *k* represents the number of probe sets that will be combined

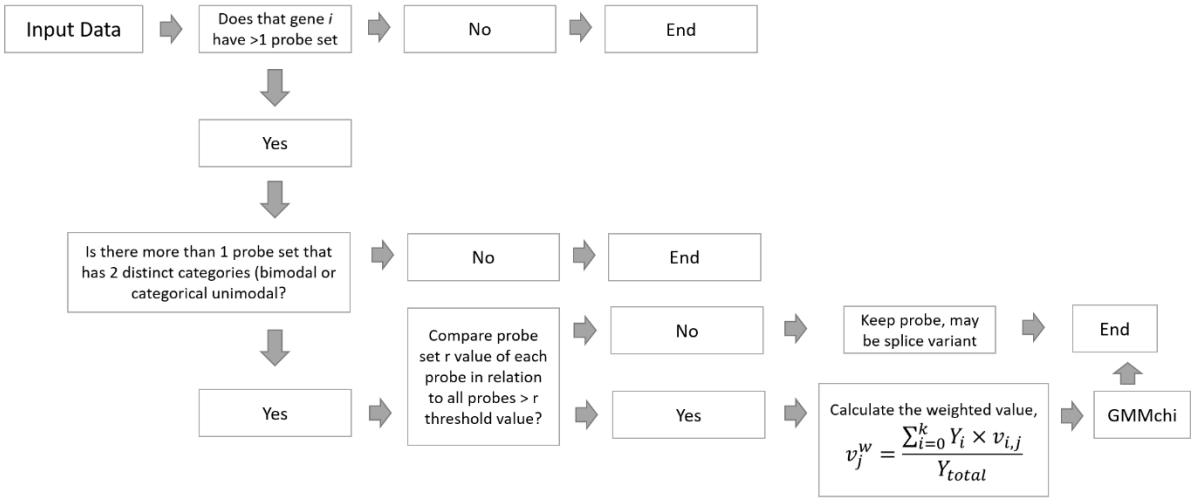
$v_{i,j}$ = expression value of probe *i* for a given cell line *j*

$$\boldsymbol{v_j^w = \frac{\sum_{i=0}^{k} Y_i \times v_{i,j}}{Y_{total}}} \quad \textbf{(5)}$$

The weighted value v_j^w shown in Equation (5) is the final output encompassing the weighted sum of each correlated probe. The inverse of the goodness of fit value Y_i is used as the weight for each probe *i* expression value. The reason for using the inverse of the goodness of fit is because the lower goodness of fit values the better the fit thus we want the smaller-value to weigh more heavily towards the overall outcome. Besides, if all the weights are the same, the sum of weighing by $\frac{1}{{\chi_i}^2}$ results in the arithmetic mean since it would equate to $\frac{\sum_{i=0}^{k} v_{i,j}}{\mathrm{k}}$ for all *i*, which is what is expected. The weighted value is then put through GMMchi to convert those values into groups of 1 (low expression) or 2 (high expression).

Lastly, probe sets that do not behave similarly with other probe sets (r-value < r threshold value) are treated as splice variants and kept as a separate probe set annotated with the gene plus the probe number (i.e., CDH1--2). The result of this pipeline will be the basis of analysis further downstream.

a)



b)

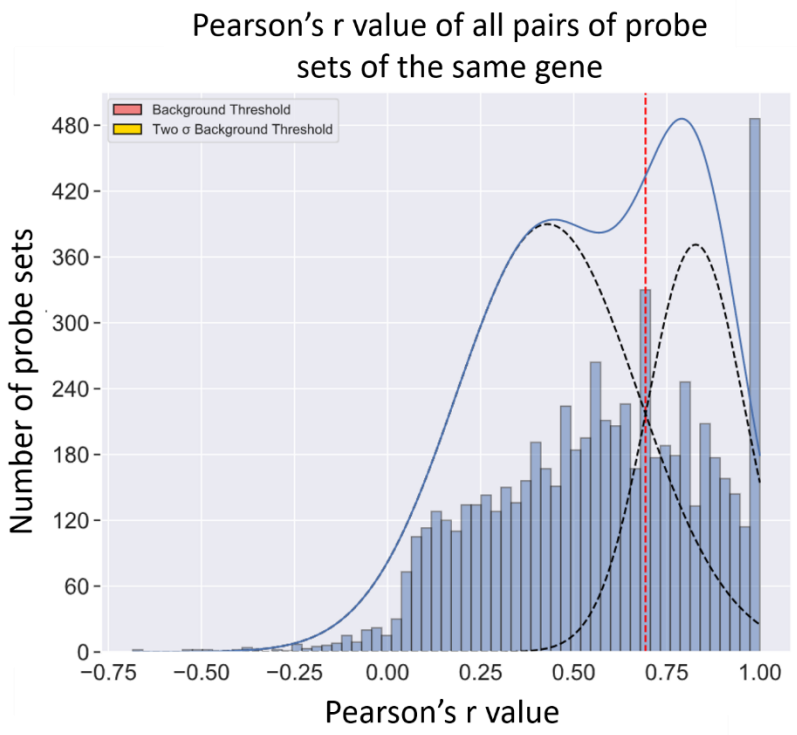


Figure 4.2 Schematic of the Probe Combining Pipeline: (a) The pipeline for probe combining is used to combine probe sets in microarray data that represent the same gene while keeping probe sets with a different expression pattern as a potential splice variant. (b) shows the total distribution of the r-values of all pairs of probe sets of the same gene. We apply GMMchi to the bimodal distribution to find the cutoff (red vertical line) that defines our r threshold value.

4.2.3 Phenotype Association

Over the years, our lab has accumulated rich phenotypic characterizations of our CRC cell lines. These phenotypes, as shown previously, include lumen-formation, EMT-like phenotype, RER status, and cell lines with goblet cell subpopulations. We will be using GMMchi-based analysis and 2x2 tables to understand and explore phenotypes and gene mutations somewhere associated with the phenotype. The built-in function **GMM.find_hits** in the ***GMMchisquare*** library finds statistically significant correlations using a 2x2 table analysis between every single gene against each phenotype. The outcome contains every gene with its respective r-value and p-value. We will also be using another built-in function **GMM.run_hits** to visualize the 2x2 table of genes that showed a statistically significant correlation with the phenotype. All p-values have been subjected to Bonferroni multiplicity testing. The resulting output is ranked based on the r-value sorted in the order of large positive correlations to negative values.

In addition, the result in the following analysis will be analyzed via a pattern-based analysis using the inclusion/exclusion criterion. The result for each phenotype will be presented in two large tables, one showing genes positively correlated and one showing genes negatively correlated. Within each large table, there will be 3 smaller tables containing genes without inclusions/exclusions and genes with inclusions/exclusions that are skewed towards either direction of the association. Another common observation is genes labeled as ‘nan’ or ‘LOC’ followed by a number. These are genes without a published symbol and orthologs not yet been determined.

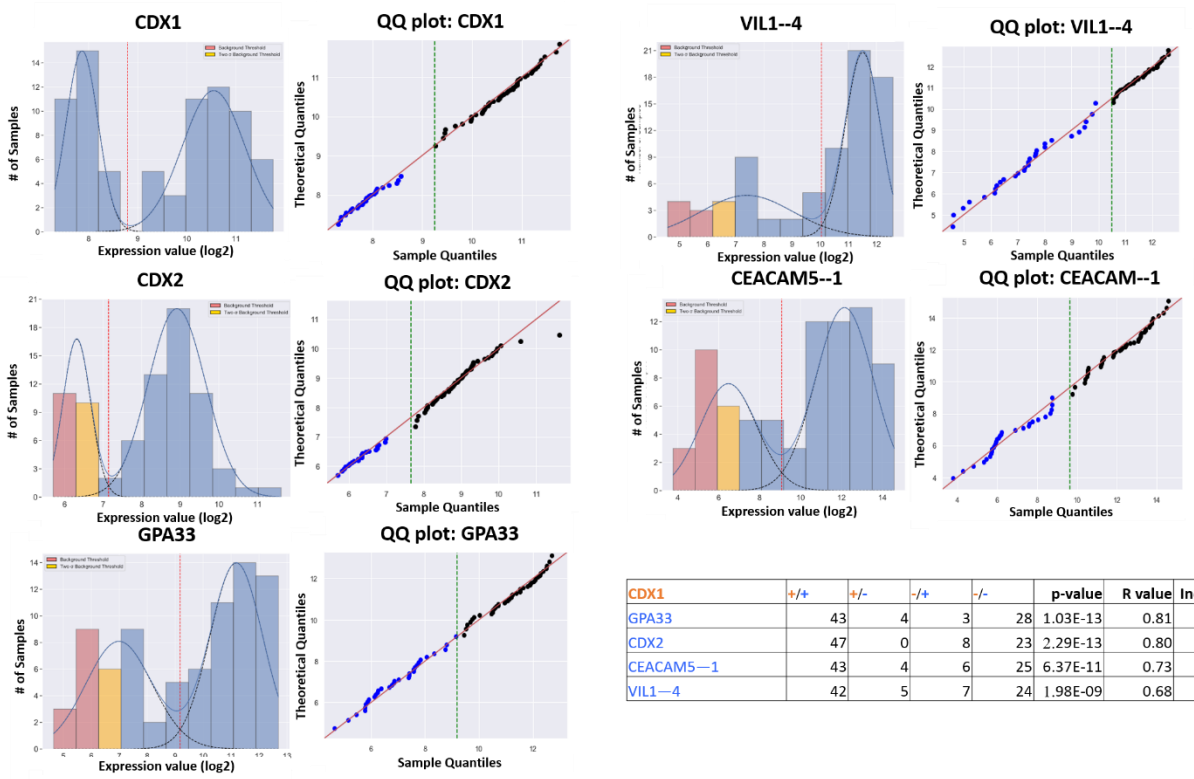
4.2.3.1 Lumen-forming CRC Cell Lines

Colorectal cancer is a mix of cell types consisting of different stages of differentiation. Even though tumors are known to have a higher percentage of the stem cell population, there are also different subpopulations of differentiated cell types. They include common differentiated colonic cell types such as enterocytes, goblet cells, and enteroendocrine cells. Here, we explore genes associated with lumen formation. The phenotype lumen formation is determined by previous work in our lab where a single cell from each CRC cell line grows in a 3D matrix while allowing it to differentiate and grow for 9 days before fixing and immunofluorescence staining. The stained samples are visually determined whether they form 3D lumen structures. Through this analysis, we expect genes related to differentiation to correlate significantly with lumen formation.

In total there are 42 cell lines with the phenotype lumen formation assessed band so this analysis will only be based on those cell lines. Our returning output from **GMM.find_hits** (our built-in function to run 2x2 table associations) is a dataframe consisting of genes that are significantly correlated with the phenotype at a p-value cutoff of 0.05 after Bonferroni multiplicity correction. We are returned with 184 genes that are significantly correlated with lumen formation.

It was previously shown that caudal-related protein 1 (CDX1) is the primary transcription factor responsible for enterocyte differentiation as well as its family member CDX2 (Ashley, Yeung and Bodmer, 2013). As shown in Figure 4.3, CDX1 and CDX2 are both significantly correlated with lumen formation. Looking closer, CDX1 and CDX2 exhibit an inclusion whereby cell lines with high CDX1 expression have high CDX2 expression whereas cell lines with high CDX2

expression do not necessarily have high CDX1 expression. Although in the literature, CDX1 and CDX2 are suggested to have an overlapping function, this observation may suggest that CDX1 plays a more specific role or that CDX2 is a precursor gene that is expressed in a broader population of cells before CDX1 expression. Furthermore, our lab has previously shown that CDX1 is sufficient for stimulating lumen formation in CDX1 low expressing cell lines. To show that the differentiation markers such as carcinoembryonic antigen-related cell adhesion molecule 5 (CEACAM5), villin 1 (VIL1), and glycoprotein A33 (GPA33) reveal a pattern of significant co-upregulation with CDX1, we ran another 2x2 table among the genes shown in Figure 4.3.



CDX1	+/+	+/-	-/+	-/-	p-value	R value	Inclusion
GPA33	43	4	3	28	1.03E-13	0.81	No
CDX2	47	0	8	23	2.29E-13	0.80	Yes
CEACAM5—1	43	4	6	25	6.37E-11	0.73	No
VIL1—4	42	5	7	24	1.98E-09	0.68	No

Figure 4.3 Examples of GMMchi Analysis on Known Differentiation Markers: Here we shown examples of known differentiation markers that are positively and significantly associated with CDX1, a biomarker for colonic epithelial differentiation identified by phenotype association with lumen formation. All genes exhibited a well-defined bimodal distribution consisting of two subpopulations. The R-value shows that CDX1 is strongly correlated with GPA33, a basolateral intestinal differentiation marker, as well as CEACAM5 and VIL1 which are all shown to be epithelial differentiation markers in the colon.

In our association study, we always apply the Bonferroni correction for multiple comparisons. Based on prior expectation, Bonferroni correction is a very stringent correction so any associations significant after correction are highly likely to be biologically significant. First, we noticed the gene SLC39A5 from the solute carrier family in Table 4.1a. SLC39A5, also known as ZIP5, is known to regulate zinc excretion and is found abundant at the basolateral surface in intestinal enterocytes (Geiser, De Lisle and Andrews, 2013). Other differentiation markers include Propsero homebox 1 (PROX1) (Conacci-Sorrell *et al.*, 2002) and cystic fibrosis transmembrane conductance regulator (CFTR). CFTR is shown to have an average 10-fold increase in expression

levels in differentiated colonic epithelium compared to undifferentiated cells (Sood *et al.*, 1992). As expected, many of the genes that correlate with lumen formation are markers found expressed in colonic epithelial differentiation. One thing we noticed is the lack of VIL1 and GPA33 despite their strong association with CDX1. We revisited their associations with lumen formation and found that they are on the verge of being significantly associated with lumen formation but are excluded after correction for multiplicity testing.

On the other hand, genes that fit the inclusion criterion skewed towards the +/- reveal a pattern that is more to just being positively associated with lumen formation. In Table 4.1b, we observed that non-lumen forming cell lines (-/+) rarely express any of the genes with at most 1 exception. Nonetheless, lumen-forming cell lines may or may not express these genes. This pattern allows us to draw the inference that these sets of genes are only expressed in lumen-forming cell lines but are not required genes for lumen formation or may play a replaceable role in lumen formation where other genes can compensate for their functions. For example, neuron-glia-related cell adhesion molecule (NRCAM) is shown to be a differentiation marker for neural development. It is most likely NRCAM is a nonessential gene for lumen formation but could potentially be expressed in lumen-forming cells due to differentiation (Conacci-Sorrell *et al.*, 2002). Moreover, we also observe distinct sets of genes that exhibit the same pattern. The observation of cell lines expressing distinct sets of genes suggests networks of related genes that work together to affect lumen formation.

In Table 4.1c, genes that fit the inclusion criterion skewing towards the opposite direction (-/+) revealed an alternative pattern where lumen-forming cell lines almost always express these sets of genes. It is important to note that non-lumen-forming cell lines could also express these

sets of genes by looking at (-/+). The pattern suggests that these genes may act as precursors genes for lumen formation or genes that regulate a broader pathway required for lumen formation such that when cell lines form lumens, they will almost always express these genes. For example, nephronectin (NPNT), serine/threonine/tyrosine kinase I (STYK1), and nuclear factor I A (NFIA) are all thought to be important regulators of cellular differentiation in epithelial cells and may precede lumen formation (Aleksandrova *et al.*, 2015; Hu *et al.*, 2015; Chen *et al.*, 2020).

We looked at genes that are inversely correlated with lumen formation (Table 4.1d) including IGFBP6 and TRIB2. Insulin-like growth factor-binding protein 6 (IGFBP6) is a glycoprotein that selectively binds to IGF-II, which is responsible for cellular proliferation and differentiation. It has been suggested IGFBP6 is observed to decrease in differentiated cells (Bach, 2015). Tribbles Pseudokinase 2 (TRIB2) is a gene that is shown to be important in maintaining pluripotency in mouse embryonic stem (ES) cells. *Kim et al.* have shown the decrease in TRIB2 as ES cells differentiate and the knockdown of TRIB2 decreased pluripotency markers Oct4, SOX2, Nanog, and KLF4 (Do *et al.*, 2017). These genes may suggest that stemness-related genes are most inversely correlated with lumen formation as opposed to differentiation markers that showed a positive correlation.

Different from the analysis of inclusion criteria shown previously, exclusion criteria for negative association looks at the patterns of -/- and +/+ instead of -/+ and +/- used in positive associations. This is because, in positive associations, most observed samples fall into the -/- and +/+ categories while in negative associations, most observed samples fall into the +/- and -/+ categories. Table 4.1e revealed a pattern where almost no lumen-forming cell lines expressed

any of the genes listed in the table. This pattern suggests that the list of genes may inhibit lumen formation or be predominantly associated with dividing cells. It is also important to note that we, again, observe multiple sets of genes that exhibit identical patterns of expression, suggesting a larger network or pathway working against lumen formation, or also positively associated with dividing cells.

Table 4.1 Genes Associated with Lumen Formation: (a) Genes positively correlated with lumen formation that do not fall into the inclusion criterion. (b) and (c) show genes positively correlated with lumen formation that are inclusions skewed towards +/- and -/+, respectively. (d) Genes negatively correlated with lumen formation that does not fall into the exclusion criterion. (e) and (f) show genes negatively correlated with lumen formation that are exclusions skewed towards -/- and +/+, respectively.

POSITIVE ASSOCIATIONS WITH LUMEN FORMATION

a) Association with lumen formation without Inclusions

Lumen Formation	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
TINAG--2	18	4	4	16	1.34E-04	0.62	No
nan--3854	18	4	6	14	1.46E-03	0.52	No
TST	15	7	4	16	2.32E-03	0.48	No
SLC39A5--1	16	6	5	15	4.80E-03	0.48	No
AIFM3	18	4	7	13	4.04E-03	0.48	No
DPEP1	17	5	6	14	4.61E-03	0.47	No
PROX1--2	13	9	3	17	4.56E-03	0.45	No
AGXT2L2--2	14	8	4	16	5.84E-03	0.44	No
CDX1	19	3	9	11	8.00E-03	0.44	No
CFTR--2	15	7	5	15	6.78E-03	0.43	No
C13orf18_weighted	15	7	5	15	6.78E-03	0.43	No
ATP10B--1	18	4	8	12	1.01E-02	0.43	No
BCL11A--3	18	4	8	12	1.01E-02	0.43	No
DDC_weighted	19	3	10	10	1.86E-02	0.39	No
SGK2--1	19	3	10	10	1.86E-02	0.39	No
TMPRSS2--4	19	3	10	10	1.86E-02	0.39	No
nan--11030	14	8	5	15	1.57E-02	0.39	No
nan--3823	14	8	5	15	1.57E-02	0.39	No
nan--9357	14	8	5	15	1.57E-02	0.39	No
CCL14-CCL15 /// CCL15	14	8	5	15	1.57E-02	0.39	No
CLDN3_weighted	18	4	9	11	2.31E-02	0.38	No
CEACAM5--1	18	4	9	11	2.31E-02	0.38	No

CYP2B6 /// CYP2B7P1	18	4	9	11	2.31E-02	0.38	No
GSPT2	17	5	8	12	2.65E-02	0.38	No
C20orf112--1	17	5	8	12	2.65E-02	0.38	No
TLN2_weighted	17	5	8	12	2.65E-02	0.38	No
SERPINI1	12	10	4	16	2.89E-02	0.36	No

b) Association with lumen formation with Inclusions (skewed to +/-)

Lumen Formation	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
NRCAM--1	12	10	1	19	6.38E-04	0.54	Significant
MME_weighted	10	12	1	19	4.17E-03	0.46	Significant
SLC6A4--1	7	15	0	20	9.19E-03	0.43	Significant
LARS--1	9	13	1	19	9.84E-03	0.42	Significant
INPP5D--2	9	13	1	19	9.84E-03	0.42	Significant
LOC339290_weighted	9	13	1	19	9.84E-03	0.42	Significant
CAB39L--1	9	13	1	19	9.84E-03	0.42	Significant
PROX1_weighted	11	11	3	17	2.32E-02	0.37	Significant
LOC285178	6	16	0	20	2.16E-02	0.39	Significant
DLX1--2	6	16	0	20	2.16E-02	0.39	Significant
nan--8245	6	16	0	20	2.16E-02	0.39	Significant
HPS4—3	6	16	0	20	2.16E-02	0.39	Significant
nan--8323	8	14	1	19	2.21E-02	0.38	Significant
TMPRSS13--1	8	14	1	19	2.21E-02	0.38	Significant
GNA14--1	8	14	1	19	2.21E-02	0.38	Significant

c) Association with lumen formation with Inclusions (skewed to -/+)

Lumen Formation	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
NPNT--1	22	0	12	8	1.07E-03	0.51	Significant
FAR2--2	21	1	12	8	7.71E-03	0.43	Significant
STYK1_weighted	21	1	12	8	7.71E-03	0.43	Significant
CDX2	20	2	11	9	1.32E-02	0.41	Significant
TUSC1	21	1	13	7	1.82E-02	0.39	Significant
QPRT--1	20	2	12	8	2.96E-02	0.36	Significant
NFIA--4	20	2	12	8	2.96E-02	0.36	Significant

NEGATIVE ASSOCIATIONS WITH LUMEN FORMATION

d) Association with lumen formation without Exclusions

Lumen Formation	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
IGFBP6	4	18	13	7	4.04E-03	-0.48	No
TFAP2A--1	9	13	17	3	4.56E-03	-0.45	No
TRIB2_weighted	3	19	11	9	8.00E-03	-0.44	No
BAG4_weighted	3	19	11	9	8.00E-03	-0.44	No
MOSC2_weighted	5	17	13	7	1.16E-02	-0.43	No
C9orf125--2	10	12	17	3	1.07E-02	-0.41	No
RASSF5--2	3	19	10	10	1.86E-02	-0.39	No
CSGALNACT2_weighted	3	19	10	10	1.86E-02	-0.39	No
FZD4--1	3	19	10	10	1.86E-02	-0.39	No
ZBED2	3	19	10	10	1.86E-02	-0.39	No
ALDH1A3	8	14	15	5	1.57E-02	-0.39	No
IFI6	7	15	14	6	2.94E-02	-0.38	No
IDS--6	7	15	14	6	2.94E-02	-0.38	No
RARRES1_weighted	5	17	12	8	2.65E-02	-0.38	No
CAV2_weighted	5	17	12	8	2.65E-02	-0.38	No
GPX8--1	5	17	12	8	2.65E-02	-0.38	No
FRAS1--4	6	16	13	7	2.86E-02	-0.38	No
PPP1R9A--2	6	16	13	7	2.86E-02	-0.38	No

e) Association with lumen formation with Exclusions (skewed to -/-)

Lumen Formation	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
PRKCDBP	0	22	7	13	2.87E-03	-0.47	Significant
PHLDB2--5	1	21	8	12	7.71E-03	-0.43	Significant
SAA1 /// SAA2_weighted	0	22	6	14	7.39E-03	-0.43	Significant
LIX1L--1	2	20	9	11	1.32E-02	-0.41	Significant
ULBP2_weighted	2	20	9	11	1.32E-02	-0.41	Significant
MT1E	2	20	9	11	1.32E-02	-0.41	Significant
LOXL4--2	1	21	7	13	1.82E-02	-0.39	Significant
ZNF193_weighted	1	21	7	13	1.82E-02	-0.39	Significant
LTBP3--2	1	21	7	13	1.82E-02	-0.39	Significant
SAMD9L--2	0	22	5	15	1.82E-02	-0.39	Significant
TSPAN14--1	0	22	5	15	1.82E-02	-0.39	Significant
ZDHHC8P1	0	22	5	15	1.82E-02	-0.39	Significant
SMARCD3--1	0	22	5	15	1.82E-02	-0.39	Significant
SPON1—3	0	22	5	15	1.82E-02	-0.39	Significant
TMEM45A	0	22	5	15	1.82E-02	-0.39	Significant
CDK14—1	0	22	5	15	1.82E-02	-0.39	Significant
FGF9--1	2	20	8	12	2.96E-02	-0.36	Significant
TTC7B	2	20	8	12	2.96E-02	-0.36	Significant
LOXL1	2	20	8	12	2.96E-02	-0.36	Significant
TUBB2A /// TUBB2B	2	20	8	12	2.96E-02	-0.36	Significant

L1CAM--1	2	20	8	12	2.96E-02	-0.36	Significant

f) Association with lumen formation with Exclusions (skewed to +/+)

Lumen Formation	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
MPZL2_weighted	14	8	20	0	3.78E-03	-0.46	Significant
NMU	15	7	20	0	9.19E-03	-0.43	Significant
ALDH5A1--1	13	9	19	1	9.84E-03	-0.42	Significant
C5orf35--1	12	10	18	2	1.67E-02	-0.39	Significant
HOXA10 /// HOXA9_weighted	14	8	19	1	2.21E-02	-0.38	Significant
KIF13A—5	14	8	19	1	2.21E-02	-0.38	Significant
PECI	11	11	17	3	2.32E-02	-0.37	Significant
CEP55	13	9	18	2	3.53E-02	-0.35	Significant
TMEM56--1	13	9	18	2	3.53E-02	-0.35	Significant
BACE2_weighted	17	5	20	0	4.92E-02	-0.35	Significant

4.2.3.2 PR5D5-expressing Goblet Cell Subpopulation in CRC Cell Lines

As already mentioned, goblet cells are one of the three differentiated cell types and are responsible for the production of mucin 2, the major mucin that lines the intestine. The phenotype for goblet cell-presenting cell lines is determined previously in our lab by PR5D5 immunofluorescence staining, an in-house antibody that targets goblet cell marker Mucin 2 (MUC2). In total, we had 30 cell lines with the phenotype for goblet cell subpopulation assessed.

We explored the returned output in Table 4.2. As expected in Table 4.2b, MUC2 is strongly associated with the phenotype. Other genes that are known to be related to goblet cells include SAM pointed domain-containing ETS transcription factor (SPDEF) and Fc- γ binding protein (FCGBP). SPDEF is thought to govern the differentiation of goblet cells in intestinal epithelium thus is expected to correlate with goblet cell formation (Noah *et al.*, 2010). FCGBP is known to crosslink with MUC2 and the trefoil factor peptides (TFF) family to form highly viscous mucus that is secreted by goblet cells to line the outer intestinal layer (Kim and Ho, 2010). Down in the list,

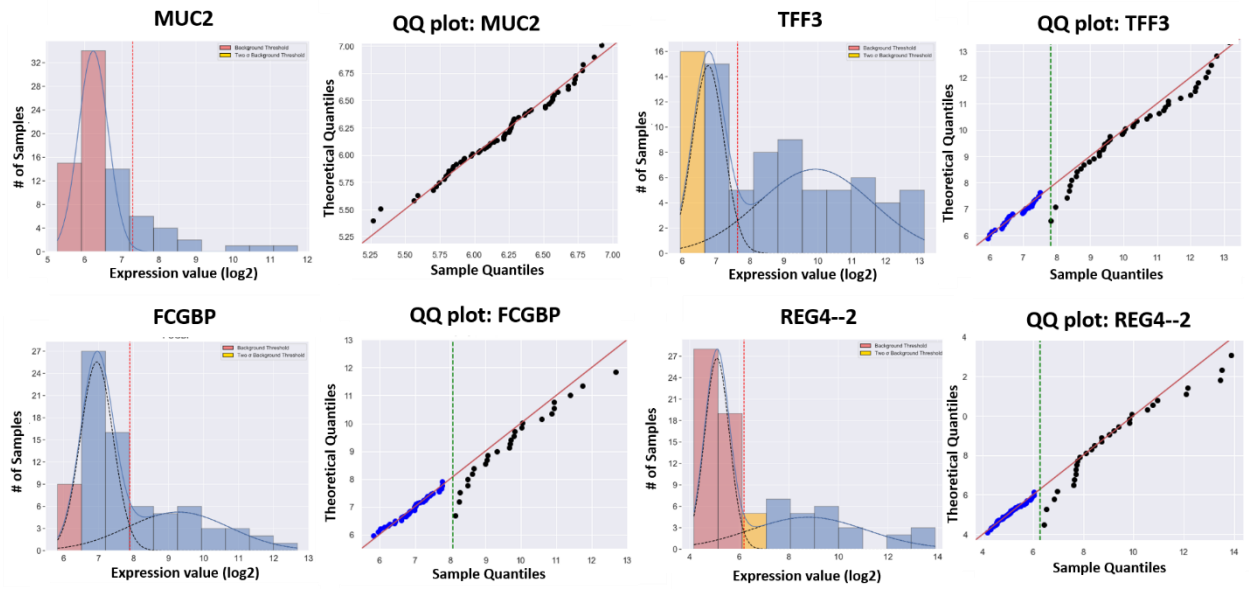
we also found TFF3 highly correlated with goblet-cell forming cell lines, further reinforcing the link between the three genes. Here, we observed that MUC2, SPDEF, and FCGBP all fell under the list of genes that fit the inclusion criterion. This observation of the inclusion of MUC2 mRNA levels even though the phenotype was determined using an antibody specifically against MUC2 can be explained by the sensitivity of microarray-based expression quantification. During IF staining of MUC2, we know that only > 3% of the cells are stained with MUC2 expression in goblet cell-forming cell lines. This means that the low proportions detectable by IF staining are too low to show up as increased mRNA levels in the microarray data. This matches with MUC2 being an inclusion skewed towards +/- where cell lines that are not stained positive with MUC2 antibody do not have MUC2 expression while cell lines stained with MUC2 antibody may or may not have increased mRNA owing to the low proportion of MUC2 expressing cells. This is an extremely important observation to note when dealing with small subsets of cells within a cell line that expresses a differentially expressed gene.

Another known goblet cell-associated gene that does not fit the inclusion criterion is REG4, a secreted protein found to associate closely with secretory cells as well as playing an important role in intestinal morphogenesis (Sasaki *et al.*, 2016). Another very important observation is the upregulation of UGT and FUT-related genes, which encode UDP glycosyltransferase and fructosyltransferase, respectively (Bergstrom and Xia, 2013). These genes are critical in glycosylating MUC2, a glycoprotein, before it being secreted into the cytoplasm from the Golgi apparatus. We also found CDX1, CDX2, and CFTR correlated with goblet cell formation in CRC cell lines.

On the other hand, AGR3, a family member of AGR2, is the most correlated gene that is also an inclusion skewed towards -/+ (Table 4.2c). AGR2 has been shown to form mixed disulfide bonds with MUC2 and is essential for MUC2 maturation (Park *et al.*, 2009). AGR3 being more positively correlated with goblet cells may suggest its role in MUC2 production similar to the role of AGR2. As a member of the disulfide isomerase (PDI) family located in the endoplasmic reticulum (ER), the AGR family had been thought to mainly play the role of catalyzing disulfide bond formation in protein folding. This pattern further reinforces this observation where goblet cell forming cell lines always expressed AGR3 while AGR3+ cell lines did not always form goblet cells since AGR plays a more general role in protein folding unlike MUC2, SPDEF, or FCGBP that exhibited the opposite pattern and played a more specific role for mucin production and maturation in goblet cells.

To further understand the interplay between identified genes related to goblet cell formation, we further performed a 2x2 table analysis between genes known to play a role in goblet cell differentiation and mucus production. We also noticed that TFF3 fit the inclusion criterion in the 2x2 table against MUC2. The inclusion criterion suggests that 13/15 (86.7%) of MUC2 expressing cell lines express TFF3 while 2/34 (0.6%) of TFF3 expressing cell lines do not express MUC2. Similar to what was discussed above about inclusion criteria, we suspect that MUC2+ expression is a subpopulation under TFF3+ expression which may further suggest TFF3 is a required protein before MUC2 maturation and secretion. A similar inclusion is also observed in FCGBP and REG4 where high MUC2 expressing cells nearly always express FCGBP or REG4 but FCGBP or REG4 expressing cells do not always express MUC2. This would, again, suggest that FCGBP and REG4 may be precursors of MUC2. This potentially suggests a new category of partially

differentiated goblet cells that expresses TFF3, FCGBP, and REG4 required prior to MUC2 maturation and subsequent secretion.



MUC2	+/+	+/-	-/+	-/-	p-value	R value	Inclusion
FCGBP	13	2	13	50	3.27E-06	0.55	Yes
REG4—2	14	1	12	51	1.17E-07	0.62	Yes
TFF3	13	2	31	32	9.49E-03	0.29	Yes

Figure 4.4 GMMchi Analysis on Genes Associated with Goblet Cells: Examples of the inter-correlations between goblet cell markers identified through genes that are significantly and positively correlated with the goblet-cell forming phenotype. As shown, MUC2 has a significant association with FCGBP as well as REG4, both known to be important goblet cell markers. Interestingly, TFF3 exhibits an inclusion pattern with most MUC2+ cells expressing TFF3 while TFF3+ cells do not necessarily express MUC2. Here, in the QQ plot of TFF3, REG4, and FCGBP, we see multiple data points deviating from the 45-degree line near the cutoff threshold (green dashed line). This is mainly because of the high overlap between the two normal distributions making the membership of the data points within the overlap harder to assign thus resulting in a poorer separation between the two distributions.

Table 4.2 Genes Associated with Goblet-Cell Formation: (a) Genes positively correlated with goblet-cell formation that do not fall into the inclusion criterion. (b) and (c) show genes positively correlated with goblet-cell formation that are inclusions skewed towards +/- and -/+, respectively. (d) Genes negatively correlated with goblet-cell formation that does not fall into the exclusion criterion. (e) and (f) show genes negatively correlated with goblet-cell formation that are exclusions skewed towards -/- and +/+, respectively.

POSITIVE ASSOCIATIONS WITH GOBLET-CELL FORMATION

a) Association with goblet-cell formation without Inclusions

Goblet cell-forming	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
RSPH1	9	2	1	18	3.51E-05	0.78	No
HNF4G—2	10	1	3	16	9.04E-05	0.73	No
TNFRSF11A_weighted	9	2	2	17	1.76E-04	0.71	No
REG4_weighted	8	3	1	18	2.23E-04	0.71	No
ST6GALNAC1	10	1	4	15	3.80E-04	0.67	No
FOXA2_weighted	10	1	4	15	3.80E-04	0.67	No
UGT1A1 /// UGT1A10 /// UGT1A4 /// UGT1A6 /// UGT1A8 /// UGT1A9_weighted	9	2	3	16	1.22E-03	0.65	No
UGT1A1 /// UGT1A10 /// UGT1A3 /// UGT1A4 /// UGT1A5 /// UGT1A6 /// UGT1A7 /// UGT1A8 /// UGT1A9_weighted	9	2	3	16	1.22E-03	0.65	No
TLN2_weighted	10	1	5	14	1.70E-03	0.62	No
CYP3A5_weighted	10	1	5	14	1.70E-03	0.62	No
CDX1	10	1	5	14	1.70E-03	0.62	No
MECOM_weighted	10	1	5	14	1.70E-03	0.62	No
PAK1_weighted	9	2	4	15	2.10E-03	0.59	No
C2orf82_weighted	9	2	4	15	2.10E-03	0.59	No
TRIM31_weighted	8	3	3	16	4.49E-03	0.57	No
AHCYL2_weighted	8	3	3	16	4.49E-03	0.57	No

b) Association with goblet-cell formation with Inclusions (skewed to +/-)

Goblet cell-forming	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
MUC2	8	3	0	19	2.82E-05	0.79	Significant
FCGBP	8	3	0	19	2.82E-05	0.79	Significant
ENTPD5--2	6	5	0	19	7.78E-04	0.66	Significant
CNTNAP2—2	6	5	0	19	7.78E-04	0.66	Significant
SPDEF_weighted	5	6	0	19	3.24E-03	0.59	Significant
nan--5067	5	6	0	19	3.24E-03	0.59	Significant

C11orf53	5	6	0	19	3.24E-03	0.59	Significant

c) Association with goblet-cell formation with Inclusions (skewed to -/+)

Goblet cell-forming	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
AGR3	11	0	4	15	5.00E-05	0.76	Significant
AKR1B10	11	0	5	14	8.66E-05	0.71	Significant
C2orf89	11	0	5	14	8.66E-05	0.71	Significant
CYP2B6 /// CYP2B7P1	11	0	6	13	3.17E-04	0.67	Significant
NOSTRIN	11	0	6	13	3.17E-04	0.67	Significant
DDC_weighted	11	0	6	13	3.17E-04	0.67	Significant
CFTR--1	11	0	6	13	3.17E-04	0.67	Significant
ATP8B1--2	11	0	6	13	3.17E-04	0.67	Significant
HEPH—2	11	0	6	13	3.17E-04	0.67	Significant
CEACAM6_weighted	11	0	7	12	6.05E-04	0.62	Significant
XK	11	0	7	12	6.05E-04	0.62	Significant
SGK2--1	11	0	7	12	6.05E-04	0.62	Significant
PPP1R1B	11	0	7	12	6.05E-04	0.62	Significant
AUTS2--1	11	0	8	11	1.56E-03	0.58	Significant
LOC100507192	11	0	8	11	1.56E-03	0.58	Significant
CBLC--1	11	0	8	11	1.56E-03	0.58	Significant
AGR2_weighted	11	0	8	11	1.56E-03	0.58	Significant
FA2H_weighted	11	0	8	11	1.56E-03	0.58	Significant
DNAJC22_weighted	11	0	8	11	1.56E-03	0.58	Significant
CLRN3	11	0	8	11	1.56E-03	0.58	Significant
KBTBD11	10	1	6	13	2.43E-03	0.57	Significant
SAMD5--2	10	1	6	13	2.43E-03	0.57	Significant
AIFM3	10	1	6	13	2.43E-03	0.57	Significant
CA12_weighted	10	1	6	13	2.43E-03	0.57	Significant
FAM108C1	10	1	6	13	2.43E-03	0.57	Significant
ATP10B--1	10	1	6	13	2.43E-03	0.57	Significant

NEGATIVE ASSOCIATIONS WITH GOBLET-CELL FORMATION

d) Association with goblet-cell formation without Exclusions

Goblet cell-forming	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
IDS--6	2	9	17	2	1.76E-04	-0.71	No
ANXA6	4	7	18	1	1.10E-03	-0.64	No
AKAP12--2	1	10	14	5	1.70E-03	-0.62	No
ETS1--3	2	9	15	4	2.10E-03	-0.59	No
TIMP2--3	3	8	16	3	4.49E-03	-0.57	No
SPAST--1	3	8	16	3	4.49E-03	-0.57	No
CDC42EP3_weighted	3	8	16	3	4.49E-03	-0.57	No
CAV2_weighted	3	8	15	4	8.61E-03	-0.51	No
nan--5064	3	8	15	4	8.61E-03	-0.51	No
CAV1_weighted	2	9	13	6	2.09E-02	-0.48	No
CALD1_weighted	2	9	13	6	2.09E-02	-0.48	No
CAP2_weighted	2	9	13	6	2.09E-02	-0.48	No
GLIS2	5	6	17	2	2.75E-02	-0.48	No

e) Association with goblet-cell formation with Exclusions (skewed to -/-)

Goblet cell-forming	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
FRMD6_weighted	0	11	13	6	3.17E-04	-0.67	Significant
CSGALNACT2_weighted	0	11	12	7	6.05E-04	-0.62	Significant
SMURF2_weighted	0	11	12	7	6.05E-04	-0.62	Significant
AMOTL1--1	0	11	10	9	4.05E-03	-0.54	Significant
UBASH3B_weighted	0	11	10	9	4.05E-03	-0.54	Significant
CYR61_weighted	1	10	12	7	6.73E-03	-0.53	Significant
GPC1--1	1	10	12	7	6.73E-03	-0.53	Significant
SRPX	1	10	12	7	6.73E-03	-0.53	Significant
WWTR1_weighted	1	10	12	7	6.73E-03	-0.53	Significant
DCBLD2--1	1	10	12	7	6.73E-03	-0.53	Significant
CALU--3	1	10	12	7	6.73E-03	-0.53	Significant
AP1S2--2	0	11	9	10	1.06E-02	-0.50	Significant
ATP10D	0	11	9	10	1.06E-02	-0.50	Significant
SLC2A3_weighted	0	11	9	10	1.06E-02	-0.50	Significant
PI4K2A--2	0	11	9	10	1.06E-02	-0.50	Significant
ZBED2	0	11	9	10	1.06E-02	-0.50	Significant
AP1S2--4	0	11	9	10	1.06E-02	-0.50	Significant
HERC5	0	11	9	10	1.06E-02	-0.50	Significant
PLAU_weighted	0	11	9	10	1.06E-02	-0.50	Significant
APOBEC3C	0	11	9	10	1.06E-02	-0.50	Significant
IQCJ-SCHIP1 /// SCHIP1	0	11	9	10	1.06E-02	-0.50	Significant
OBSL1--2	1	10	11	8	1.82E-02	-0.48	Significant
SCARA3_weighted	1	10	11	8	1.82E-02	-0.48	Significant
OBSL1--1	1	10	11	8	1.82E-02	-0.48	Significant
DSTYK_weighted	0	11	8	11	1.40E-02	-0.46	Significant

GADD45B--2	0	11	8	11	1.40E-02	-0.46	Significant
TMEM136--1	0	11	8	11	1.40E-02	-0.46	Significant

f) Association with goblet-cell formation with Exclusions (skewed to +/+)

Goblet cell-forming	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
HSPA8_weighted	6	5	19	0	3.24E-03	-0.59	Significant
FAM171B--1	1	10	13	6	2.43E-03	-0.57	Significant
STXBP1	7	4	19	0	1.20E-02	-0.52	Significant
SCAMP1_weighted	7	4	19	0	1.20E-02	-0.52	Significant
BAT3--2	7	4	19	0	1.20E-02	-0.52	Significant
PRKD3_weighted	6	5	18	1	1.56E-02	-0.48	Significant
TBC1D1--3	6	5	18	1	1.56E-02	-0.48	Significant
IFITM3	6	5	18	1	1.56E-02	-0.48	Significant
GNPDA2	6	5	18	1	1.56E-02	-0.48	Significant

4.2.3.3 Epithelial to Mesenchymal Transition-like CRC Cell Lines

Epithelial to mesenchymal transition (EMT) is a process where epithelial cells undergo changes that results in a mesenchymal-like phenotype such as motility, invasiveness, and resistance to apoptosis. The hallmark of the EMT phenotype is high vimentin (VIM) expressing but low E-cadherin (CDH1) in cancer thus we look for the same markers in CRC cell lines. It is important to note that the low CDH1 expression in EMT phenotype does not originate from the loss of expression due to a mutation in the CDH1 gene. It is mostly due to EMT transcription factors (TFs) that bind to the promoter region of CDH1, therefore, inhibiting CDH1 expression. This distinction is critical since EMT TFs are capable of translocating into the nucleus and drive the expression of other genes that are critical to the characteristic of a mesenchymal cell.

In Table 4.4a, two of the zinc finger E-box binding homebox 1 (ZEB1) probe sets are two of the most positively correlated genes with EMT. ZEB1 is an EMT TF under the ZEB family that binds to the TATA-box located on the promoter region of CDH1. The most intriguing part is that

no other EMT TFs are identified within the 800 genes returned that are significantly correlated with the EMT phenotype. In the literature, well-known EMT TFs including the ZEB family, snail family zinc finger (SNAI) family, and twist-related protein (TWIST) family are shown to co-upregulate or coordinate to induce EMT (Ansieau, Collin and Hill, 2014). Yet in our analysis, ZEB1 is not only the only EMT TF candidate but also the most positively correlated gene, making it an interesting gene that is hypothesized to be a colorectal cancer-specific EMT TF. We further followed up this observation and discuss our experimental results in Chapter 5. Here, we found that there are 3 probe sets for ZEB1 including probe set 1, probe set 2, and probe set 3. Between the three of them, there are slight differences in expression patterns that are cell line-specific. In Table 4.3a below, we show that some of the probe sets disagreed with others on whether a cell line is categorized as in a high or low expression group. We also compared the 3 probe sets in duplicate cell lines to look at the consistency of categorization in Table 4.3b. Based on the table, ZEB1—1 and ZEB1—3 are the most consistent probe sets while ZEB1—2, for example, exhibiting differences between SW480 and SW620. Interestingly, in Table 4.3c, we show that all of the cell lines listed below are classified as ZEB1 high based on the threshold determined by GMMchi; however, some are deemed as ZEB1 low due to the background expression threshold colored in red and yellow. Due to the threshold of our probe combining protocols, all three probe sets are deemed different enough to be classified as splice variants. Nonetheless, ZEB1—3 and ZEB1—1, despite the differences, correlated strongly with the EMT phenotype.

a) Discrepancies in ZEB1 probe sets

	ZEB1--1	ZEB1--2	ZEB1--3
CC20	2	1	2
CCO7	1	2	2
COLO320DM	1	1	2
NCIH716	2	1	2
OUMS23	1	1	2
OXCO1	2	2	2
RKO	2	2	2
SW480	1	1	2

b) Comparing ZEB1 probe sets in duplicate cell lines

	ZEB1--1	ZEB1--2	ZEB1--3
COLO320DM	1	1	2
COLO320HSR	1	2	2
SW480	1	1	2
SW620	1	1	2

c) Visualization of GMMchi categorization of ZEB1 probe sets

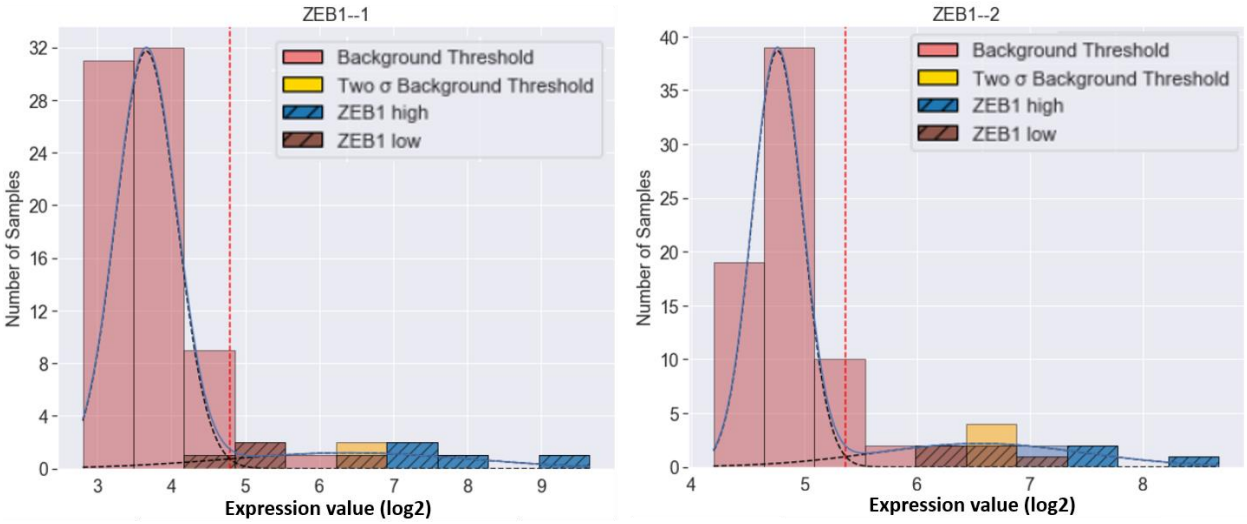


Table 4.3 Differences in ZEB1 Probes: (a) shows the categorization of CRC cell lines that differ between the three probe sets of ZEB1, the gene encoding an EMT transcription factor. All three probe sets were different enough to be deemed as splice variants and kept separate for analysis. In this comparison, ZEB1—3 seems to have the strongest consistency among EMT cell lines. (b) Comparison between duplicate cell lines shows that ZEB1—1 and ZEB1—3 are the most consistent in ZEB1 categorization. (c) GMMchi analysis shows that background threshold colored in red and yellow is the main reason behind the difference between cell lines categorized as ZEB1 high or ZEB1 low in different ZEB1 probe sets.

As expected, we found upregulated VIM and downregulated CDH1 in the returned output (Table 4.4a and 4.4d). Moreover, we also found serine/threonine-protein kinase 3 (AKT3) and mitogen-activated protein kinase 12 (MAPK12). PI3K-AKT and MAPK pathways are known to be downstream effectors of tyrosine kinase receptors (RTKs) (Osaki and Gama, 2013; Manning and Toker, 2017). At the same time, they are suggested to be the upstream activators of EMT TFs including ZEB1. Thus, we could infer AKT3 and MAPK12 as possible upstream activators of ZEB1 that will be further explored in Chapter 5.

Inversely correlated genes (Table 4.4d) such as AP-1 complex subunit mu-2 (AP1M2), Ras-related protein 25 (RAB25), CDH1 as well as epithelial cell adhesion molecule (EpCAM) are genes that are epithelial-specific markers that are thought to be downregulated when cells undergo EMT (Lapierre *et al.*, 2011; Takahashi *et al.*, 2011; Gao *et al.*, 2015). Furthermore, genes including claudin 7 (CLDN7) and plkophilin-3 (PKP3) are shown to be directly inhibited by ZEB1 through promoter binding (Larsen *et al.*, 2016). Lastly, Ovo-like transcriptional repressor 1 (OVOL1) is implicated to play an important role in mesenchymal to epithelial transition (MET), which is the opposite of EMT where mesenchymal-like cells gain epithelial-like phenotype. OVOL1 is shown to bind to the promoter region of ZEB1 to inhibit ZEB1 expression (Roca *et al.*, 2013; Saxena *et al.*, 2020). This results in the promotion of MET thus is downregulated in EMT cell lines. Studies and work done on EMT will be discussed in depth in Chapter 5.

Table 4.4 Genes Associated with EMT: (a) Genes positively correlated with EMT that do not fall into the inclusion criterion. (b) and (c) show genes positively correlated with EMT that are inclusions skewed towards +/- and -/+, respectively. (d) Genes negatively correlated with EMT that do not fall into the exclusion criterion. (e) show genes negatively correlated with EMT that are exclusions skewed towards -/-.

POSITIVE ASSOCIATIONS WITH EMT

a) Association with EMT without Inclusions

EMT	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
ZEB1--3	6	0	2	62	2.14E-07	0.85	No
nan--711	5	1	1	63	2.94E-06	0.82	No
DNAJC18--2	4	2	0	64	1.64E-05	0.80	No
ZEB1--1	4	2	0	64	1.64E-05	0.80	No
DCAF5--3	3	1	1	38	1.27E-03	0.72	No
MEX3B	4	2	1	63	7.98E-05	0.71	No
VIM--2	5	1	3	61	2.67E-05	0.69	No
REEP2	5	1	3	61	2.67E-05	0.69	No
LIX1L--2	3	3	0	64	3.65E-04	0.69	No
SNAP25_weighted	3	3	0	64	3.65E-04	0.69	No
DLL3--1	3	3	0	64	3.65E-04	0.69	No
SALL2	5	1	4	60	5.93E-05	0.64	No
APLP1	5	1	4	60	5.93E-05	0.64	No
FBXO17 /// SARS2	5	1	4	60	5.93E-05	0.64	No
NAP1L5_weighted	5	1	4	60	5.93E-05	0.64	No
MAPK12--2	5	1	4	60	5.93E-05	0.64	No
LTBP2--2	4	2	2	62	2.34E-04	0.64	No
EFNB3--1	4	2	2	62	2.34E-04	0.64	No
FAM101B_weighted	4	2	2	62	2.34E-04	0.64	No
CNTNAP1	4	2	2	62	2.34E-04	0.64	No
FKBP7--2	2	0	3	54	5.84E-03	0.62	No
YPEL1--1	3	3	1	63	1.41E-03	0.58	No
FZD8--3	3	3	1	63	1.41E-03	0.58	No
ERCC1--3	3	3	1	63	1.41E-03	0.58	No
ATP8B2--1	3	3	1	63	1.41E-03	0.58	No
C18orf54--5	3	1	3	55	2.03E-03	0.58	No
RGS4_weighted	4	2	3	61	5.31E-04	0.58	No
NRGN	4	2	3	61	5.31E-04	0.58	No
FSD1	5	1	6	58	2.11E-04	0.57	No

b) Association with EMT with Inclusions (skewed to +/-)

EMT	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
PCDHB7	2	4	0	64	6.21E-03	0.56	Significant
ST6GALNAC6	2	4	0	64	6.21E-03	0.56	Significant
GPC6_weighted	2	4	0	64	6.21E-03	0.56	Significant
ZNF474	2	4	0	64	6.21E-03	0.56	Significant
MCL1_weighted	2	4	0	64	6.21E-03	0.56	Significant

PI4K2A--3	2	4	0	64	6.21E-03	0.56	Significant
FAM13A--2	2	4	0	64	6.21E-03	0.56	Significant
MGC16121 /// MIR503	2	4	0	64	6.21E-03	0.56	Significant
DPF1--2	2	4	0	64	6.21E-03	0.56	Significant
PVT1--2	2	4	0	64	6.21E-03	0.56	Significant
BEAN1	2	4	0	64	6.21E-03	0.56	Significant
PCDH9--1	2	4	0	64	6.21E-03	0.56	Significant
NPTX1	2	4	0	64	6.21E-03	0.56	Significant
FAM84B--2	2	4	0	64	6.21E-03	0.56	Significant
AKT3--5	2	4	0	64	6.21E-03	0.56	Significant

c) Association with EMT with Inclusions (skewed to -/+)

EMT	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
LOC100506844	6	0	4	60	1.60E-06	0.75	Significant
IKBIP--1	6	0	5	59	3.52E-06	0.71	Significant
CCDC88A_weighted	6	0	8	56	2.29E-05	0.61	Significant
LRRC8C--3	4	0	7	47	7.78E-04	0.56	Significant
AKT3_weighted	6	0	10	54	6.11E-05	0.56	Significant

NEGATIVE ASSOCIATIONS WITH EMT

d) Association with EMT without Exclusions

EMT	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
CDH1_weighted	0	6	63	1	5.34E-08	-0.92	No
AP1M2_weighted	0	6	63	1	5.34E-08	-0.92	No
OVOL1--2	0	6	62	2	2.14E-07	-0.85	No
RAB25	0	6	62	2	2.14E-07	-0.85	No
GRTP1--2	0	5	58	2	2.54E-06	-0.83	No
TMEM125	1	5	63	1	2.94E-06	-0.82	No
DSC2--3	1	5	63	1	2.94E-06	-0.82	No
CLDN7	1	5	63	1	2.94E-06	-0.82	No
GALNT3--1	1	5	63	1	2.94E-06	-0.82	No
PKP3_weighted	1	5	63	1	2.94E-06	-0.82	No
IRF6_weighted	1	5	63	1	2.94E-06	-0.82	No
TJP3--1	0	4	57	2	2.52E-05	-0.80	No
SLPI	0	4	55	2	2.87E-05	-0.80	No
F11R--3	0	6	61	3	6.41E-07	-0.80	No
GRHL2	0	6	61	3	6.41E-07	-0.80	No
ERCC6 /// PGBD3--2	0	4	22	2	7.33E-04	-0.78	No
PRSS16	0	5	43	3	2.38E-05	-0.76	No
PRRG2	0	3	35	2	1.01E-03	-0.75	No
nan--8774	1	5	62	2	1.01E-05	-0.75	No
ST14_weighted	1	5	62	2	1.01E-05	-0.75	No
TMC4	1	5	62	2	1.01E-05	-0.75	No
TJP2--1	1	5	62	2	1.01E-05	-0.75	No

EPCAM	1	5	62	2	1.01E-05	-0.75	No
RDH13--1	0	4	50	3	8.86E-05	-0.73	No
MAL2	2	4	63	1	7.98E-05	-0.71	No
CDS1_weighted	2	4	63	1	7.98E-05	-0.71	No
F2RL1--2	2	4	63	1	7.98E-05	-0.71	No
LAD1--2	2	4	63	1	7.98E-05	-0.71	No
KRT8--1	2	4	63	1	7.98E-05	-0.71	No
RAB17	2	4	63	1	7.98E-05	-0.71	No
KRT19--1	2	4	63	1	7.98E-05	-0.71	No
CNKSR1	2	4	63	1	7.98E-05	-0.71	No
OVOL2--2	2	4	62	1	8.46E-05	-0.71	No
MAP7_weighted	1	5	61	3	2.67E-05	-0.69	No
STAP2	3	3	64	0	3.65E-04	-0.69	No
MARVELD2--1	3	3	64	0	3.65E-04	-0.69	No
LSR	3	3	64	0	3.65E-04	-0.69	No
VAMP8	3	3	64	0	3.65E-04	-0.69	No
MYO5B--2	1	5	55	3	4.22E-05	-0.69	No
C20orf54--2	0	2	24	2	1.59E-02	-0.68	No

e) Association with EMT with Exclusions (skewed to -/-)

EMT	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
LOC283278	0	6	60	4	1.60E-06	-0.75	Significant
nan--7559	0	6	60	4	1.60E-06	-0.75	Significant
ELMO3	0	6	60	4	1.60E-06	-0.75	Significant
PRSS8	0	6	60	4	1.60E-06	-0.75	Significant
ELL3 /// SERINC4--2	0	5	47	4	3.30E-05	-0.72	Significant
KRTCAP3	0	6	59	5	3.52E-06	-0.71	Significant
SH3YL1--2	0	6	59	5	3.52E-06	-0.71	Significant
C1orf210	0	6	43	5	1.79E-05	-0.70	Significant
PRR15	0	6	58	6	7.05E-06	-0.67	Significant

4.2.3.4 Mismatch Repair Deficiency

It is known that at least 15% of sporadic colorectal tumors exhibit DNA mismatch replication error repair (MMR) deficiency (Sinicrope, 2010). MMR deficiency is a genetic instability involving defects in the mismatch repair pathway that results in an increased mutation rate due to unrepaired DNA replication errors. This is especially found in regions of repetitive DNA such as microsatellites. Moreover, it is previously found in our lab that genes that are most differentially expressed in MMR deficient cell lines are genes with 3'UTR T repeats that are longer

than 11 base pairs (Wilding *et al.*, 2010). This observation is mainly thought to be linked to the accumulation of mutations in repetitive regulatory 3'UTR sequences. In our in-house data, 71 of the cell lines have known replication error (RER) status. 18 (25%) are MMR deficient labeled as RER+.

MutL Homolog 1 or *MLH1* is a gene that encodes a key DNA mismatch repair protein important to the normal function of the DNA mismatch repair pathway (McCarthy *et al.*, 2019). As expected, RER status is significantly and negatively associated with *MLH1* expression, as well as reduced transforming growth factor beta receptor 2 (*TGFBR2)*. This further reinforces the role of MLH1 in the mismatch repair pathway. Furthermore, TGFBR2 is known to be mutated in RER+ or repair mismatch deficient colorectal tumors, specifically at repetitive sequences (Xu and Pasche, 2007). Another gene Villin 1 or VIL1 is also shown to have reduced promoter activity in RER+ cancer tumors.

Compared to the findings of Jennifer L. Wilding (Wilding *et al.*, 2010), a previous post-doctoral researcher in our lab, published on replication error deficiency, we found that 22 genes overlapped with the top 30 genes that are most differentially expressed due to RER shown in the literature. We carried out the same analysis as in the paper and demonstrated the result in a volcano plot shown in Figure 4.5.



Figure 4.5 Volcano Plot of RER-related Differentially Expressed Genes: Shows a volcano plot of differentially expressed genes between RER+ and RER- CRC cell lines. Genes labeled are shown to be differentially expressed in the paper Jennifer L. Wilding *et al.* published on differentially expressed genes in RER CRC cell lines.

The 22 overlapping genes between our association study and the paper's finding include PAFAH1B2, CDC42SE1, RAPGEF6, STRN3, TMTC3, TOB1, THRAP3, TOR1AIP2, IQGAP3, LRRC16A, CTNNB1, ITCH, LETM1, MTA2, HNRNPL, RIT1, CASP2, USF1, CDC42, TNRC6, MARCH6, and HNRNPH1 highlighted in grey in Table 4.5.

Table 4.5 Genes Associated with RER Status: (a) Genes positively correlated with RER that do not fall into the inclusion criterion. (b) and (c) show genes positively correlated with RER that are inclusions skewed towards +/- and -/+, respectively. (d) Genes negatively correlated with RER that do not fall into the exclusion criterion. (e) and (f) show genes negatively correlated with RER that are exclusions skewed towards -/- and +/+, respectively. The 22 genes that overlapped with the top 30 RER differentially expressed genes shown in *Jennifer L. Wilding et al*. (Wilding *et al.*, 2010) are highlighted in grey.

POSITIVE ASSOCIATIONS WITH RER STATUS

a) Association with RER without Inclusions

RER status	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
CDC42SE1--3	14	1	1	49	3.62E-12	0.91	No
HNRNPL_weighted	16	2	1	51	1.01E-12	0.89	No
STRN3--2	14	1	2	46	4.63E-11	0.87	No
TMEM33--5	8	1	1	47	4.81E-08	0.87	No
MTA2	15	3	1	51	1.72E-11	0.85	No
SLC30A7--3	6	2	1	31	4.85E-05	0.76	No
BOD1L--3	6	3	1	45	1.92E-05	0.72	No
TMTC3--1	8	2	3	44	4.01E-06	0.71	No
NSD1--2	8	1	5	33	3.26E-05	0.67	No
EWSR1--3	9	2	5	28	1.18E-04	0.62	No
ING5--3	6	1	5	30	5.43E-04	0.61	No
DOCK5--5	6	6	1	49	9.56E-05	0.60	No
PCGF3--3	8	2	5	25	5.56E-04	0.59	No
FGFR1_weighted	12	6	5	47	6.43E-06	0.58	No
ADAM22--5	6	4	2	29	1.06E-03	0.58	No
FDXACB1	7	4	3	34	4.09E-04	0.57	No
TRAPPC10--1	10	2	6	21	8.90E-04	0.57	No
NUP43--2	6	4	3	40	6.10E-04	0.55	No
KLHL17--1	7	1	5	11	2.72E-02	0.53	No
CLSPN--4	5	5	2	40	1.69E-03	0.52	No
GTPBP3--1	6	4	3	28	2.86E-03	0.52	No
TNRC6A--1	6	12	5	47	2.68E-02	0.28	No

b) Association with RER with Inclusions (skewed to +/-)

RER status	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
TOR1AIP2--4	5	2	0	46	7.32E-06	0.83	Significant
MAP6--2	4	2	0	14	3.10E-03	0.76	Significant
SMU1--1	11	7	1	51	1.57E-07	0.69	Significant
USF1	9	9	0	52	7.48E-07	0.65	Significant
ITCH--4	5	9	0	52	2.24E-04	0.55	Significant
LYG1	6	12	0	52	1.42E-04	0.52	Significant
ARL9	6	12	0	52	1.42E-04	0.52	Significant
IQGAP3--1	3	7	0	51	3.33E-03	0.51	Significant

c) Association with RER with Inclusions (skewed to -/+)

RER status	+/+	+/-	-/+	-/-	p-value	r-value	Inclusion Criterion
DCAKD--1	14	0	6	20	3.34E-06	0.73	Significant
CASP2_weighted	18	0	11	41	1.49E-09	0.70	Significant
C1orf58--2	13	0	9	23	6.81E-06	0.65	Significant
LETM1--1	18	0	14	38	2.03E-08	0.64	Significant
C19orf23--1	6	0	5	15	2.01E-03	0.64	Significant
C17orf65	12	0	8	17	7.13E-05	0.64	Significant
AMFR--1	16	2	10	42	2.23E-07	0.63	Significant
HNRNPH1_weighted	17	1	12	40	9.32E-08	0.63	Significant
THRAP3--1	16	0	8	12	1.72E-04	0.63	Significant
CTNNB1—1	15	0	13	28	4.61E-06	0.60	Significant
TAF6L--1	4	0	5	20	5.31E-03	0.60	Significant
VPS37D	6	0	5	9	1.41E-02	0.59	Significant
INCENP	11	0	10	18	2.29E-04	0.58	Significant
ARL5A—3	14	1	11	26	3.15E-05	0.58	Significant
nan--8757	16	1	12	22	6.18E-05	0.56	Significant
GPR27—1	8	0	6	7	1.80E-02	0.55	Significant
NAA15--5	15	1	12	22	1.45E-04	0.55	Significant
EIF2C1--2	13	0	11	13	8.69E-04	0.54	Significant
TOR1AIP2--6	17	0	12	12	3.81E-04	0.54	Significant
CNTROB--3	5	0	5	7	4.41E-02	0.54	Significant
nan--5187	13	0	14	21	1.46E-04	0.54	Significant
PAFAH1B2--2	18	0	18	14	6.81E-04	0.47	Significant
FNIP1 /// RAPGEF6--1	16	1	23	16	1.06E-02	0.35	Significant
CCDC42B--1	10	2	16	17	4.65E-02	0.31	Significant

NEGATIVE ASSOCIATIONS WITH RER STATUS

d) Association with RER without Exclusions

RER status	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
TXNDC9_weighted	4	14	50	3	7.03E-09	-0.74	No
PPP1CB--4	5	13	50	3	6.40E-08	-0.69	No
B4GALT1--5	6	12	51	2	1.08E-07	-0.69	No
CFLAR_weighted	4	14	44	9	5.40E-06	-0.57	No
ZNF148--3	9	9	51	2	2.71E-05	-0.56	No
nan--7225	7	11	48	5	3.03E-05	-0.54	No
ZBTB38--2	6	12	45	8	7.92E-05	-0.50	No
AMIGO2	5	13	40	13	5.14E-04	-0.43	No
TUSC1	9	9	47	6	1.37E-03	-0.41	No
MALAT1--11	9	9	47	6	1.37E-03	-0.41	No
CGN_weighted	8	10	45	8	1.42E-03	-0.40	No
RBPMS_weighted	7	11	43	10	1.88E-03	-0.40	No
TGFBI	7	11	42	11	2.65E-03	-0.38	No
MUC20_weighted	7	11	42	11	2.65E-03	-0.38	No
PPP1R12A--1	11	7	49	4	4.07E-03	-0.38	No

e) Association with RER with Exclusions (skewed to -/-)

RER status	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
PTCD2--2	1	17	37	16	2.54E-06	-0.56	Significant
APOOL--5	3	15	40	13	1.58E-05	-0.52	Significant
MARCKS_weighted	4	14	40	13	1.29E-04	-0.48	Significant
nan--11384	4	14	39	14	1.95E-04	-0.46	Significant
HSPA1A	0	18	26	27	1.05E-04	-0.44	Significant
PDK3_weighted	1	17	29	24	2.10E-04	-0.43	Significant
RIT1--3	1	17	28	25	2.80E-04	-0.42	Significant
SPINT1_weighted	5	13	39	14	1.43E-03	-0.41	Significant
C6orf132_weighted	2	16	31	22	7.19E-04	-0.41	Significant
GPX2--2	2	16	31	22	7.19E-04	-0.41	Significant
AGXT2L2--2	1	17	27	26	5.74E-04	-0.40	Significant
DAB2_weighted_1	0	18	23	30	3.04E-04	-0.40	Significant
SYTL5--2	1	17	26	27	7.25E-04	-0.39	Significant
RHOBTB3--1	5	13	38	15	1.73E-03	-0.39	Significant

f) Association with RER with Exclusions (skewed to +/+)

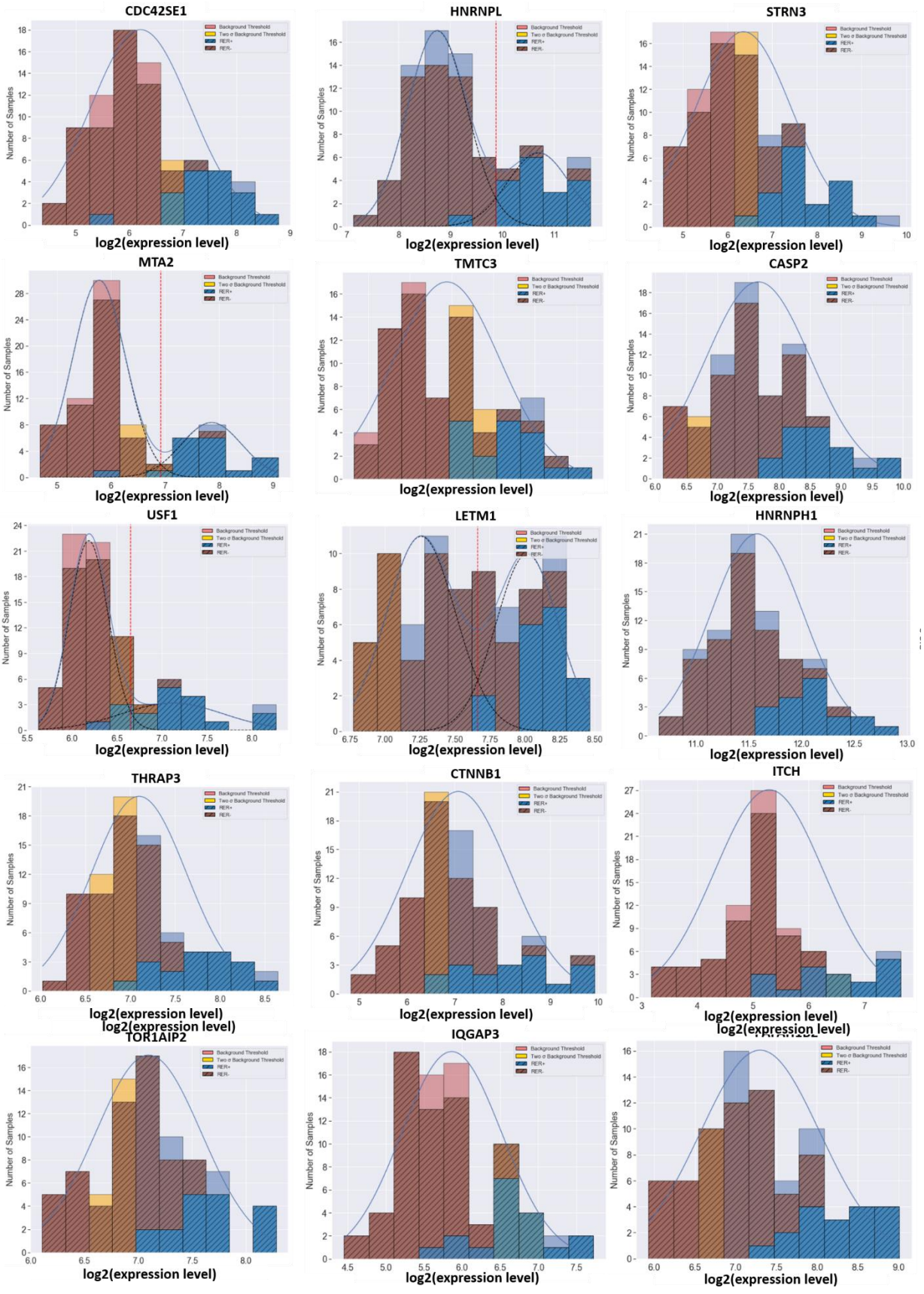
RER status	+/+	+/-	-/+	-/-	p-value	r-value	Exclusion Criterion
MDM4--5	5	13	53	0	1.47E-10	-0.81	Significant
SEC22B--3	6	12	53	0	1.45E-09	-0.77	Significant
COCH--1	6	12	53	0	1.45E-09	-0.77	Significant
LOC100505937	6	12	53	0	1.45E-09	-0.77	Significant
IAH1--2	8	10	53	0	9.48E-08	-0.69	Significant
TOB1--2	6	7	52	0	2.46E-06	-0.69	Significant
HNRNPH1--4	10	8	53	0	4.11E-06	-0.61	Significant
LTN1--1	10	8	52	1	3.18E-05	-0.56	Significant
CTTN--6	12	6	53	0	1.30E-04	-0.52	Significant
HSPH1--1	12	6	53	0	1.30E-04	-0.52	Significant
MLH1	12	6	53	0	1.30E-04	-0.52	Significant
ZNF140--1	13	5	53	0	6.58E-04	-0.47	Significant
LRRC16A—4	11	4	49	0	2.15E-03	-0.47	Significant
SLC38A1--1	12	6	52	1	7.64E-04	-0.46	Significant
C1orf59	11	7	51	2	6.21E-04	-0.46	Significant
SEC31A_weighted	14	4	53	0	3.15E-03	-0.42	Significant
PURA--1	14	4	53	0	3.15E-03	-0.42	Significant
WTAP--5	11	7	50	3	1.75E-03	-0.42	Significant
SMAP1	11	7	50	3	1.75E-03	-0.42	Significant
DDX27--2	12	6	51	2	2.57E-03	-0.41	Significant
MARCH6--3	9	3	40	1	3.25E-02	-0.36	Significant

6 of the 22 genes highlighted in grey in Table 4.5a had very little few cell lines that fell into the +/- or -/+ category, suggesting that the expression level is heavily influenced by RER status. These genes include *HNRNPL*, a protein of the heterogeneous nuclear ribonucleoparticle family thought to regulate RNA stability, as well as *MTA2*, a chromatin remodeling protein thought to mediate cytoskeletal reorganization in tumor cells, consistent with prior studies of the effects of this highly perturbative characteristic on gene expression, transcript stability, and cell behavior.

3 of the 22 genes, which are inclusions skewed towards +/- (Table 4.5b), exhibited a pattern where only RER- cell lines expressed these genes. This is because in these cases the mutations in the repeats suppress expression so that it resulted in a whole family of associated gene expressions which are only expressed when they are mutated. These genes include USF1, ITCH, and IQGAP. The 3'UTR regions of the three genes contained T repeats that are greater than 25 base pairs, putting them at the higher end of the spectrum. The greater the number of repeats, the more likely mutation could occur and suppress gene expression.

9 of the 22 genes that fit the inclusion criterion skewed towards -/+ shown in Table 4.5c revealed an opposite pattern where these genes are almost always expressed in RER+ cell lines but not always expressed in RER- cell lines. Here, we found HNRNPH1, which belongs to the same family of hnRNP proteins with HNRNPL; nonetheless, exhibiting a different pattern of association. Although hnRNP proteins are thought to play an important role in regulating RNA stability, the difference in the pattern of association suggests further heterogeneity within the effects on gene expression based on RER deficiency.

All of these genes are thought to be highly enriched with sequences containing mononucleotide repeats within the 3'UTR regions. All of them had 3'UTR T repeats with lengths between 10 to 32 base pairs. The GMMchi analysis of the 22 genes is shown in Figure 4.6. In addition to the 22 overlapping genes, we explored the 8 genes that were not returned in our association study. We found that other than the gene SET///hCG_1644608, which did not exist in our microarray data, the remaining 7 genes were classified as unimodal and thus were excluded in our association study. Nonetheless, the GMMchi analysis shown in Figure 4.7 reveals a general pattern of association, supporting the findings in the paper.


CDC42SE1
HNRNPL
STRN3
MTA2
TMTC3
CASP2
USF1
LETM1
HNRNPH1
THRAP3
CTNNB1
ITCH
TOR1AIP2
IQGAP3
Number of Samples
log2(expression level)
Background Threshold
Two σ Background Threshold
RER+
RER-

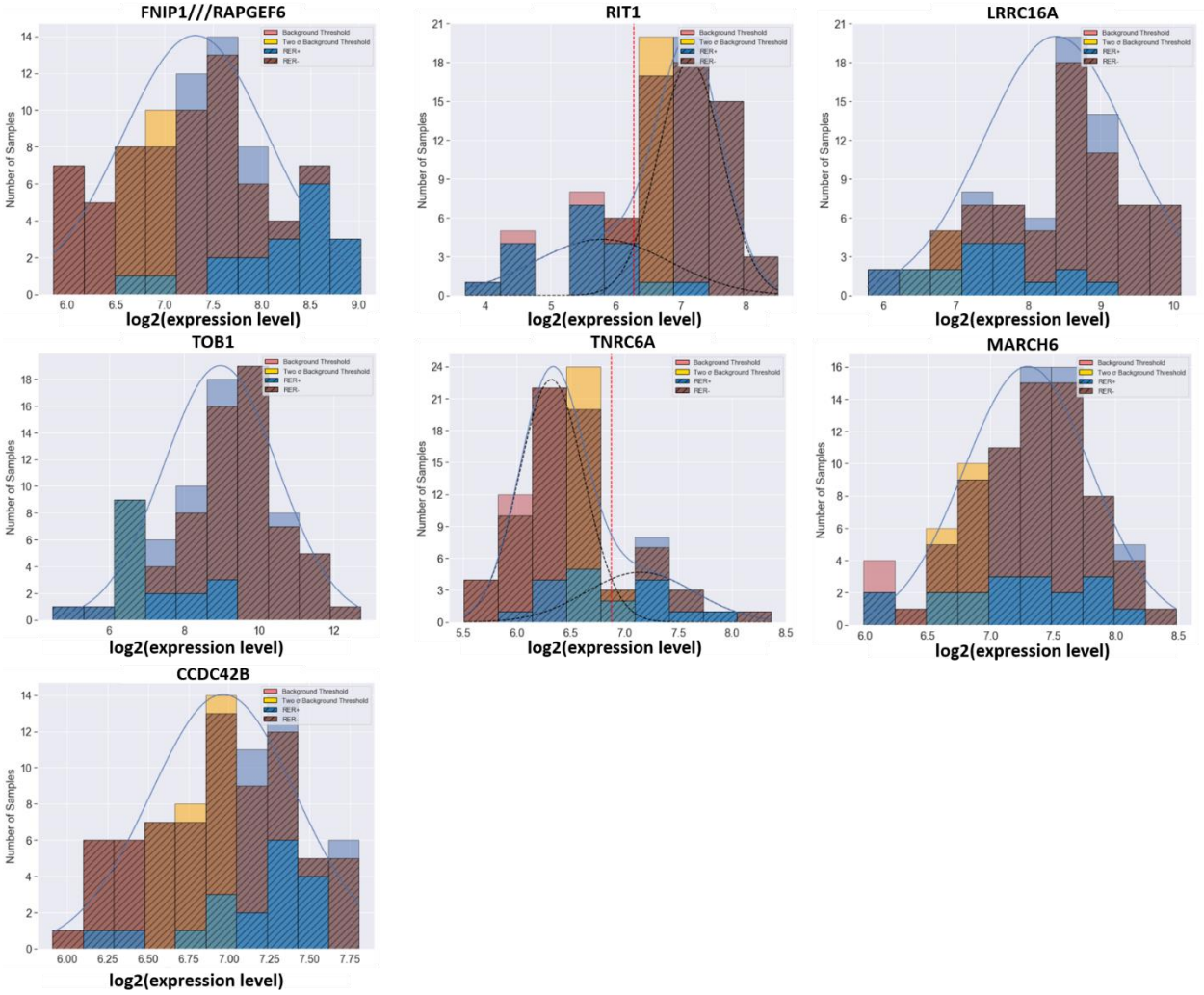
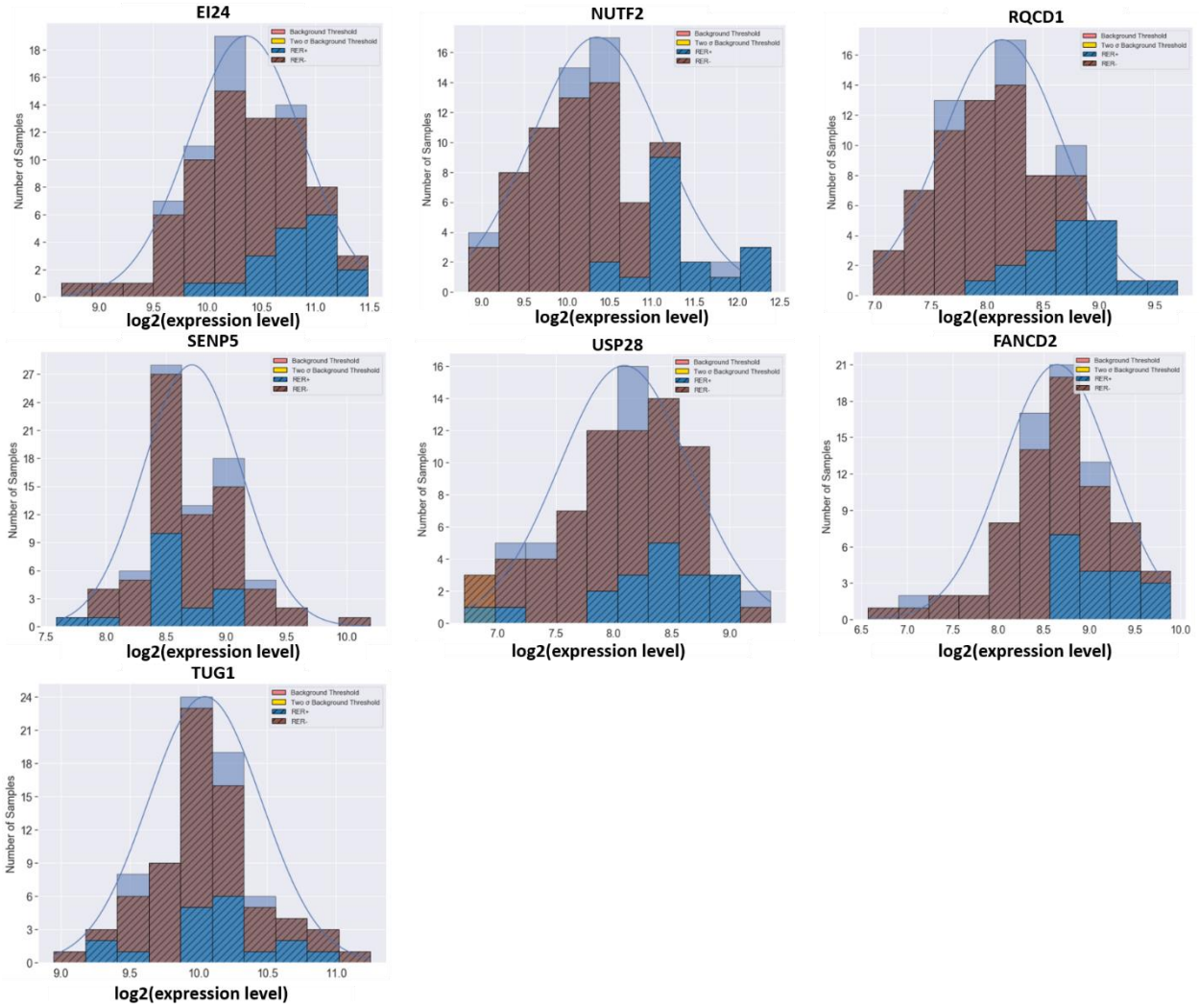


Figure 4.6 GMMchi Analysis on the 22 Overlapping RER-related Genes: GMMchi analysis on the 22 genes returned in our association study with RER status that overlapped with the top 30 genes that are most differentially expressed due to RER show in the paper Jennifer L. Wilding *et al.* (Wilding *et al.*, 2010) published on differentially expressed genes in RER CRC cell lines. Cell lines that are RER- are labeled in brown and those that are RER+ are labeled in blue.

Figure 4.7 GMMchi Analysis on the 7 Non-Overlapping RER-related Genes Reveal a General Trend in Association: GMMchi analysis on the 7 genes that did not return in our association study with RER status but found as differentially expressed due to RER show in the paper Jennifer L. Wilding *et al.* (Wilding *et al.*, 2010) published on differentially expressed genes in RER CRC cell lines. All the 7 genes exhibit a unimodal distribution, explaining why they were excluded from our association study. Nonetheless, we observe a general pattern of association based on the distribution of RER+ and RER- in the higher and lower end of the gene expression distribution. SENP5 and TUG1 revealed the least observable trend in the association. Cell lines that are RER- are labeled in brown and those that are RER+ are labeled in blue.



4.2.4 Well-defined Bimodal Distribution Criterion

Next, we explored genes with very clearly defined bimodal distributions based on the GMMchi categorization. As indicated in Chapter 3, the motivation for establishing the GMMchi

pipeline is to identify well-defined bimodal distributions that allow us to analyze objectively the correlation of bimodal patterns with novel mutations or epigenetic changes. We explored two datasets including our in-house microarray data with 78 CRC cell lines as well as the cell line RNA-seq data downloaded from CCLE, which includes 56 CRC cell lines (Ghandi *et al.*, 2019). 37 CRC cell lines overlapped across the two datasets. We found that 9 out of the 19 cell lines that did not overlap and are not included in our microarray data are Seoul National University (SNU) CRC cell lines from Korea.

We established a criterion to identify well-defined bimodal distributions. We started by visually identifying distribution patterns that exhibit a well-defined bimodal distribution. The list of genes that represented a well-defined bimodal distribution came down to be MUC13 probe set 1, CDX1, and CDX2 shown in Figure 4.8a. Then, we objectively characterized these distributions using a list of measurable attributes:

1. Category: The modality of the distribution (unimodal vs bimodal)
2. Chiscore: Chi-square goodness of fit value
3. High_percentage: Percentage of the distribution categorized as high expressing
4. Overlap_percentage: Percentage of the overlap between the two distributions if the gene is categorized as a bimodal distribution
5. Weight: The ratio of the weight of the two distributions

By measuring these attributes (Figure 4.8b) of every gene distribution pattern, the criterion's goal is to identify bimodal distributions (using **category**) that have a good fit (using **chiscore**) in which there is not a large overlap (using **overlap_percentage**) between the two

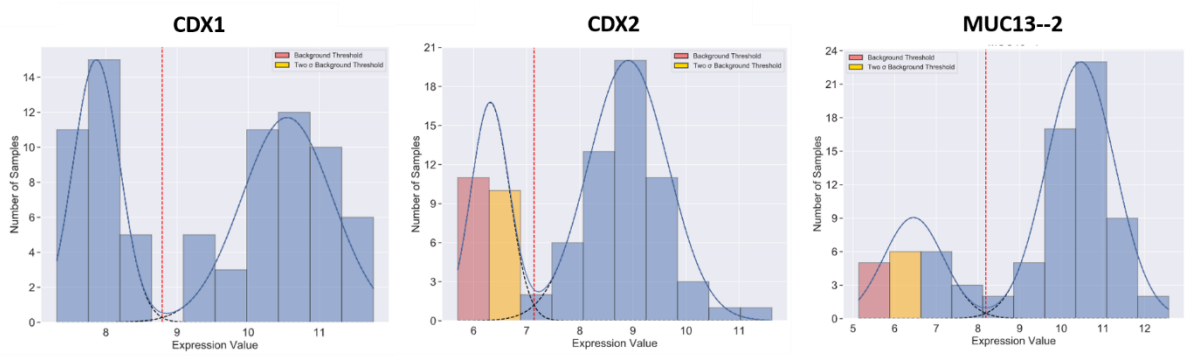
distributions. It also needs a minimum percentage of the data lying in each distribution (using **high_percentage**) while not having any of the distribution being too flat or non-normally distributed (using **weight**). The **weight** variable is outputted by GMM which represents the ratio of the two normal distributions, allowing us to access the relative height of the two distributions. It is different from the **high_percentage** variable where it measures the absolute percentage of total datapoints that lie in each of the distributions.

Next, we took the parameters of the 3 genes (CDX1, CDX2, MUC13) to determine the thresholds for each attribute. This resulted in a chi-score value threshold of 800, a minimum of 10% of total cell lines residing in the high expressing group, less than 7.8% of overlap, and a weight threshold of 4. The result of our criterion is shown in Table 4.6a, which resulted in 64 genes. The same process is applied to RNA-seq data downloaded from CCLE. The threshold for each category resulted in a chi-score value threshold of 60, a minimum of 19% of total cell lines residing in the high expressing group, less than 2% of overlap, and a weight threshold of 6. The result of our criterion is shown in Table 4.6b, which gave rise to 82 genes. Examples of GMMchi analysis on genes that fit the well-defined bimodal criterion in both microarray data and CCLE are shown in Figures 4.11a and 4.11b. One thing to note is that the threshold for each category for our criterion is different between the CCLE and microarray data. This is reasoned to be due to the difference in the number of cell lines (CCLE = 56, in-house microarray = 78) as well as the technology used (RNA-seq vs microarray), which resulted in a difference in the patterns of gene expression even for the same gene. Another obvious distinction is in the **chiscore** value of the two datasets. In our microarray data, the **chiscore** threshold is set at 800 while in the CCLE data the **chiscore** threshold is set at 60. The reason why we are seeing such a large discrepancy in the

chi-square value is due to the nature of the datasets. Microarray data, for example in Figure 4.9, normally has a log2(expression value) range of 6-14 while RNA-seq data has a log2(expression value) range of -8 to 3. Since the χ^2 value is calculated based on the difference in the area under the curve, this means that when we apply the chi-square goodness of fit formula of $\chi^2 = \frac{(Observation - Expectation)^2}{Expectation}$ the χ^2 value of microarray data will always be much larger than that of RNA-seq data due to the difference in the range of expression values on the x-axis.

Among the 64 genes and 82 genes identified as having well-defined bimodal distributions in the microarray data and CCLE cell line data, respectively, 10 genes overlapped between the two lists (Table 4.6c). The association study between well-defined bimodal distributions against mutation, methylation, and enhancers data will be done on both our microarray and CCLE cell line data.

a)



b)

	category	chiscore	high_percentage	overlap_percentage	Weight
CDX1	bimodal	27.56942	60.3	0	1.527
CDX2	bimodal	790.9386	70.5	7.7	2.474
MUC13--2	bimodal	155.6665	75.6	1.3	3.005

Figure 4.8 Genes Used to Determine the Criteria for Well-Defined Bimodal Distributions: (a) CDX1, CDX2, and MUC13 are genes where the pattern distribution is visually determined to fit the well-defined bimodal distribution criterion. They are used as benchmarks to establish quantified attributes that defined the well-defined bimodal distribution criterion, which consists of category, chiscore, high_percentage, overlap_percentage, and weight shown in (b). A snippet of the resulting matrix of the attributes quantified for each gene.

a) **CCLE well-defined bimodal genes**

SLC38A5	STC2	DTX3	TXLNG2P	MAP7D2	NLRP2	ZNF550	GJB1	TMSB4Y	RP11-1L12.3	RP11-745O10.2	KDM5D	ZNF71
SLFN13	CTD-3157E16.1	TBX20	NPSR1	ZNF844	BATF	ZNF530	UTY	AP001065.2	TRIM60P18	ALCAM	ZNF420	NPIPB15
AKT3	PCSK1N	GATA6-AS1	HUNK	DOK4	NR1I2	FBXO17	ZNF518B	AREG	DDX3Y	RP4-765C7.2	HNF1A-AS1	PLA2G3
ZFY	RPS4Y1	PARVB	GJB4	CDK14	RP11-297P16.4	AC073283.7	AC108142.1	LINC00649	TRIM58	AC104135.3	LINC00675	MTERF
EIF1AY	ANKRD19P	PVRL4	CLDN11	HNF1B	WBP5	AF131215.2	RP5-881L22.5	CTD-2021H9.3	DUSP27	AFAP1-AS1	LGALS4	CDX1
TRAPPC2P1	CEL	ANKS4B	SCAND3	HS6ST2	RP11-362F19.1	PRKY	MUC13	ZNF606	GSPT2	FAM21B	ZNF548	NGFRAP1
AC016700.6	PRSS8	RP11-440D17.3	HEPH									

b) **Microarray well-defined bimodal genes**

C9orf152	CFTR--1	ENTPD5--2	TRPM2	PPP1R9A--2	DISC1 /// TSNAX-DISC1	PPP1R1B	XIST--5	KRT20	KIAA1199--2	MUC13--1	DDX3Y--3
USH1C--2	ABP1	DDC--2	CCND2--3	GSTT1--1	GPX2--1	CYorf15B--2	ABHD12B--2	VAV3--2	C10orf99--1	MUC13--2	CDX2
ZDHHC2--3	CDX1	EIF1AY--1	LOC654433--1	XIST--4	TOX3--1	CFTR--2	CLRN3	POF1B--3	XIST--1		
C2orf89	AKT3--1	DDC--1	LGALS4	GIPC2--1	SNRPN /// SNURF--1	HSPA1A	HEPH--2	EPDR1	SGCE		
HOXB2	LY75	VIL1--3	HS6ST2--2	C10orf99--2	ADAM19--1	XIST--6	TOX3--2	GUCY2C	TP53--2		
PROM1	HS6ST2--3	ZNF655--3	TOX3--3	XIST--2	KDM5D	SLC38A5	TP53--1	XIST--3	CDH17		

c) **Overlapping well-defined bimodal genes in both CCLE and microarray data**

HS6ST2	CDX1	EIF1AY	HEPH	DDX3Y	KDM5D	SLC38A5	LGALS4	MUC13	AKT3

Table 4.6 Summary of Well-Defined Bimodal Genes in Our Microarray and CCLE Data: (a) Full list of the 82 genes that fitted the well-defined bimodal distribution criterion in the CCLE RNA-seq data. (b) Full list of the 64 genes that fitted the well-defined bimodal distribution criterion in our microarray data. (C) The full list of the 10 genes that fitted the well-defined bimodal distribution criterion and overlapped in both CCLE and microarray data.

Figure 4.9 Examples of GMMchi Analysis on Well-Defined Bimodal Genes: (a) Examples of genes identified as well-defined bimodal distributions in our microarray cell line data. (b) Examples of genes identified as well-defined bimodal distribution in the CCLE RNA-seq cell line data.

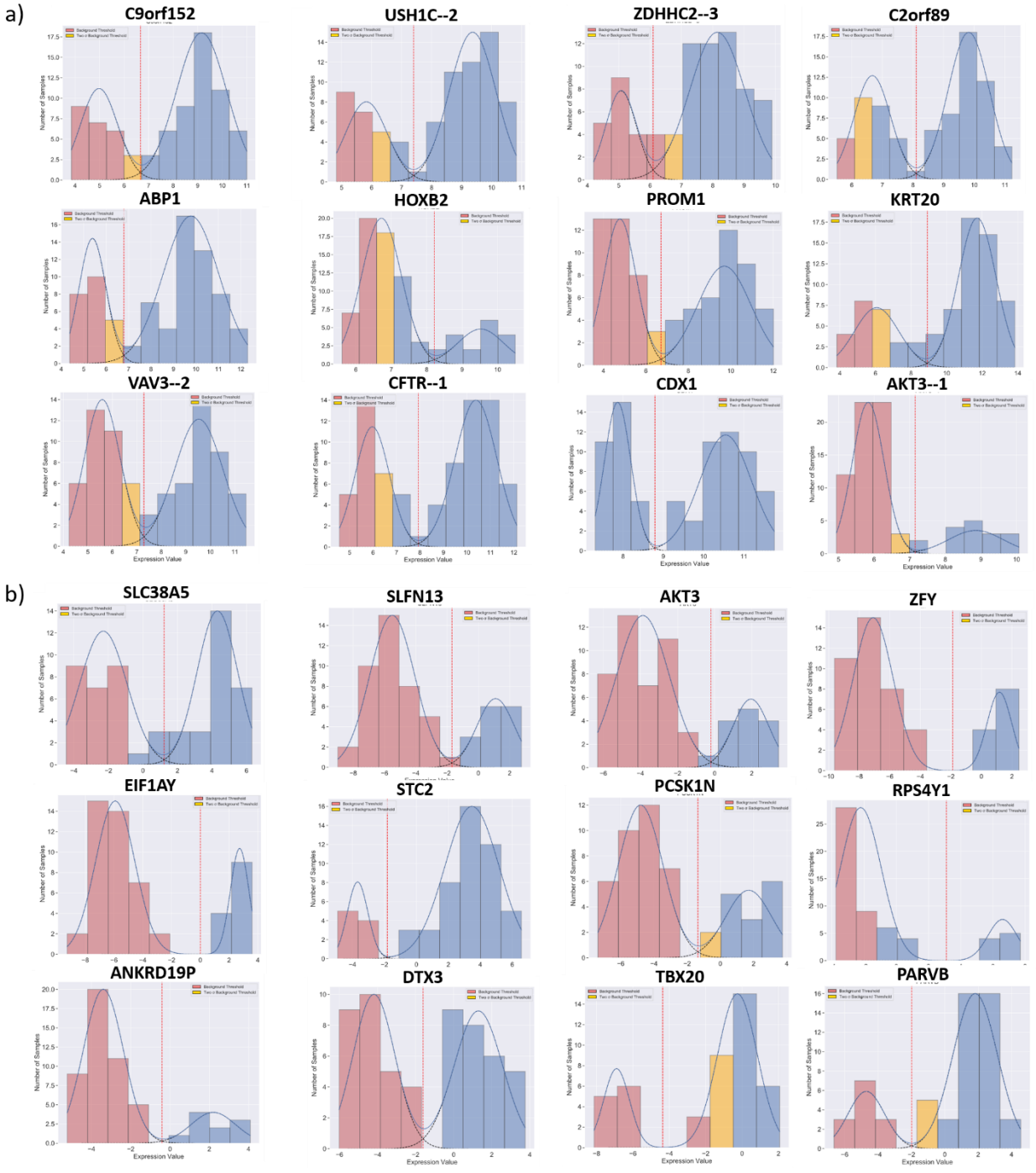
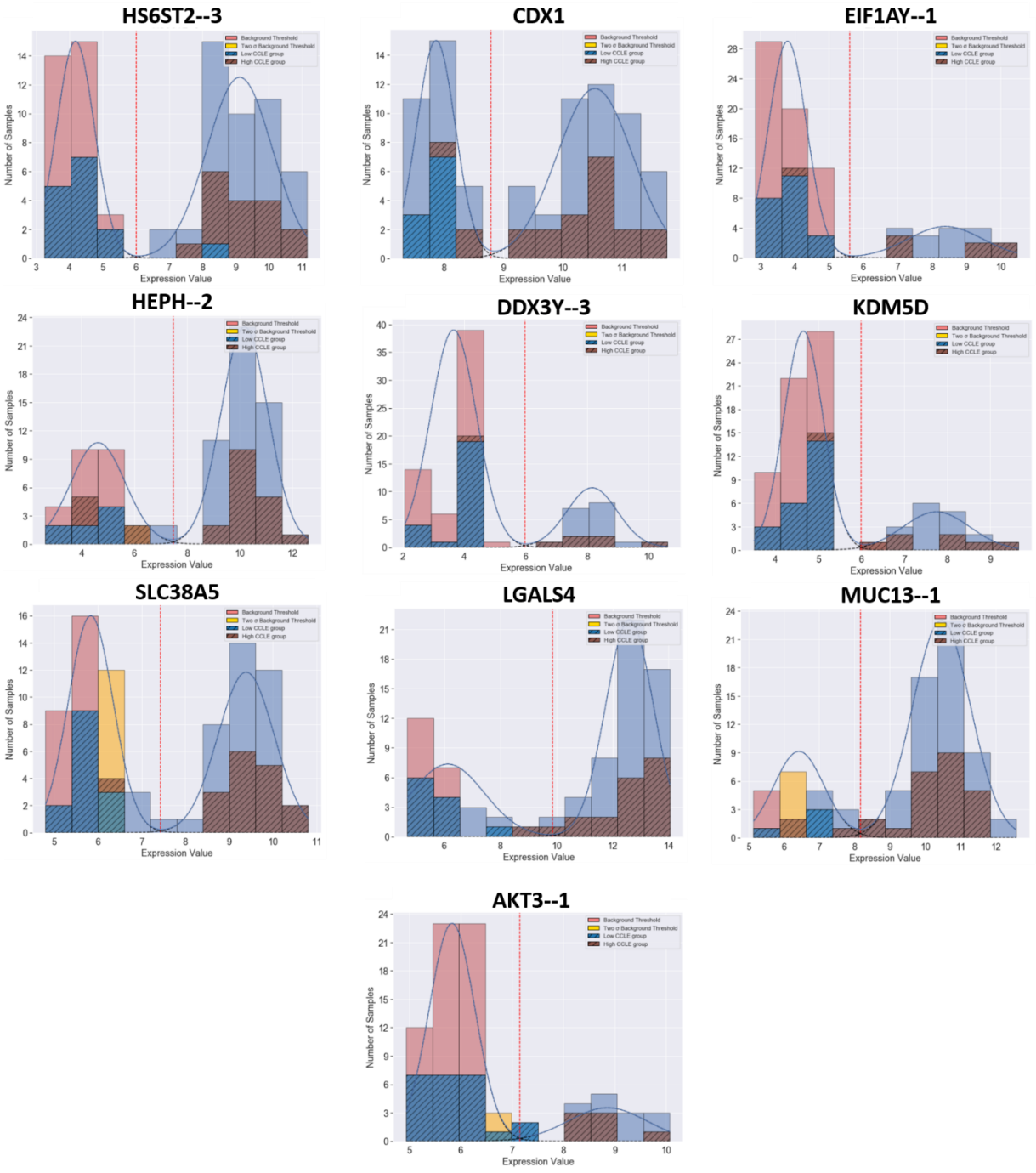
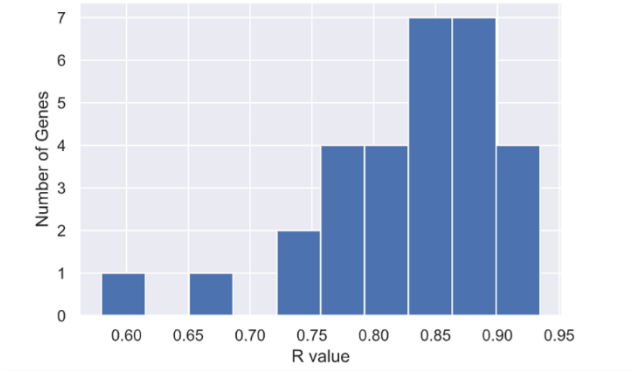


Figure 4.10 10 Overlapping Well-Defined Bimodal Genes in Both Datasets: 10 genes from the microarray data that are deemed to be well-defined bimodal distributions in both CCLE and our microarray data. The GMMchi analysis is done on our microarray data. The hatched cell lines are overlapping cell lines found in both CCLE and our microarray data. Blue indicates low expressing cell lines and brown indicates high expressing cell lines in CCLE. The distributions of the blue and brown hatches representing CCLE data revealed a similar pattern compared to our microarray data in the categorization of low and high expressing cell lines across both datasets.



With only 10 overlapping genes (shown in Figure 4.10), we further explored the low percentage of overlap identified as well-defined bimodal across both datasets. To understand this observation, we compared to see whether the genes identified as well-defined bimodal in one dataset exhibited a different pattern in the other dataset. We only looked at cell lines that are shared across the two datasets. In Figure 4.11a, we found that genes that are identified as well-defined bimodal patterns in our microarray data had a relatively strong correlation or r-value > 0.6 with the same genes in CCLE, suggesting a high degree of agreement in classification. Nonetheless, in Figure 4.11b, we found that genes that are defined as well-defined bimodal in CCLE had a much wider spread of r-value when compared against the same genes in our microarray data, suggesting a wider degree of differences across the two datasets. We suspect that this is because of a larger sample size we had in our microarray data (n=78). This would mean that because of our larger collection of expression data on more CRC cell lines, our gene expression distributions exhibited patterns that are more generalizable and closer to the ground truth, which is the case if one can collect expression data on an infinite amount of CRC cell lines that fairly represented the true distribution. This observation agreed with the law of large numbers in which stated that in probability and statistics, the mean gets closer to the average of the whole population as your sample size grows larger. This would explain why well-defined bimodal distributions found in our microarray data are generally agreed by CCLE, while the other way around is not true where well-defined bimodal patterns found in CCLE are not always agreed by our microarray data. Nonetheless, the 10 overlapping genes shown in Figure 4.10 showed a very strong agreement in the categorization of expression levels across the two datasets.

a)



b)

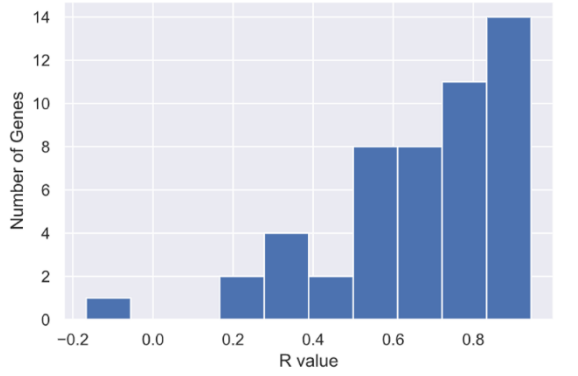


Figure 4.11 Correlations of the Same Genes Across Both Datasets: (a) Histogram showing the correlation of the expression levels between genes identified as well-defined bimodal patterns in our microarray data and the same genes in CCLE (b) Histogram showing the correlation of the expression levels between genes identified as well-defined bimodal patterns in the CCLE RNA-seq data and the same genes in our microarray data. This shows that the r-value in (a) is consistently above 0.6, suggesting a strong correlation of the gene expression level of the same genes across the two datasets. (b), however, shows a much larger spread from -0.2 to 1, suggesting a weaker overall correlation of the gene expression level of the same genes across the two datasets.

4.2.4.1 Association with Mutation

To explore novel correlations between our two lists of well-defined bimodal genes, we investigated the association against known driver mutations in colorectal cancer using a 2x2 table. Known driver mutations include APC (found in 65-80% of colorectal cancer), KRAS (30-40%), TGFRB2 (30%) PIK3CA (10-20%), BRAF (10%), CTNNB1 (10%), FBXW7 (2.8%) and SMAD4 (10%) (Chung, 2000). The mutation status of each driver mutation of our cell lines has been previously collected from multiple sources and recorded in our lab. The correlation of the 64 genes in our microarray and 82 genes in the CCLE cell line data against driver mutations are shown in Table 4.7a and 4.7b, respectively. Only p-values with ≤ 0.05 after Bonferroni multiplicity correction is shown.

a) Bodmer Lab Microarray Cell Line Data

BRAF hotspot mutations	+/+	+/-	-/+	-/-	p-value	R value
CDX1	1	9	38	20	1.40E-03	-0.39
EPDR1	3	7	46	12	3.51E-03	-0.26

SMAD4 all mutations	+/+	+/-	-/+	-/-	p-value	R value
TP53--2	4	10	34	9	9.73E-04	-0.46
TP53--1	4	10	35	8	5.07E-04	-0.49

p53 all mutations	+/+	+/-	-/+	-/-	p-value	R value
TP53—2	34	24	13	0	3.01E-03	-0.34

b) CCLE Cell line Data

FBXW7 all mutations	+/+	+/-	-/+	-/-	p-value	R value
RP11-745O10.2	0	13	14	14	1.31E-03	-0.49

Table 4.7 Mutations Associated with Well-Defined Bimodal Genes: (a) The full list of well-defined bimodal genes identified that are significantly associated with driver mutations in our microarray cell line data ranked by r-value. (b) The full list of well-defined bimodal genes that are significantly associated with driver mutations in the CCLE cell line data ranked by r-value. As shown, there are two types of driver mutation status we explored labeled as 'hotspot mutations' and 'all mutations', which represented mutations in known "hotspot" areas or locations with exceptionally high mutation frequency or mutation cluster region and all mutation types, respectively. All p-value is corrected with Bonferroni multiple testing correction. The general lack of agreement with mutations is not surprising as we would expect such associations to be more likely with methylation changes.

In total, we found 4/64 (0.06%) and 1/82 (0.01%) genes associated with driver mutations in the microarray data and the CCLE cell line data, respectively. In line with previous studies, we found a negative association between the mutations in tumor suppressor p53 and SMAD4 and the expression level of p53 (TP53) (Kawaguchi *et al.*, 2019). Note that in the 2x2 table shown between p53 mutations and the gene TP53, we see a striking exclusion where there are no cell lines that fell into the -/-category. This suggests not every p53 mutation inactivates the gene expression of p53. Furthermore, in the literature, it has been shown that $BRAF^{V600E}$, which is known to be the most common BRAF mutation in cancer, correlated significantly with dedifferentiation and the suppression of CDX2, an important transcription factor for epithelial

differentiation in the intestine (Reischmann *et al.*, 2020). Interestingly, we found that BRAF mutation is also correlated with CDX1 in our analysis which is thought to share a similar role as a differentiation transcription factor with CDX2. In addition, previous studies have shown that BRAF mutated CRC tumors are associated with high levels of undifferentiated cells, suggesting the loss or downregulation in differentiation markers such as CDX1. Another gene EPDR1, encoding ependymin related 1 that plays a role in calcium-dependent cell adhesion, is found associated with BRAF mutations. It is also shown in previous studies that BRAF and TGFβR2 mutations are negatively associated with EPDR1 (Chu *et al.*, 2018).

The low percentage of genes found significantly associated with driver mutations suggested other possible mechanisms of regulation. We further investigated whether any bimodal distribution is due to a genetic mutation. We found that no genes are associated with a genetic mutation. This result is in fact in line with expectation since high-frequency mutations such as driver mutations had already been rigorously identified thus, we do not expect to find novel mutations with a high enough mutation that would result in a bimodal pattern.

4.2.4.2 Association with Methylation

Next, we explored the association against gene methylation retrieved from the CCLE database. The CCLE methylation data are quantified using Illumina methylation arrays. The output data for methylation is a matrix of β values ranging from 0 to 1, indicating the degree of methylation of a given region of a gene. Each of the CpG_sites_hg19 in Supplementary Table S1 represents a set of CpG islands and their chromosome location. The CCLE methylation data consists of 51 CRC cell lines. 28 cell lines overlapped with our microarray data and 50 cell lines overlapped with the CCLE RNA-seq data. There are two separate methylation datasets including

methylation data specific to CpG islands in promoter regions and methylation data specific to 1000kb upstream of the transcription starting site. These are two separate datasets downloaded separately as excel files. We separated our analysis into two sections. The first part focused on the association of methylation and well-defined bimodal genes. In the second part, we are interested in exploring novel associations between methylation and gene expression levels in all bimodal genes.

Methylation pattern on well-defined bimodal genes

5/64 (7.8%) and 3/82 (3.7%) of the well-defined bimodal genes are significantly correlated with methylation on the CpG islands within the promoter region in our microarray and CCLE cell line data, respectively. The data are shown in Table 4.8.

Immediately we noticed CDX1 which was previously shown in our lab to be controlled by direct methylation in the promoter region (Wong *et al.*, 2004). The expression of the homeobox protein CDX1 is often lost or reduced in a large percent of colorectal cancer tumors. The loss of differentiation and increased stemness in many tumors would lead to a decrease in the differentiation marker CDX1. Our analysis reinforces previous experimental work on CDX1 methylation. Another gene KDM5D, which encodes a protein containing zinc finger domains and regulates metalloproteinase family genes (MMP), is consistently associated with promoter methylation both in our microarray and CCLE cell line data (Li *et al.*, 2019). KDM5D is a male-specific gene that resides on the Y chromosome. It is shown to be a pro-metastasis gene that works as a demethylase of histone medication protein H3K4me2/3. Also, inhibition of KDM5D in male mice, as well as CRC cell line LOVO, has shown significant inhibition of migration capability

(Li *et al.*, 2019). An important observation to note is that KDM5D is upregulated when methylated. Many studies have reported hypermethylation leading to upregulation of its associated gene (Arechederra *et al.*, 2018; Rauluseviciute, Drabløs and Rye, 2020), which is in many ways counterintuitive to our understanding of methylation. The mechanism behind such observation is still unclear; however, here we observe an example of such a relationship with KDM5D which may provide useful insights and important implications in the future. The novel correlation with promoter methylation in the context of CRC has not been studied in the literature and is thus a novel relationship in need of further exploration.

In the analysis against methylation occurring 1000kb upstream of the transcription starting site, we found 2/64 (3.1%) and 1/82 (1.2%) significant associations in our microarray and CCLE cell line data, respectively (Table 4.8) KDM5D methylation is again significant in both bodies of data. In addition, HSPA1A, a gene under the heat shock protein family, has been experimentally shown to be hypermethylated in ovarian and bladder cancer cells (Ban *et al.*, 2019). Although HSPA1A has not been suggested in colorectal cancer, our analysis reveals a potential insight into the underlying mechanism of regulating HSPA1A.

a) Bodmer Lab Microarray Cell Line Data

Methylation at CpG islands in promoter region	+/+	+/-	-/+	-/-	p-value	R value
KDM5D	4	0	0	24	4.88E-05	1.00
CTFR--1	1	11	8	8	3.90E-02	-0.44
HSPA1A	7	3	18	0	3.66E-02	-0.46
USH1C--2	7	11	9	1	1.59E-02	-0.49
SLC38A5	8	7	13	0	6.88E-03	-0.54

Methylation 1kbp upstream of TSS of gene	+/+	+/-	-/+	-/-	p-value	R value
KDM5D	2	2	0	24	1.59E-02	0.68
HSPA1A	4	6	18	0	5.57E-04	-0.70

b) CCLE Cell line Data

Methylation at CpG islands in promoter region	+/+	+/-	-/+	-/-	p-value	R value
KDM5D	6	6	0	38	5.81E-05	0.66
CDX1	4	28	7	11	4.06E-02	-0.31
SLC38A5	16	10	23	1	5.05E-03	-0.41

Methylation 1kbp upstream of TSS	+/+	+/-	-/+	-/-	p-value	R value
KDM5D	4	8	0	38	2.15E-03	0.52

Table 4.8 Methylations Associated with Well-Defined Bimodal Genes: Table showing the gene and Bonferroni multiplicity corrected p-value of genes that fitted the well-defined bimodal distribution criterion. The table shows the genes that are significantly associated with methylation at the CpG islands in the promoter region and methylation 1kbp upstream of TSS. (a) shows the result in our microarray data and (b) shows the result in CCLE cell line data.

Methylation pattern on all bimodal genes

We found 105 bimodal genes significantly associated with methylation. Other than the well-defined bimodal genes found associated with methylation shown above, we also found other genes that we suspected associated with methylation. For example, in Chapter 5 we explored the role of methylation as a potential upstream regulator of EMT transcription factor ZEB1. In Supplementary Table S1, ZEB1 is shown to be negatively associated with a CpG island located on chromosome 10:31606675-10:31606814 near ZEB1 with a significant r-value of -0.81. Another gene we noticed is the methylation of CDH1. Interestingly, CRC cell line RKO and

Colo320DM are the only two cell lines that were methylated. This pattern of methylation fits with our experimental work shown in Chapter 5 where CRISPR knockout experiments of ZEB1, which was hypothesized to be inhibiting CDH1, failed to lead to a re-expression of CDH1 in Colo320DM. CDH1 methylation may thus be the reason behind the absence of CDH1 expression instead of ZEB1 promoter-binding. We will further discuss ZEB1 and CDH1 methylation in Chapter 5.

Other interesting genes include NCAM1, MMP14, MMP28, and KRT15, KRT16, KRT17, and KRT9. Downregulation of adhesion molecule NCAM1 is associated with poorer prognosis in CRC (Fernández-Briera *et al.*, 2010). In our result, SW480 and NCIH1716 are the only cell lines downregulated in NCAM1 as well as methylated in a CpG region close to the NCAM1 promoter region. Both cell lines are categorized as EMT cell lines, suggesting that the observation of NCAM1 in cells undergoing EMT is more linked to methylation changes instead of the direct inhibition from EMT transcription factors. The full list of significantly associated genes is listed in Supplementary Table S1.

4.2.4.3 Association with Enhancers

In the analysis of the association with enhancers (n = 28), we found that there are no significant associations in our microarray data and 3/82 (3.7%) significant associations found in the CCLE cell line data. One methylation-associated gene GJB1 is a gap junction gene that is responsible to form gap junction channels that facilitate the transfer of ions and small molecules between cells (Tomaselli *et al.*, 2017). Another, LINC00675, a long non-coding RNA that has been suggested to regulate cell proliferation and migration, and thirdly NLRP2, a gene that encodes a nucleotide-binding and leucine-rich receptor, are also shown to correlate with enhancer regulation in Table 4.19 (Tilburgs *et al.*, 2017; Ma *et al.*, 2018).

CCLE Cell line Data

Enhancers	+/+	+/-	-/+	-/-	p-value	R value
NLRP2	9	0	15	26	5.22E-03	0.49
GJB1	16	15	19	0	2.40E-04	-0.51
LINC00675	11	12	26	1	1.63E-04	-0.55

Table 4.9 Enhancers Associated with Well-Defined Bimodal Genes: The table shows the genes whose expression is significantly associated with enhancer methylation. There were no genes significantly associated in our microarray data and 3 genes shown in the table were significantly associated with the CCLE cell line data.

4.2.5 GMMchi-based Clustering

One powerful application of GMMchi is in clustering genes based on categorized values instead of the continuous value most are familiar with. The advantage of building a correlation matrix using categorical values instead of using continuous values is the ability to take a biological *a priori* into the analysis. For continuous data, since each number represents the expression level of a gene, there are no numbers specifically represent any particular biologically defined groups such as a background signal, non-expressing, low-expressing, and high-expressing group. Furthermore, clustering based on categorical values is a pattern-based clustering where any significant correlation between two genes originated from a similar pattern in cell line categorization is inherently different from a trend-based correlation seen in continuous datasets. Thus, this method provides new insights into understanding the relationship between genes and clusters of genes. GMMchi-based clustering will be analyzed in two parts: correlation matrix of genes and correlation matrix of gene clusters. To prevent confusion, we will from here on refer to clusters of genes as ‘gene clusters’ and clusters of gene clusters as ‘clusters’.

In brief, GMMchi-categorized microarray data is inputted into the built-in function ***df.corr*** from the *pandas* library. ***df.corr*** creates a correlation matrix from the categorized expression data.

Clustering is performed on the correlation matrix based on standard agglomerative clustering using ***sklearn.agglomerativeclustering***. The documentation for this package is found in:

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html.

Agglomerative clustering is a type of hierarchical clustering technique using a bottom-up approach. In this method, each observation is initially treated as an individual cluster. The clusters are subsequently merged by minimizing the linkage criteria. The distance between two genes in our case uses the Euclidean distance calculated based on the r-value between the two genes. Here, we used the default criteria which minimizes the sum of squared differences within all clusters. The clusters are extracted by clustering genes with the closest linkage distance returned from the output of agglomerative clustering using a distance threshold. The distance threshold is determined by applying GMMchi on the total distribution of linkage distances. The threshold is determined to be 28.8 for gene clusters and 7.2 for clusters as shown in Figure 4.12. All analysis is based on Python.

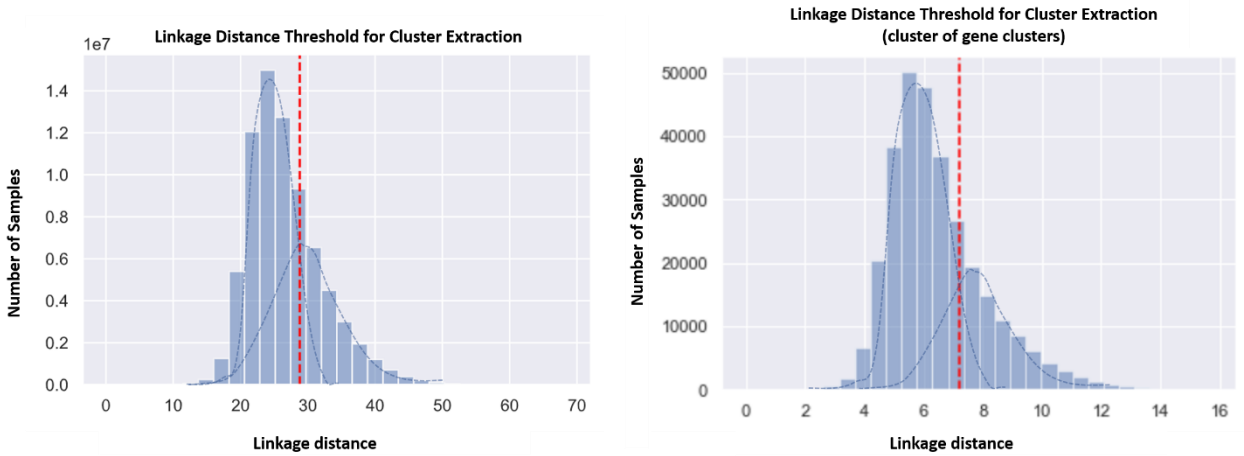


Figure 4.12 GMMchi Analysis on Identifying Threshold for Linkage Distance: Identifying the threshold used for cluster extraction in the clustering of genes (left) as well as the clustering of gene clusters (right). The red line separates bimodal patterns of the linkage distance when plotted out on a histogram via GMMchi.

4.2.5.1 Gene Clusters

Clustering was done on genes that did not exhibit a unimodal pattern. In other words, genes that had two distinct groups classified as one of the following were included for analysis: bimodal, unimodal with a tail, bimodal with a tail, and categorical unimodal patterns (discussed in Chapter 3). Using the distance threshold determined by GMMchi, we were returned with 1418 clusters. Next, to eliminate loose clusters, we took the following steps:

1) Remove clusters with 1 member (gene)
2) Remove clusters that exhibit a unimodal pattern
3) In any given cluster, identify and remove genes that fit the inclusion criterion

The modality as well as whether a gene is an inclusion to the cluster is determined in the following way:

1) Assign a phenotype to the cluster by adding up 1s and 2s for all the genes in the cluster for all cell lines. For example, if there are *m* genes in a cluster and *N* cell lines the then value ranges from *mN* to 2*mN*.
2) Apply GMMchi onto the total counts and recategorize them into 1s and 2s (lows and highs) and determine the modality of the distribution
3) Run 2x2 table analysis of each gene to the cluster phenotype to determine whether a gene is an inclusion to the cluster

We ended up reducing 1418 gene clusters to 777 tight gene clusters that were further studied for downstream analysis. The full list of genes in each gene cluster is shown in Supplementary Table S2.

To analyze clusters more systematically, we ran 2x2 tables between each cluster and all known phenotypes including mutation status, goblet cell formation, EMT, lumen formation, MF recruitment, and RER status. Only associations that were significant after multiplicity correction was of interest and shown in Table 4.10.

EMT was most positively associated with gene cluster 856 which is consisted of known EMT-related genes including ZEB1, VIM, MAPK12, and RGS4. Within this gene cluster, we also found previously unidentified genes including NAP1L5, IKBIP, and FAM101B. In the literature, NAP1L5, a gene under the Nucleosome assembly protein family, has been shown to be linked to aggressive and metastatic tumors in hepatocellular as well as pancreatic carcinoma (Schimmack *et al.*, 2014). Furthermore, IKBIP, a gene encoding Kappa B kinase interacting protein, and FAM101B, a gene encoding the protein Refilin B, have been implicated to promote invasion and metastasis in glioma and prostate cancer, respectively (Gordon *et al.*, 2017; Ting-Yu *et al.*, 2020). On the other hand, EMT phenotype was most negatively associated with gene cluster 48 consisted of epithelial-related markers including EPCAM, RAB25, AP1M2, OVOL1, and CDH1.

In this analysis, goblet cell-forming cell lines, to our surprise, are only significantly associated with gene cluster 402. In gene cluster 402, other than genes such as REG4, TFF1, and SLC9A2 that are thought to be important markers found in goblet cell differentiation, AKAP7, and PDE3B had not been previously linked to goblet cells. It has been previously shown that mucus

secretion is dependent on a Ca^{2+}-mediated pathway and that Ca^{2+} is an important component in forming MUC2. In the literature, AKAP7, the gene encoding A-kinase anchoring protein 7, and PDE3B, gene encoding phosphodiesterase 3B, both play a role in cAMP/PKA regulation of Ca^{2+} cycling. AKAP7, which is shown to play an important role in modulating cardiac function, is thought to target cAMP-dependent protein kinase (PKA) to the plasma membrane and further couples it to the Ca^{2+} channel (Lygren *et al.*, 2007; Jones *et al.*, 2012). On the other hand, PDE3B mediates Ca^{2+} by downregulating intracellular cAMP (Walz *et al.*, 2007; Chung *et al.*, 2015). Therefore, both genes may be playing a role in the Ca^{2+}-mediated pathway in mucus production, providing an interesting insight worth further exploration. We noticed that MUC2 was not associated with this phenotype and found that MUC2 belonged to gene cluster 23 where genes such as SPINK4 and FCGBP are also part of. We found that this gene cluster is significantly associated with this phenotype before multiplicity correction and that MUC2 is an inclusion of this gene cluster. As previously discussed, goblet cell-related genes such as FCGBP and SPINK4 (not shown) were an inclusion of MUC2, where MUC2-expressing cells were a subset of FCGBP-expressing cells. We suspect that many of these genes such as FCGBP and SPINK4 formed a gene cluster such that MUC2 is a subset of. In other words, the phenotype of the gene cluster revealed a pattern in line with FCGBP and SPINK4, thus did not result in a strong association with goblet cell-forming cell lines compared to MUC2 alone.

RER status, as expected, is associated with gene clusters consisting of genes that are previously identified to be differentially expressed in RER CRC cell lines. For example, gene cluster 1416 included genes such as STRN3, and CDC42SE1. In addition, many other known RER-related genes such as HNRNPL, MTA2, SMU1, USF1, TOR1AIP2, and CTNNB1 are classified as inclusions

of this gene cluster. This pattern coincides with what was described above where STRN3 and CDC42SE1 did not fit the inclusion criterion while the others mentioned did. Moreover, gene cluster 909 consisted of genes such as CASP2, LETM1, and HNRNPH1, and gene cluster 910 consisted of genes such as TOR1AIP2 and THRAP3 which were also, as expected, associated with RER status. Interestingly, both gene clusters themselves are an inclusion of RER status. This was an important observation because gene clusters 910 and 909, which are inclusions of RER status, is a different scenario from gene cluster 1416 which is not an inclusion of RER status but contains a list of known differentially expressed genes that are inclusions to the gene cluster. This further reveals a more sophisticated relationship between the genes and how they may potentially relate to one another.

Other phenotypes such as the WNT expression in cell lines and TGFBR2 mutation are also associated significantly associated with other gene clusters shown in Table 4.10. Here, we found that all of the gene clusters associated with TGFBR2 mutations are also gene clusters associated with RER status, reconfirming the strong correlation between the two phenotypes discussed previously. Besides, most of the gene clusters related to WNT expression are genes known to play an important role in differentiation.

Association of EMT with Gene Clusters

EMT	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Key Genes Associated
Gene cluster856	6	0	3	69	3.27E-07	0.80	Significant	ZEB1--3, VIM--2, IKBIP--1, MAPK12—2, FAM101B_weighted
Gene cluster43	0	6	62	10	3.12E-05	-0.57	Significant	RBM47_weighted, SSBP3_weighted, CD24_weighted, FGD3--2
Gene cluster47	1	5	71	1	1.69E-06	-0.82	No	CLDN7, TMEM125, GALNT3--1, PKP3_weighted, ST14_weighted, RAB17, KRT8--1, KRT19—1, OVOL2--2
Gene cluster48	0	6	71	1	2.73E-08	-0.92	No	AP1M2_weighted, OVOL1--2, RAB25, IRF6_weighted, DSC2--3, EPCAM, MAP7_weighted, ESRP1_weighted, TMEM159, CDH1_weighted

Association of Goblet Cell-forming Cell Lines with Gene Clusters

Goblet cell-forming	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Key Genes Associated
Gene cluster402	9	2	1	18	3.51E-05	0.78	No	REG4_weighted, TFF1, SLC9A2, AKAP7—1, ATOH1

Association of RER status with Gene Clusters

RER	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Key Genes Associated
Gene cluster1416	17	1	4	49	9.67E-12	0.83	No	STRN3--2, CDC42SE1—3, HNRNPL_weighted, MTA2, SMU1--1, TMEM33--5, USF1, BOD1L--3, RAB12--1, TOR1AIP2--4, CTNNB1--1, LYG1
Gene cluster909	17	1	12	41	7.13E-08	0.64	Significant	CASP2_weighted, DCAKD--1, LETM1--1, HNRNPH1_weighted
Gene cluster923	12	6	7	46	3.54E-05	0.53	No	TMEM19--4, KLHL17—1, METT10D--1, RPL23--2
Gene cluster1253	16	2	17	36	2.89E-05	0.50	Significant	FOXA1--1, TBC1D30_weighted, CYP1B1_weighted
Gene cluster910	18	0	24	29	1.33E-05	0.48	Significant	TOR1AIP2--6, THRAP3--1, FUT10--2
Gene cluster118	2	16	36	17	2.89E-05	-0.50	Significant	KPNA1--3, ARHGEF7—2, DHRS12--1, RAB11FIP4--3
Gene cluster85	5	13	44	9	3.28E-05	-0.52	No	PRPF38B--2, C2orf67--4, SOX6--8, ZNF224--1
Gene cluster116	0	18	33	20	1.11E-06	-0.54	Significant	NFIB--8, C1orf56--1, TMEM8B--1
Gene cluster705	1	17	39	14	3.48E-07	-0.60	Significant	SPAG4, TAF9B--5, PDP2, GNL3L--4, PTCD2--2
Gene cluster702	1	17	40	13	1.61E-07	-0.62	Significant	CAB39L--3, GLS--2, SHANK2—3, MAP3K12--2, IGFBP4
Gene cluster701	1	17	42	11	3.01E-08	-0.66	Significant	ZMYM2--2, TXNDC9_weighted, NR1D2--1, RSF1--4
Gene cluster703	1	17	42	11	3.01E-08	-0.66	Significant	CFLAR_weighted, CSRNP1, SMAP1, TRIM13--3
Gene cluster707	1	17	48	5	4.18E-11	-0.80	No	RBM6—3, MDM4--5, IAH1--2, HNRNPH1--4, TOB1--2, CTTN--6, SLC38A1--1

Association of TGFBR2 Mutation with Gene Clusters

TGBR2 mutation	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Key Genes Associated
Gene cluster1416	13	3	3	36	1.75E-07	0.74	No	STRN3--2, CDC42SE1—3, HNRNPL_weighted, MTA2, SMU1--1, TMEM33--5, USF1, BOD1L--3, RAB12--1, TOR1AIP2--4, CTNNB1--1, LYG1
Gene cluster909	13	3	7	32	1.78E-05	0.60	No	CASP2_weighted, DCAKD--1, LETM1--1, HNRNPH1_weighted
Gene cluster701	2	14	29	10	4.21E-05	-0.57	No	ZMYM2--2, TXNDC9_weighted, NR1D2--1, RSF1--4
Gene cluster703	3	13	31	8	4.32E-05	-0.57	No	CFLAR_weighted, CSRNP1, SMAP1, TRIM13--3
Gene cluster707	3	13	37	2	3.53E-08	-0.78	No	RBM6—3, MDM4--5, IAH1--2, HNRNPH1--4, TOB1--2, CTTN--6, SLC38A1--1

Association of WNT-expressing Cell Lines with Gene Clusters

WNT-expressing	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Key Genes Associated
Gene cluster40	66	3	3	6	2.46E-05	0.62	No	ARID5B--3, ZNF703--1, NEBL--1, SYT7--3, HPDL, DUSP16--3
Gene cluster15	53	16	0	9	1.12E-05	0.53	Significant	UGT1A1 /// UGT1A10 /// UGT1A3 /// UGT1A4 /// UGT1A5 /// UGT1A6 /// UGT1A7 /// UGT1A8 /// UGT1A9--3, PIP5K1B--1, POF1B_weighted, FA2H_weighted, FUT3_weighted
Gene cluster4	57	12	1	8	4.10E-05	0.52	Significant	LGALS4, AKR1C2_weighted, GPX2--1, SEMA4G--2, CDH17, CLDN3--1, FLJ46875, CLRN3, HEPH--2, FOXA3, CDX1, GPA33, GUCY2C
Gene cluster21	50	19	0	9	3.79E-05	0.48	Significant	MECOM_weighted, ARL14, ACSM3_weighted
Gene cluster13	50	19	0	9	3.79E-05	0.48	Significant	ATP8B1--2, BCL2L14--3, PPP1R1B, CEACAM5--1, FUT2_weighted, CEACAM6_weighted, AGR2_weighted, ACSL5_weighted, VILL,
Gene cluster1169	20	49	9	0	5.49E-05	-0.47	Significant	COL27A1_weighted, MLLT11, NRD1—2, GM2A--5, FRMD4A_weighted, EEPD1_weighted, PHLDB3--1, BMP6--1
Gene cluster873	9	60	8	1	8.26E-06	-0.59	Significant	HOXB2, EFNA5_weighted, FAM129A_weighted

Table 4.10 Gene Clusters Associated with Cell Line Phenotype: Table showing gene clusters that are significantly associated with phenotypes including EMT, goblet cell-forming cell lines, TGFBR2 mutation, and WNT-expressing cell lines.

4.2.5.1 Cluster of Gene Clusters

To further understand how different gene clusters relate to one another, we performed the same cluster extraction using the phenotype (1s and 2s) assigned to each gene cluster, remove clusters with < 1 member (gene cluster), remove gene clusters that are inclusions of the cluster, and remove clusters that exhibit a unimodal pattern. We resulted in 181 clusters which are reduced to 44 clusters after eliminating loose clusters. The full list of gene clusters in each cluster is shown in Supplementary Table S3. The correlation matrix of gene clusters is represented as a heat map in Figure 4.13. The heat map is presented in the range of r-values from -1 to 1 colored from dark blue to bright red.

Here, we once again systematically ran associations between clusters and known phenotype mentioned in the previous section. EMT was associated with cluster 93 consisting of gene clusters 809, 856, 863, 880, and 862. We found that gene cluster 856, a previously identified gene cluster associated with EMT that includes genes ZEB1 and VIM, formed a larger cluster with other gene clusters that were previously unidentified to exhibit a strong association with EMT. Within the other 4 gene clusters, under cluster 862, we found genes such as Mannose receptor C type 2 (MRC2) and dual-specificity phosphatase-1 (DUSP-1) both found to be correlated with promoting tumor invasiveness and aggressiveness (Sinha *et al.*, 2013; Shen *et al.*, 2016). Moreover, other genes such as LGALS1, ZNF510, and ETV1 are also found to be associated with EMT (Li *et al.*, 2013; Bacigalupo *et al.*, 2015; Gualtero and Suarez Castillo, 2016). On the other hand, EMT was negatively associated with clusters 19 and 22 which consisted of gene clusters 2, 3, 39, and 22, 48, respectively. Previously identified epithelial marker-associated cluster 48 is, again, retained in the larger cluster.

In the association with goblet cell formation, we again see that gene cluster 402 is retained in the larger cluster 115, which also included gene clusters 1036 and 1317. Under this larger cluster, other genes that are known to be associated with goblet cells not included in cluster 402 emerged in both clusters. These genes include SPDEF, UGT2B15, and SCG5. Moreover, other positively associated clusters 3, 9, and 4 included genes previously known to be related to goblet-cell formation such as KRT20, MUC1, SEPRPINB1, AGR3, AG42, GALNT12, CA12, CTFR, UGT family, and FUT3.

One important observation from this analysis is the ability to identify genes previously not found to be highly associated with a given phenotype. Here, we were able to identify novel genes based on the association with the overall phenotype of a given cluster, thus providing a powerful insight in identifying novel genes that may be functionally relevant in the cluster they belong to. This not only allows novel gene exploration as well as a bigger picture of the tight cluster each gene belonged to, the relationship and biological context the gene has with other members of the cluster, and the relationship of the gene cluster to other gene clusters within the larger cluster. Other phenotypes such as the WNT expression in cell lines, TGFBR2 mutation, p53 mutation, and RER status are also associated significantly associated with other clusters shown in Table 4.11. Once again, clusters associated with TGFBR2 mutation are clusters associated with RER status, suggesting a striking relationship in the genes controlled by the two phenotypes.

Association of EMT with Clusters

EMT	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Clusters Associated
Cluster93	5	1	5	67	6.75E-05	0.61	No	809, 856, 863, 880, 862
Cluster19	1	5	61	11	1.09E-03	-0.45	Significant	2, 3, 39
Cluster22	1	5	71	1	1.69E-06	-0.82	No	22, 48

Association of Goblet Cell-forming Cell Lines with Gene Clusters

Goblet cell-forming	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Clusters Associated
Cluster115	11	0	5	14	8.66E-05	0.71	Significant	402, 103, 131
Cluster3	11	0	5	14	8.66E-05	0.71	Significant	15, 65, 1, 25, 4, 71, 72, 69, 70
Cluster9	11	0	6	13	3.17E-04	0.67	Significant	14, 51, 473, 44, 37, 474
Cluster4	11	0	7	12	6.05E-04	0.62	Significant	16, 13, 17, 22, 12, 5, 18, 20, 19, 439, 8, 440
Cluster98	1	10	15	4	3.80E-04	-0.67	No	890, 1116, 835, 869, 893, 895, 841, 968, 1198, 894

Association of RER status with Gene Clusters

RER	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Clusters Associated
Cluster83	15	3	8	45	2.82E-07	0.63	No	1416, 1251, 1153, 1007
Cluster103	17	1	22	31	6.96E-05	0.46	Significant	910, 845, 909, 923, 736, 846
Cluster32	8	10	49	4	5.64E-05	-0.52	No	597, 391, 640, 122, 116, 705

Association of TGFBR2 Mutation with Gene Clusters

TGBR2 mutation	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Clusters Associated
cluster83	10	1	7	32	1.75E-05	0.64	Significant	1416, 1251, 1153, 1007
Cluster103	11	0	15	24	2.74E-05	0.51	Significant	910, 845, 909, 923, 736, 846
Cluster32	4	7	36	3	3.06E-04	-0.58	No	597, 391, 640, 122, 116, 705

Association of p53 Mutation with Gene Clusters

P53 mutation	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Clusters Associated
Cluster83	13	45	10	3	3.68E-04	-0.45	No	1416, 1251, 1153, 1007

Association of WNT-expressing Cell Lines with Gene Clusters

WNT-expressing	+/+	+/-	-/+	-/-	p-value	r value	Inclusion/exclusion	Clusters Associated
Cluster19	60	9	2	7	0.000123	0.51	Significant	2, 3, 39
Cluster3	48	21	0	9	7.85E-05	0.46	Significant	15, 65, 1, 25, 4, 71, 72, 69, 70
Cluster9	45	24	0	9	0.000211	0.42	Significant	14, 51, 473, 44, 37, 474
Cluster92	25	44	9	0	0.000288	-0.41	Significant	1196, 1203, 881, 873, 1135
Cluster97	19	50	8	1	0.000647	-0.41	Significant	871, 874, 891
Cluster98	22	47	9	0	0.000111	-0.44	Significant	890, 1116, 835, 869, 893, 895, 841, 968, 1198, 894

Table 4.11 Clusters Associated with Cell Line Phenotype: Table showing clusters that are significantly associated with phenotypes including EMT, goblet cell-forming cell lines, TGFBR2 mutation, p53 mutation, and WNT-expressing cell lines.

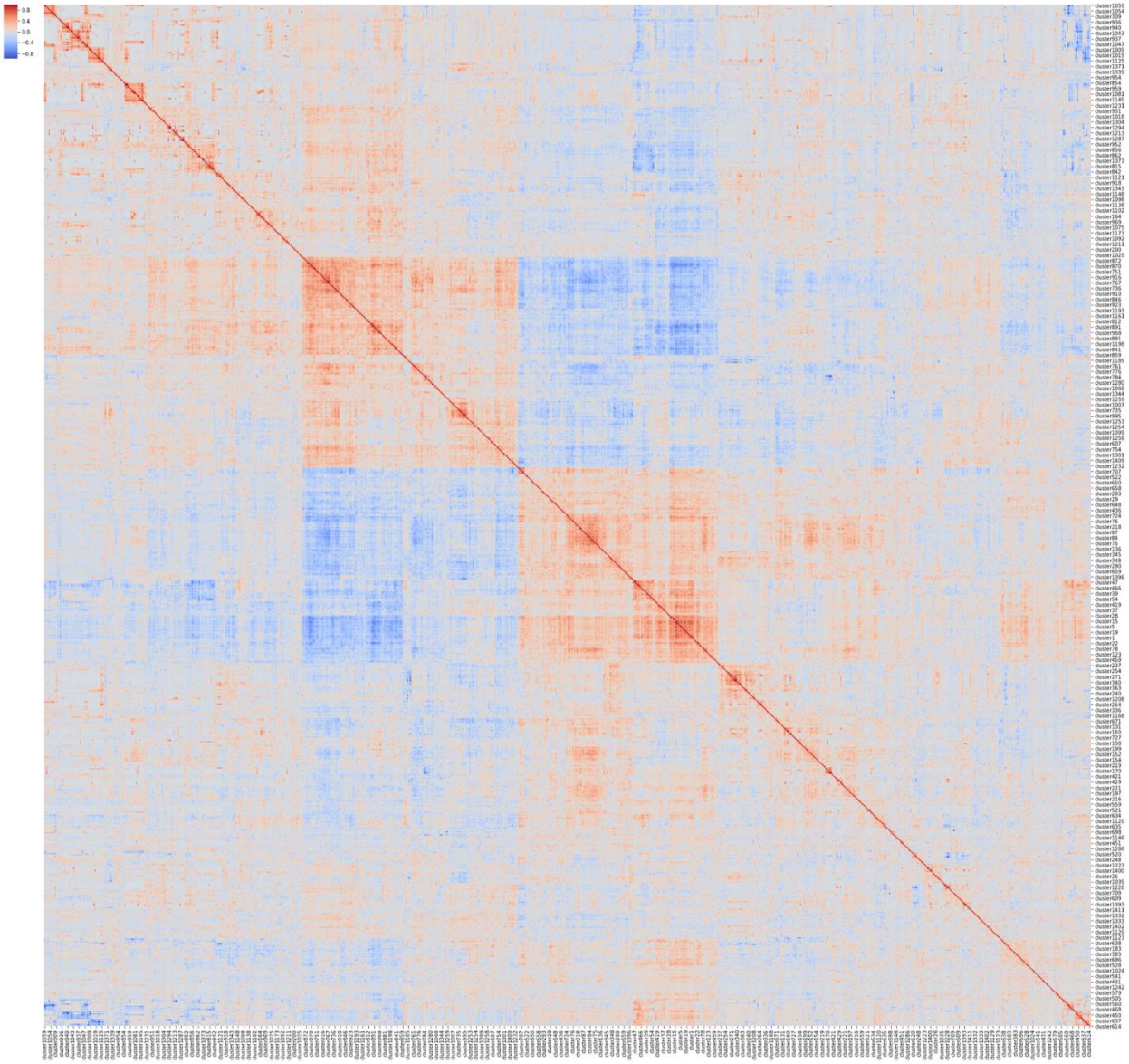
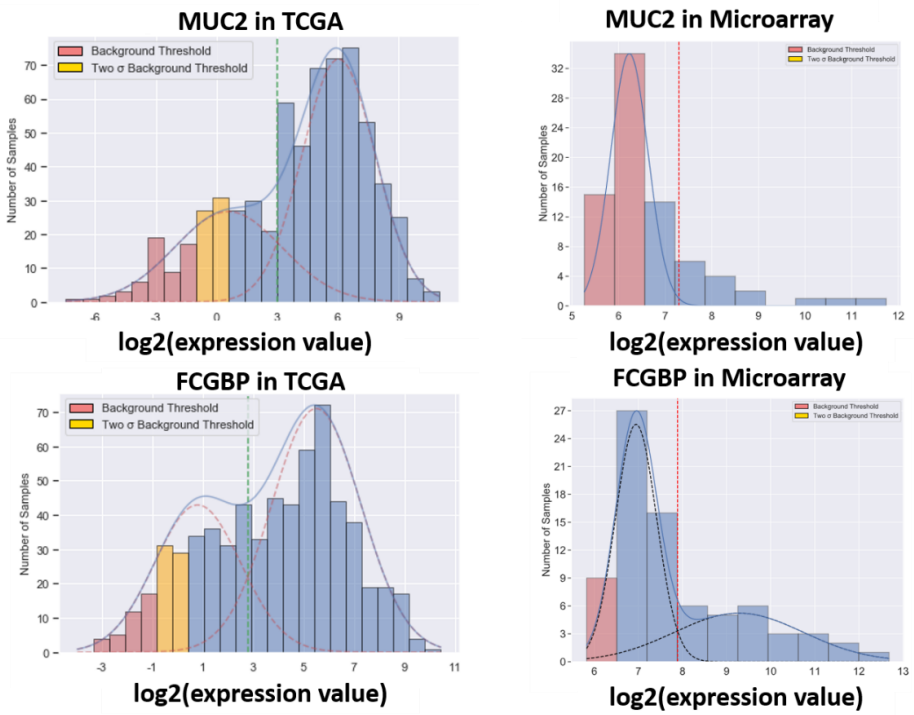


Figure 4.13 Heatmap of the Correlation Matrix Between Gene Clusters: Shows the heatmap of the correlation matrix between gene clusters. The range of r-value ranges from -1 to 1 shown in the legend.

4.2.6 GMMchi-based TCGA Purification

The Cancer Genome Atlas (TCGA) is a vast and diverse database that is consists of a variety of data on patient tumor samples including gene mutation, expression, methylation, and clinical data. One of the biggest faults of the TCGA lies in the inclusion criterion for the clinical samples. Essentially, the tumor sample only needs to be consist of ≥ 80% tumor nuclei, leaving up to 20% of patient samples subjected to non-tumor tissue which may include normal healthy tissue, myofibroblasts, blood cells, immune cells, and endothelial cells (Aran, Sirota and Butte, 2015). Unfortunately, most researchers analyze TCGA data assuming that it consists of a pure tumor sample without realizing the non-tumor cell type contamination that can greatly affect any downstream analysis. Previously in our lab, we have identified that in the TCGA there is an overwhelming population of myofibroblast contamination as seen in Figure 4.14b based on the standard GMMchi analysis. In addition, we can also see that there is a high level of goblet cells in the TCGA data that potentially originated from healthy tissue around the tumor in Figure 4.14a. To address these issues, we established a model to remove non-tumor expression from the bulk expression TCGA data. In total, we have downloaded 689 samples from the TCGA on colorectal cancer. Initially, we ran a GMMchi analysis on the flattened gene expression level of all genes across all patient samples to determine the background threshold of the TCGA CRC dataset shown in Figure 4.15.

a)


MUC2 in TCGA
Background Threshold
Two σ Background Threshold
Number of Samples
log2(expression value)
MUC2 in Microarray
Background Threshold
Two σ Background Threshold
Number of Samples
log2(expression value)
FCGBP in TCGA
Background Threshold
Two σ Background Threshold
Number of Samples
log2(expression value)
FCGBP in Microarray
Background Threshold
Two σ Background Threshold
Number of Samples
log2(expression value)

b)

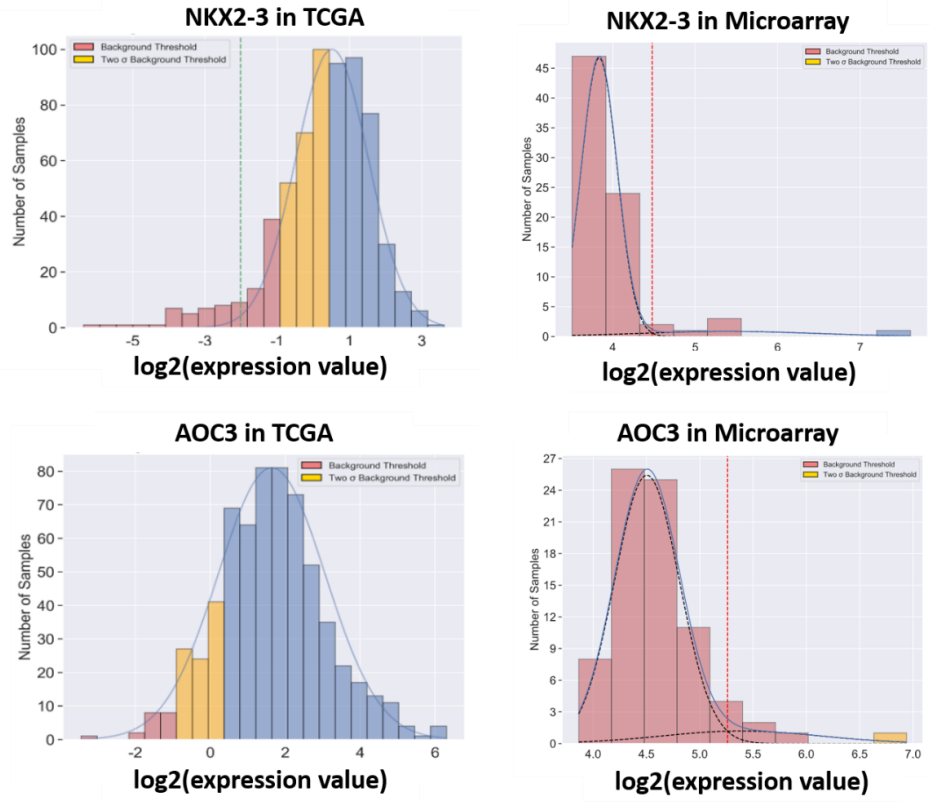

NKX2-3 in TCGA
Background Threshold
Two σ Background Threshold
Number of Samples
log2(expression value)
NKX2-3 in Microarray
Background Threshold
Two σ Background Threshold
Number of Samples
log2(expression value)
AOC3 in TCGA
Background Threshold
Two σ Background Threshold
Number of Samples
log2(expression value)
AOC3 in Microarray
Background Threshold
Two σ Background Threshold
Number of Samples
log2(expression value)

Figure 4.14 Comparison of Gene Expression Distributions in Our Microarray Data vs TCGA: (a) Shows a parallel comparison of MUC2 and FCGBP, markers of goblet cells, in the TCGA and microarray. The percentage of cell lines/patient samples in each low and high group is strikingly different suggesting the possibility of substantial non-tumor tissue contamination. (b) NKX2-3 and AOC3 are expressed in a large group of patient samples in the TCGA data, while not expressed at all in the tumor epithelial CRC cell line data. Note that although both genes were identified as a bimodal distribution via GMMchi analysis, all categories fell under the background threshold, and thus both genes are not expressed in any of the cell lines (except 1 cell line on the right in NKX2-3).

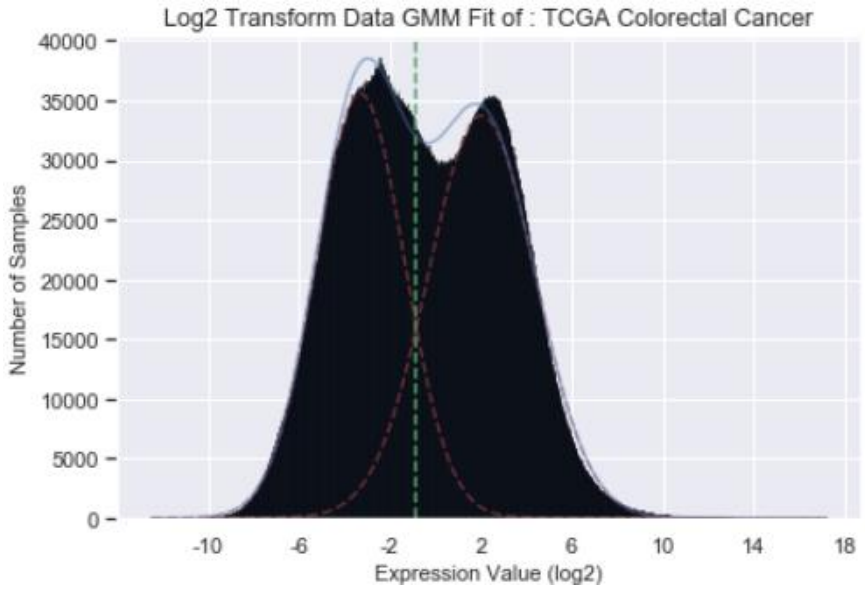


Figure 4.15 Determining the Background Threshold for TCGA RNA-seq Data: Graph shows the distribution of the expression levels of all genes combining all 637 CRCs patient samples from the TCGA. The green line delimits where the background noise threshold is that separates the background noise distribution (left) and the true positive distribution (right).

4.2.6.1 TCGA Purification

The TCGA purification pipeline goal is to remove contamination originating from non-tumor contamination in RNA-seq data from the TCGA. In the TCGA, sets of paired normal/cancer data are utilized to create a model for non-tumor cell-type expression. To do this, we will need to make the following assumptions:

1) The cell-type composition of the paired normal expression data is equivalent to the cell-type composition of non-tumor tissue in tumor expression data.
2) Copy number variation (CNV) provides an accurate approximation of tumor to normal tissue ratio since copy number variation is essentially observed in all the CRC tumors while none in normal tissue.

The model is then represented by Equation (6):

$$\boldsymbol{C = T - N \times (1 - p_{cancer})} \quad \textbf{(6)}$$

C = tumor expression

T = total expression (tumor + non-tumor)

N = non-tumor expression

p_{cancer} = percentage of tumor expression out of the total expression

The total expression *T* represents the individual expression level of each patient sample obtained from the tumor data, while the percentage of tumor expression out of the total expression p_{cancer} is estimated from the extent of copy number variation in each tumor sample. p_{cancer} is a constant indicating the purity of each tumor sample extracted from a study done by *Aran et al.* where they studied the tumor purity using copy number variation as well as other methods on data available in the TCGA (Aran, Sirota and Butte, 2015). The non-tumor expression *N* is calculated by finding the mean estimate by applying GMMchi on each gene on the paired normal sample downloaded from TCGA (Figure 4.16). Out of the 689 TCGA CRC samples downloaded, 51 samples are paired normal samples thus we ended up having 637 tumor samples

and 51 non-tumor normal samples. We inputted all the known variables into our model and retrieved tumor expression *C*.

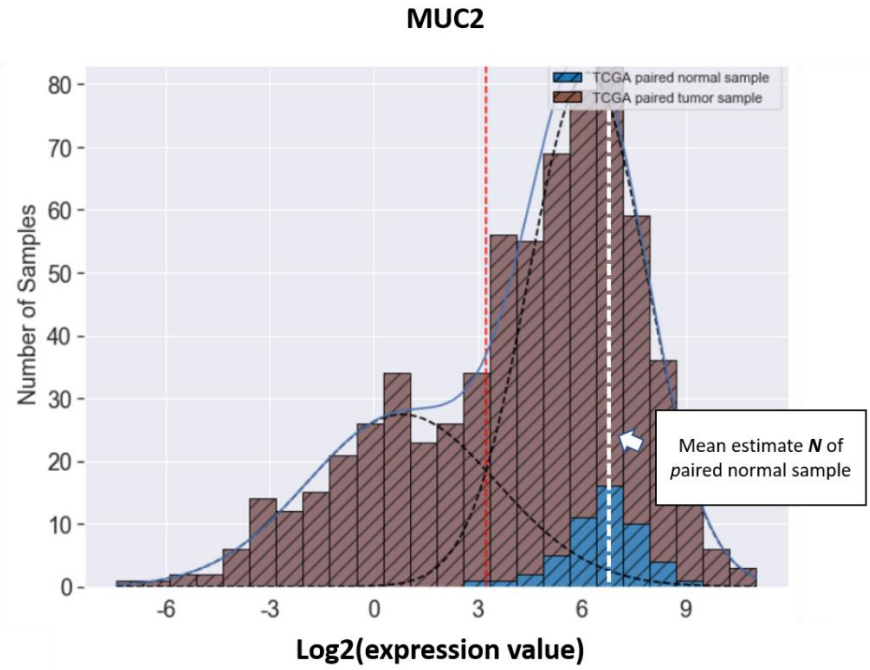


Figure 4.16 Example of MUC2 Gene Expression Distribution in TCGA: Shows an example of the expression distribution of MUC2 combining paired tumor and normal sample. The white line is the mean estimate *N* of the paired normal sample (n = 51) used for determining the average expression level of MUC2 in non-tumor sample.

In Figure 4.17, we show comparisons of gene expression before and after purification. AOC3 and NKX2-3 are two important biomarkers of myofibroblasts that line the colonic crypt. Since AOC3 and NKX2-3 are not expressed in epithelial cancer tumors, we would not expect any expression in pure tumor data. In the TCGA data, AOC3 and NKX2-3 are expressed in nearly 33% and 71%, respectively, of tumor samples. After purification, we see a massive reduction in expression to nearly no expression in all samples for both AOC3 and NKX2-3. Moreover, prior to purification, MUC2 expression is detectable in nearly 63% of all 672 TCGA colorectal patient

tumors; however, both the literature and our cell line data have shown that colorectal cancer tends to de-differentiate thus retaining much less differentiated cell types. In our microarray data, we only detected MUC2 expression in 23% of our cell lines. After purification, the percentage of MUC2 expressing samples in purified tumor sample went down to 30%, which is similar to that observed in our microarray data. Lastly, we investigated housekeeping genes such as GAPDH and ACTB where we would expect a minor reduction in expression level but not a complete removal of the signal. As expected, we have demonstrated that all the tumor samples still express housekeeping genes after purification thus showing that the model is not over-reducing signal from non-tumor tissue. The slight shift of expression level is only proportional to the removal of expression level contributed by non-tumor contamination.

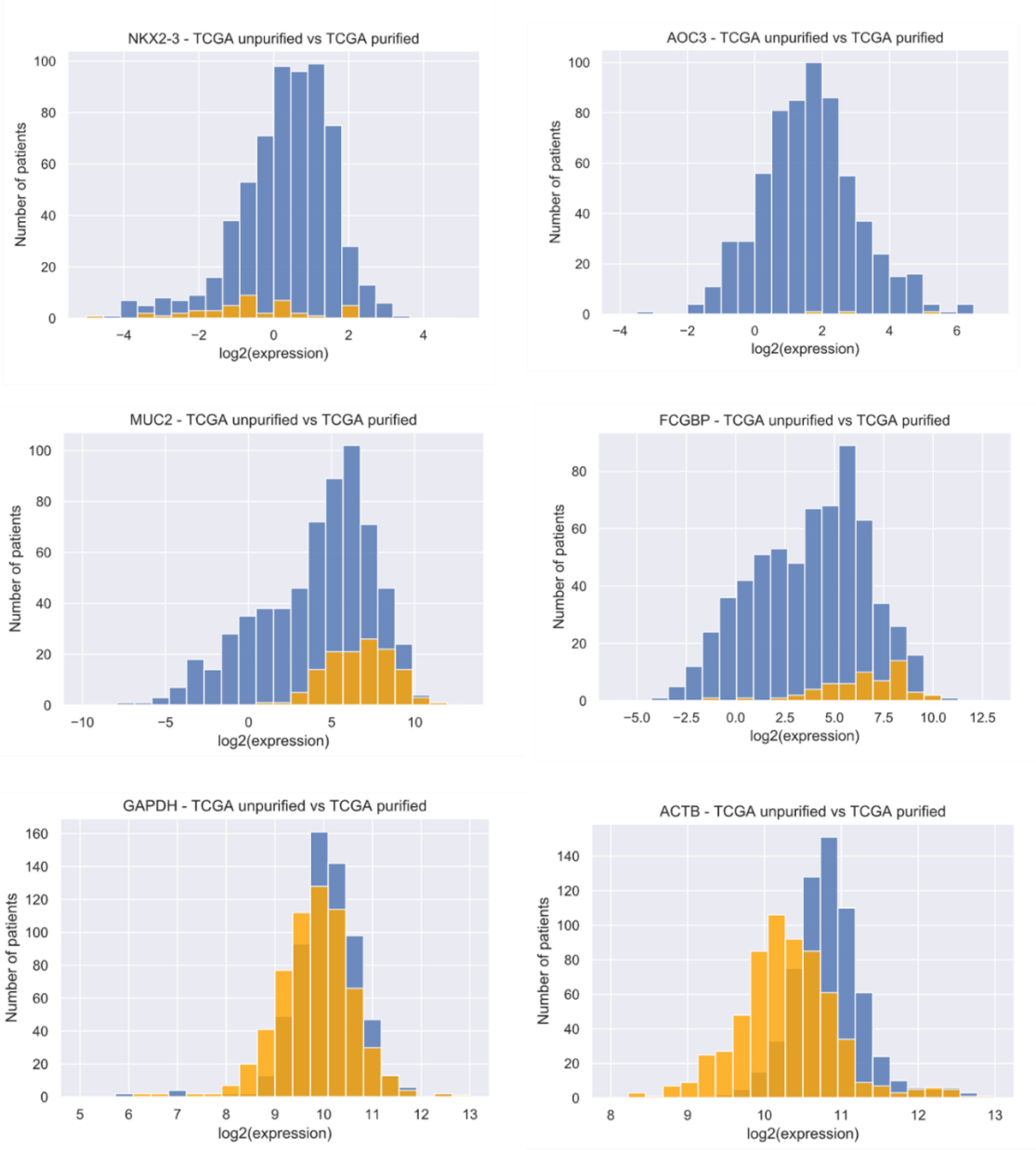


Figure 4.17 Examples of Before and After Purification: In all the examples, blue labels the original expression data before purification and, orange labels the expression after purification. As shown, both myofibroblast marker AOC3 and NKX2-3 have close to no detectable expression after purification. Also, MUC2, a goblet cell-specific biomarker, and FCGBP returned to levels closer to pure epithelial epithelium expression. GAPDH and ACTB revealing only reduced but not ablated levels of expression suggest the pipeline is not over-filtering non-tumor expression. It is important to note that the graph only shows patients with expression levels > 0, so patient samples where the expression level have reduced to zero after purification will not appear on the graph. The goal is to only visualize patients with expression levels > 0. Thus, a decrease in the total number of orange bars vs blue bars indicates a reduction in contaminated patient samples.

Chapter 5: Methylation Controls EMT Transcription Factor ZEB1

5.1 Introduction

Cancer cells undergoing epithelial-to-mesenchymal transition (EMT) have long been observed in various formats and it has been suggested that EMT plays a role in cancer metastasis. EMT transcription factors, including especially SNAI1, SNAI2, SNAI3, TWIST1, ZEB1, ZEB2 have been shown to be upregulated and to regulate the process of EMT in many cancer models (Garg, 2013; Stemmler *et al.*, 2019). These EMT TFs inhibit epithelial-like cancer cells from adhesion molecules and allow them to detach from their neighboring niche and migrate to a distant organ (Banyard and Bielenberg, 2015). However, it is important to note that whether EMT is the primary path cancer cells escape from their primary niche and metastasize requires further investigation. Nonetheless, this makes EMT an inherently important subject of interest in studying tumor cell migration, cellular plasticity, and anti-apoptotic properties conferred in a mesenchymal-like cell type (Aiello *et al.*, 2018; Lu and Kang, 2019).

The hallmark biomarkers for EMT are the expression loss of CDH1, the gene encoding adhesion protein E-cadherin, and the gain in expression of VIM, the gene encoding a type III intermediate filament protein that is commonly found in mesenchymal cells such as fibroblasts (Schliekelman *et al.*, 2015; Cheng *et al.*, 2016; An, Zheng and Dann, 2017). In the literature, it has been shown that the loss of CDH1 expression is due to the directing binding of EMT TFs such as SNAI, TWIST, and ZEB family onto the promoter region of CDH1 (Serrano-Gomez, Maziveyi and Alahari, 2016). Thus, the loss of CDH1 expression is not to be confused with expression lost due to mutation in the CDH1 gene. This important distinction needs to be well-accounted for during further investigation in this chapter. Also, it is noteworthy that most of the EMT TFs are zinc-

finger binding proteins that bind onto a specific region known as the TATA box located in the promoter region of many genes (Dykes *et al.*, 2017).

Furthermore, several upstream regulators of EMT TFs have been implicated in major pathways including transforming growth factor-beta (TGFβ), WNT, Notch, Hedgehog, and the hypoxic signaling pathway (Zhang, Tian and Xing, 2016; Zhu *et al.*, 2018). Most upstream regulators are shown to trigger a cascade of signaling pathways that led to the nuclear translocation and activation through direct binding onto the EMT TFs promoter region. In the literature, it has also been shown that mi-RNAs can directly bind and inhibit EMT TFs as a negative feedback regulation (Title *et al.*, 2018).

In this chapter, we explored our panel of 70 CDH1 wildtype Colorectal cancer cell lines using in-house microarray data to identify EMT-like phenotypes. By using cell lines as a model for exploring the EMT phenotype, we applied GMMchi as the basis of understanding genes upregulated in EMT-like cell lines, as preluded in Chapter 4. We observed ZEB1, a member of the ZEB family of EMT TFs, as the only EMT TF that is consistently upregulated in all the EMT-like cell lines which lead to our investigation in the regulation of the mesenchymal-like phenotype using CRISPR-knockout models of ZEB1. Furthermore, we are interested in the upstream regulator of ZEB1 expression in CRC cell lines and discovered a novel mechanism through epigenetic methylation in CpG islands proximal to ZEB1.

5.2 Results

5.2.1 Identifying EMT cell lines through GMMchi

141 cell lines are screened for CDH1 mutation and those carrying such mutations are removed from the initial screening to prevent misinterpretation of CDH1 downregulation due to genetic mutation instead of EMT TF (Figure 5.1). The 141 cell lines include duplicate cell lines as well as clones. Other than CDH1-mutated cell lines, both duplicate cell lines and clones are removed to remove any bias in the expression distribution due to their similar properties with existing cell lines. In Figure 5.1, cell lines with mutations in CDH1 are labeled with 'Homo', 'Het M', and 'FS' which refers to homozygous mutation, heterozygous mutation, and frameshift mutation. The 'Dup' represents duplicate cell lines and 'clone' represents clones of existing cell lines. We remove CDH1 mutated cell lines, clones as well as duplicate cell lines, returning us 70 CRC cell lines available for further analysis.

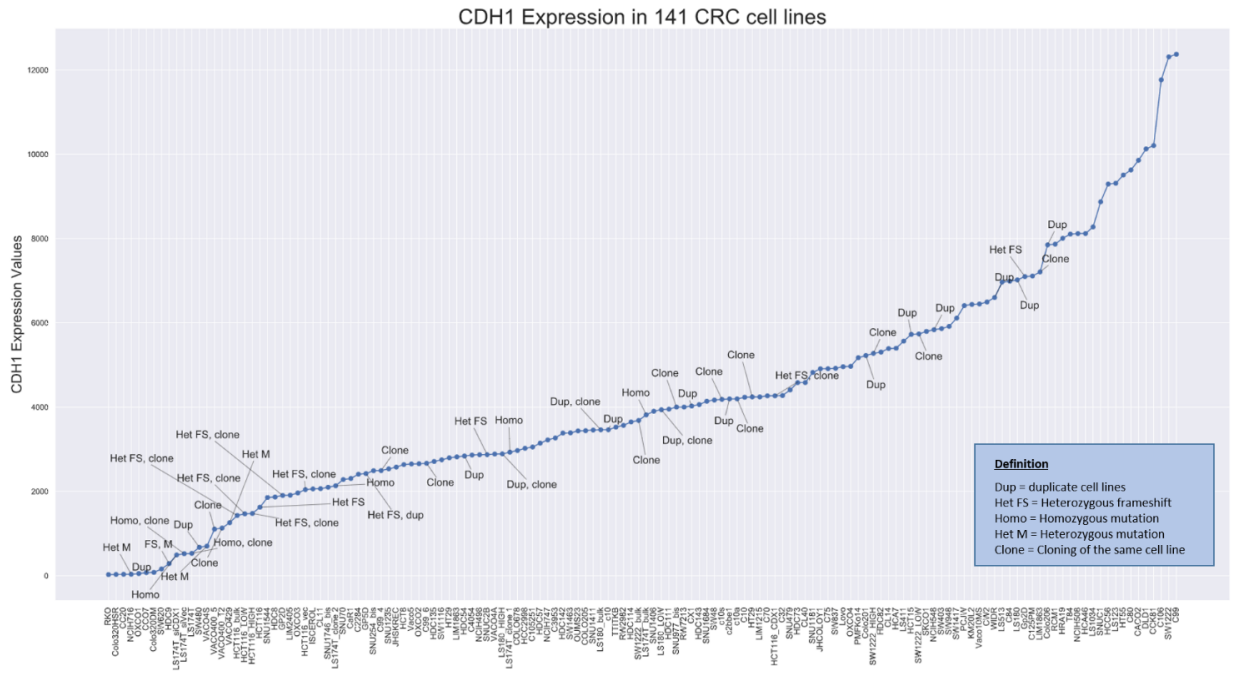


Figure 5.1 CDH1 Expression of 141 CRC Cell Lines: The CDH1 expression of 141 Colorectal cancer cell lines generated from our in-house microarray data are plotted from low to high. Duplicate cell lines, clones as well as cell lines with CDH1 mutations are labeled on the graph. Cell lines that fall onto any one of these conditions are removed from downstream analysis.

In Figure 5.2, we show the distribution of both CDH1 and VIM mRNA expression. As seen in Figure 5.2a, there exists a group of cell lines in an orange dashed box that represents a significantly low level of CDH1 expression level, which is also where the cutoff delineates the two groups in the GMMchi analysis in Figure 5.3c. On the other hand, in Figure 5.2b, the orange dashed box reveals a group of cell lines with a significantly higher VIM expression, also evident in Figure 5.3a where the cutoff is drawn to delineate the two groups. In our microarray data, 5 out of 70 Colorectal (CRC) cell lines exhibit hallmarks of EMT markers: upregulation of vimentin and downregulation of E-cadherin (Figure 5.3). 2x2 table analysis of all 70 CRC cell lines shows a highly significant negative correlation between the two genes (VIM and CDH1). Those cell lines that express low CDH1 and high VIM include Colo320DM, CCO7, CC20, NCIH716, SW620, and

OXCO1. These cell lines that exhibit EMT hallmark biomarkers will thus be referred to as 'EMT cell lines' for this thesis. Here, we observe two unique patterns with RKO having a low CDH1 and low VIM expression as well as LOVO and PMFK014 having a high CDH1 and high VIM expression. Since our focus is on identifying EMT cell lines, we did not explore them further. They suggest other potential mechanisms that allow the co-upregulation or co-downregulation of CDH1 and VIM.

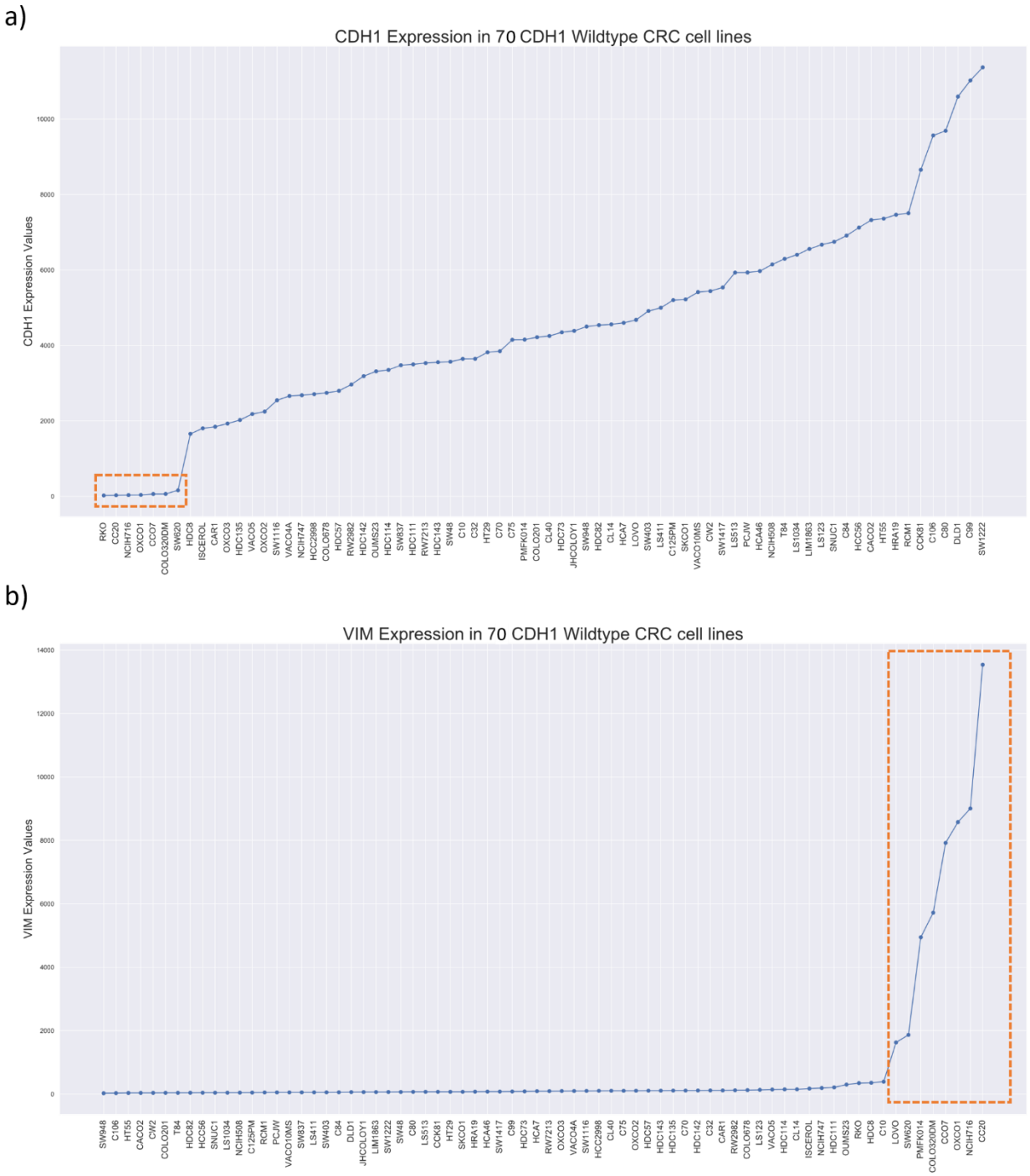


Figure 5.2 CDH1 and VIM Expression on 70 CDH1 Wildtype CRC Cell Lines: (a) CDH1 expression in 70 CDH1 wildtype CRC cell lines. The orange box delineates a group of cell lines that has a significantly lower CDH1 expression. (b) VIM expression in 70 CDH1 wildtype CRC cell lines. The orange box delineates a group of cell lines that has a significantly higher VIM expression.

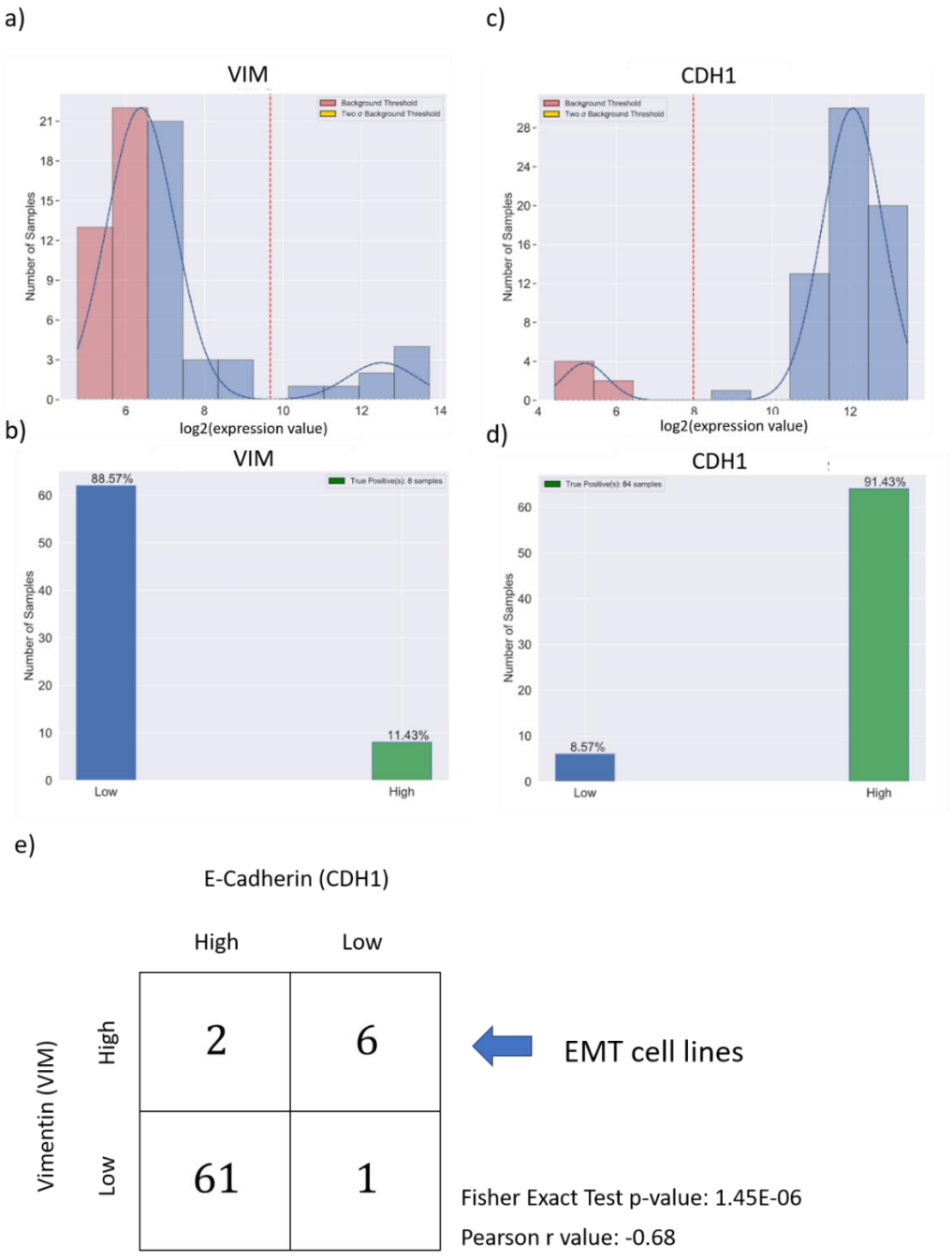


Figure 5.3 GMMchi and 2x2 Table Analysis on CDH1-/VIM+ CRC Cell Lines: (a) and (c) show the visual representation of the GMMchi analysis on VIM and CDH1, respectively; (b) and (d) show the percentage of cell lines within each low-expressing and high-expressing group in VIM and CDH1, respectively. (e) A 2x2 table summarizing the number of cell lines that fall into each category. The category with low CDH1 and high VIM is termed 'EMT' cell lines. CRC cell line RKO falls into the category of low CDH1 and low VIM, while LOVO and PMFK014 fall into the category of high CDH1 and high VIM, suggesting other potential patterns worth exploring.

5.2.2 Identifying candidate genes

Gene expression of EMT cell lines and non-EMT cell lines are compared with differential expression analysis using fold change and p-value to understand upregulated and downregulated genes between EMT and non-EMT groups. The p-value is calculated by a two-tailed independent t-test with equal variance corrected with Bonferroni correction for multiplicity comparisons. Only genes with a p-value ≤ 0.05 are considered statistically significant and included for further analysis (Table 5.1). The $\log_2$(fold change) is plotted against the $-\log_{10}$(p-value) for a volcano plot that provides easy visualization of genes that are both statistically significant and differentially up-or downregulated across the two groups (Figure 5.4). We imposed a criterion of only showing genes that are both below a p-value threshold of 10^{-8} and a fold change greater than $2^{1.5}$. To explore whether any EMT transcription factors are potentially responsible for the EMT signature observed within the previously defined EMT cell lines, we looked for SNAI families, TWIST families, and ZEB families within the list of statistically significant upregulated genes. We observed an extraordinary pattern where ZEB1 is the only EMT TF uniformly upregulated across the EMT cell lines while any other known EMT TF seems to have virtually little or no expression. This striking observation suggests that ZEB1 is a strong candidate that potentially plays a specific role as the only EMT transcription factor observed in Colorectal cancer cell lines. Many of the genes down- or up-regulated associated with the EMT phenotype is discussed in Chapter 4.



Figure 5.4 Volcano Plot of EMT vs Non-EMT Cell Lines: Volcano plot of EMT vs non-EMT CRC cell lines. Volcano plot of $\log_2$(fold change) vs $-\log_{10}$(p-value) of EMT vs non-EMT cell lines revealing ZEB1 as the only EMT TF that is significantly upregulated in cell lines exhibiting EMT-phenotypes. In addition, VIM is also one of the most upregulated genes in EMT cell lines while CDH1 is one of the most downregulated genes, fitting the hallmark biomarkers of EMT.

Table 5.1 Top 25 Differentially Upregulated and Downregulated Genes in EMT Cell Lines: (a) Top 25 differentially upregulated genes in EMT cell lines. (b) Top 25 differentially upregulated genes in non-EMT cell lines. Both tables prioritized sorting using p values to explore highly significant genes.

a) Genes upregulated in EMT Cell Lines

	Fold Change	P-value
MSRB3--3	21.37	3.87E-16
VIM--2	56.28	3.67E-14
ZEB1--3	17.23	2.33E-10
FAM101B--2	12.19	6.10E-10
CLDN11--2	30.52	9.33E-10
EMP3	15.15	1.02E-09
GNG11	16.31	3.98E-09
HDGFRP3--3	14.05	7.31E-07
WNT5A--1	13.40	1.21E-06
HDGFRP3--4	11.63	1.81E-06
IGFBP7--2	11.73	3.26E-06
MAP1B--3	10.97	3.41E-06
FERMT2--2	20.80	3.42E-06
HDGFRP3--1	18.69	5.52E-06
SMARCA1--4	10.68	1.61E-05
AKAP12--3	12.43	1.73E-05
GJC1--3	13.93	2.35E-05
CALD1--3	10.71	2.43E-05
MSN	12.93	7.06E-05
LGALS1	11.36	2.35E-04
TMEM47--2	14.50	2.55E-04
CALD1--5	16.52	2.71E-04
AKAP12--2	11.16	1.45E-03
PEG10--2	11.54	2.43E-03

b) Genes upregulated in Non-EMT Cell Lines

	Fold Change	P-value
RAB25	39.54	2.58E-15

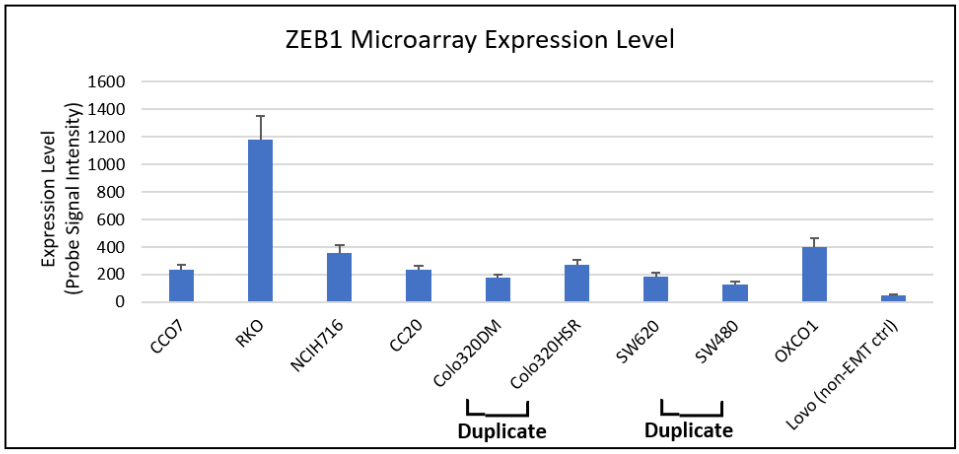
CLDN7	15.25	1.74E-13
MAL2	25.97	4.40E-13
GALNT3--1	32.18	5.47E-13
CDH1--2	40.21	1.34E-10
TMEM30B--1	15.89	9.80E-10
AIM1	13.83	1.88E-09
ERBB3--3	13.82	2.88E-09
ERBB3--5	13.54	9.83E-09
EPCAM	13.58	1.72E-08
S100A14	14.88	4.36E-08
ESRP1--2	14.86	7.34E-08
CD24--5	13.20	5.97E-07
HNMT--4	15.79	8.37E-07
CD24--6	13.56	1.39E-06
AGR2--1	46.48	4.18E-06
LY75	16.14	7.92E-06
S100P	14.98	1.03E-05
TSPAN8	31.51	1.26E-04
CEACAM6--1	13.79	7.09E-04
GPX2--1	14.25	8.91E-04
CEACAM5--1	18.07	1.11E-03
CDH17	16.34	1.27E-03
LGALS4	20.52	1.31E-03

5.2.3 mRNA and protein expression level screen in EMT cell lines

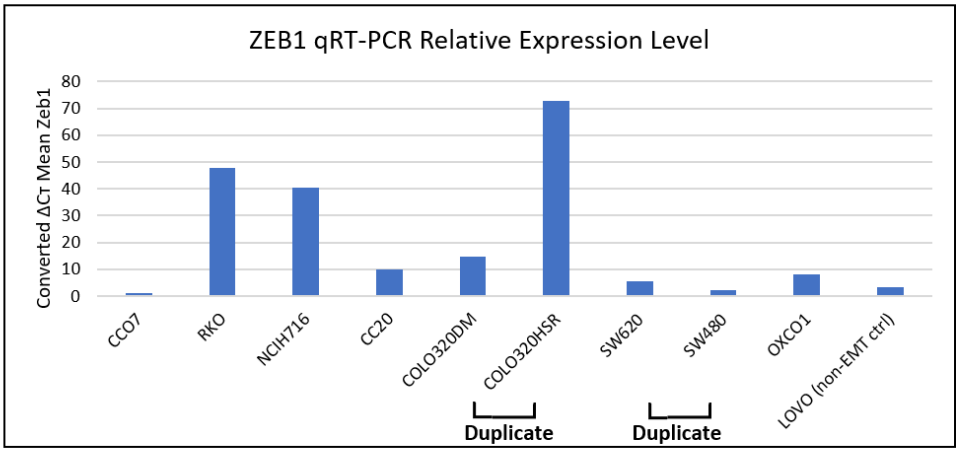
Basal level expression of ZEB1 is screened among the identified EMT cell lines. In our list of EMT cell lines, two of the cell lines Colo320DM and SW620 have matching duplicate cell lines that are isolated from the same patient, but from different origins i.e. primary vs metastatic sites in the case of SW480 and SW620, respectively. These duplicate cell lines are included in the screening due to their similarity in origin with the identified EMT cell lines. Eight EMT CRC cell lines, namely Colo320HSR, Colo320DM, SW480, SW620, CCO7, CC20, OXCO1, NCIH716, and one

non-EMT CRC cell line, Lovo, which has no ZEB1 expression is included as a negative control, and one borderline EMT cell line, RKO, that exhibits a unique pattern of a high ZEB1 expression, low CDH1 expression, but also a low VIM expression is incorporated for the screen. mRNA message level and the functional protein level are assessed by qRT-PCR and western blot, respectively (Figure 5.5). We have spotted a discrepancy between the relative expression level of the microarray data and the qRT-PCR which matches the result from western blot. In Figure 5.5, it is evident that the qRT-PCR and western blot data agree with each other closely. The screen revealed that CCO7 and SW480 both exhibited a similar expression level with Lovo, suggesting little to no ZEB1 expression. We reinvestigated and found out that both SW480 and Colo320HSR are on the borderline of being classified under the EMT cell lines with an upregulation in VIM and downregulation in CDH1 in our microarray data. We will thus include both cell lines for further investigation in the next section.

a)


ZEB1 Microarray Expression Level
Expression Level (Probe Signal Intensity)
1600
1400
1200
1000
800
600
400
200
0
CCO7
RKO
NCIH716
CC20
Colo320DM
Colo320HSR
SW620
SW480
OXCO1
Lovo (non-EMT ctrl)
Duplicate
Duplicate

b)


ZEB1 qRT-PCR Relative Expression Level
Converted ΔCT Mean Zeb1
80
70
60
50
40
30
20
10
0
CCO7
RKO
NCIH716
CC20
COLO320DM
COLO320HSR
SW620
SW480
OXCO1
LOVO (non-EMT ctrl)
Duplicate
Duplicate

c)

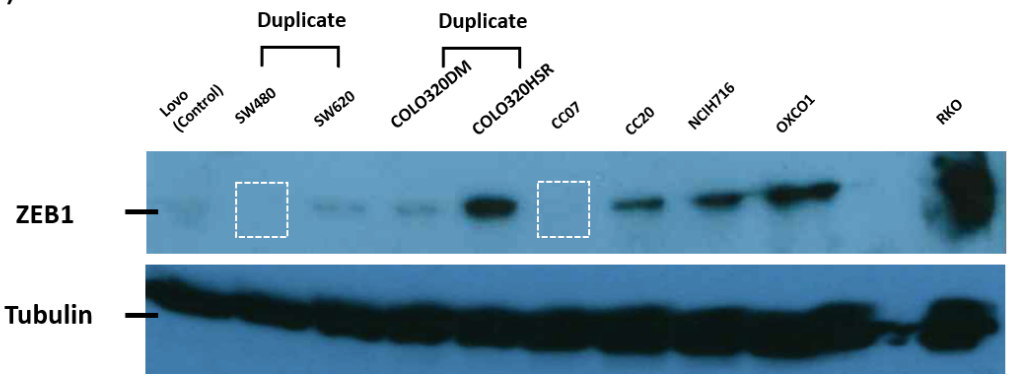

Duplicate
Duplicate
Lovo (Control)
SW480
SW620
COLO320DM
COLO320HSR
CC07
CC20
NCIH716
OXCO1
RKO
ZEB1
Tubulin

Figure 5.5 ZEB1 Microarray, WB, and qRT-PCR Characterization of EMT Cell Lines: (a) Relative ZEB1 expression in our microarray data of EMT cell lines including duplicate cell lines SW480, Colo320HSR, and a negative control non-EMT cell Lovo. (b) and (c) Screen of cell lines for ZEB1 expression with qRT-PCR and western blot, revealing a similar pattern among both baseline characterization of ZEB1 expression. SW480 is a CRC cell line isolated 1 year after the recurrence of CRC. SW620 was established 1 year before the recurrence. Colo320DM is the parental CRC cell line of Colo320HSR

5.2.4 siZEB1 knockdown in selected EMT cell lines

To understand whether ZEB1 is responsible for the EMT expression pattern we observed, we performed a siZEB1 knockdown experiment in OXCO1 to investigate the effect on CDH1. In Figure 5.6, ZEB1 knockdown in OXCO1 showed a 13-fold downregulation at 72 hours and 16-fold downregulation at 96 hours. We also observed a 14-fold upregulation of CDH1 at 72 hours and a 42-fold upregulation at 96 hours. The strong upregulation of CDH1 showed that ZEB1 is indeed regulating CDH1 expression in this EMT cell line. However, as expected, the effect of a siZEB1 knockdown only persisted for 144 hours before losing effect. To study the effect of a ZEB1-knockout in a stable cell line, we established ZEB1-knockouts with CRISPR-Cas9 from the EMT cell lines.

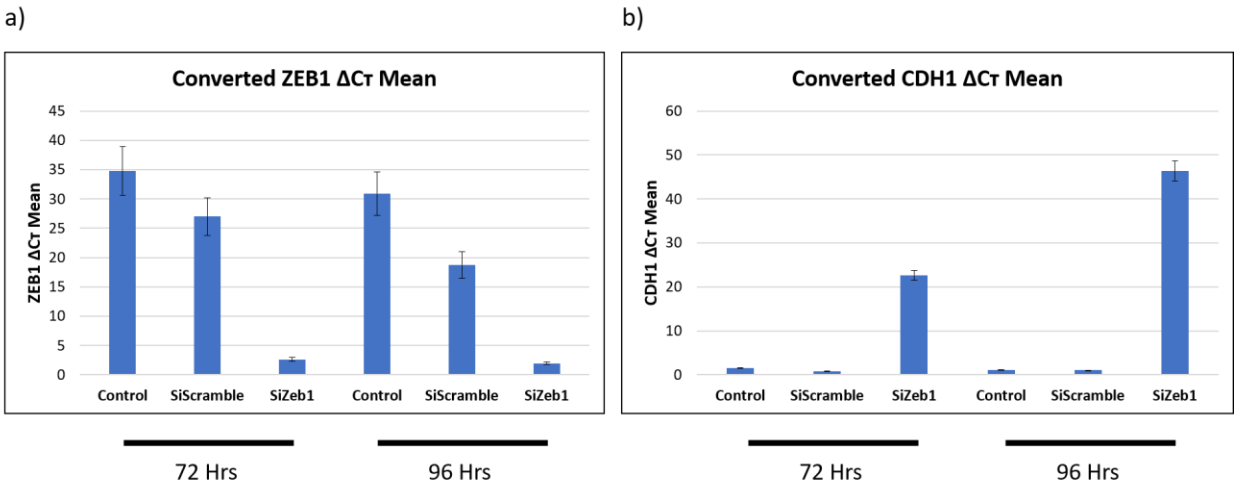


Figure 5.6 siZEB1 Knockdown Experiments in OXCO1: (a) and (b) reveal the ZEB1 and CDH1 ΔCT mean of siZEB1 knockdown in OXCO1, respectively. This qRT-PCR kinetic study of siZEB1 knockdown in CRC cell line OXCO1 from 72 hours to 96 hours reveals a strong upregulation of CDH1. A continuous increase in CDH1 upregulation is also observed from 72 hours to 96 hours

Before deciding the final set of EMT cell lines to perform ZEB1 knockouts on, we were curious whether the detectable level of ZEB1 mRNA level in CCO7 and SW480, shown in Figure 5.5, previously identified during baseline characterization had any functional property in the cell lines. Figure 5.7 is shown the result of siZEB1 knockdown in CCO7 and SW480 to investigate the CDH1 level. siZEB1 knockdown on both cell lines showed observable yet minimal effect both on ZEB1 and CDH1 expression levels; although, careful analysis does seem to suggest that the slight downregulation of ZEB1 did induce a slight upregulation in CDH1. Nonetheless, due to the lack of ZEB1 expression both on the mRNA message and protein level as well as the minimal impact on CDH1, CCO7 and SW480 are excluded from the list of EMT cell lines to study further. The new list of EMT cell lines will now include Colo320DM, Colo320HSR, SW620, CC20, OXCO1, and NCIH716. RKO will be included in the knockout experiments due to its unique pattern of expression. Since NCIH716 is a suspension cell line, we excluded it from our subsequent CRISPR Cas-9 knockout experiment.

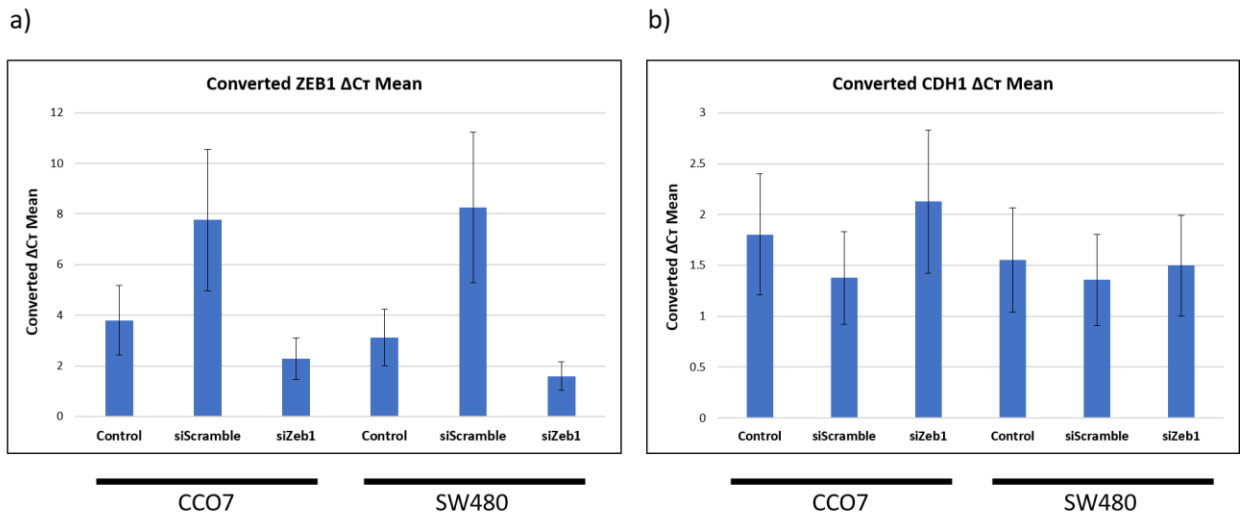


Figure 5.7 siZEB1 Knockdown Experiments in SW480 and CCO7: (a) qRT-PCR of siZEB1 knockdown in SW480 and CCO7 revealing a lack of significant ZEB1 downregulation. (b) qRT-PCR of siZEB1 knockdown in SW480 and CCO7 did not show any significant upregulation in CDH1 suggesting the lack of a functional role of the detectable ZEB1 expression in both cells lines by qRT-PCR. Here, we noticed that there is a larger discrepancy between the control and siScramble. Based on the error bar, there is a high variance in the measured ZEB1 level which we suppose is due to the very low ZEB1 levels in both cell lines. In addition, we believe the siScramble may have introduced some level of variation that resulted in the difference between siScramble and the control sample adding on to the statistical variation of the low ZEB1 level.

5.2.5 ZEB1-KO CRISPR-Cas9 Design

We have designed a double CRISPR-Cas9 system where two guides are designed to target the 5’ upstream and 3’ downstream of the critical exon. The critical exon is defined as the exon that is shared among all isoforms according to the USCS genome browser. The critical exon cannot be the first exon because even if the knockout is successful, there is a high chance we would result in a truncated protein. The major advantage of having a double CRISPR-Cas9 system is the ease of screening a large batch of clones due to a detectable deletion of the gene via PCR and gel-electrophoresis which does not rely on the more expensive Sanger sequencing. Another advantage is the double selection system during FACS that allows more stringent selection

criteria compared to a single selection. Transfected cell lines are single-cell sorted to ensure clonogenicity of the established knockout cell line. Up to 50 single-cell clones, depending on the survival rate of single clones of different cell lines, are first screened by PCR of the targeted critical exon for a confirmed genetic knockout. The selection of the critical exon as well as the gRNA is shown in Figure 5.8. Clones that have a genetic knockout are further characterized with western blotting of ZEB1 to ensure that the functional protein level is completely ablated compared to the wild-type control.

a)

1st set (5' upstream):

Guides	Sequence	PAM	Score	Exonic Off-Target/Max Score
#1	CCAAAAAGGATAGTGATAGT	GGG	72	19 OT sites/1.1
#2	TATACTTTGCTAGGAACTAG	AGG	69	8 OT sites/0.6
#3	ACCAAAAAGGATAGTGATAG	TGG	67	14 OT sites/0.8
#4	CAAAAAGGATAGTGATAGTG	GGG	59	22 OT sites/0.6
#5	ACTCTGATACTTTAGAAAGT	TGG	51	20 OT sites/1.5

2nd set (3' downstream):

Guides	Sequence	PAM	Score	Exonic Off-Target/Max Score
#1	TTCTTTTGGCATACACTAAT	TGG	66	19 OT sites/0.9
#2	TCTTTTGGCATACACTAATT	GGG	61	19 OT sites/0.7
#4	TGCTATATTATCTTCTGATT	AGG	47	21 OT sites/0.8
#5	TGTATGCCAAAAGAAATGCA	AGG	0.56	9
#3	TATCTTCTGATTAGGAAATC	TGG	52	22 OT sites/2.5

b)

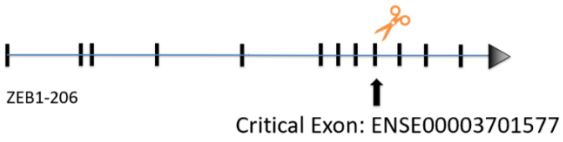


Figure 5.8 Designed Guides for ZEB1-KO: (a) 5 guides designed 5' upstream and 3' downstream using gRNA designing tools including MIT E-Crisp as well as Harvard Chop Chop (b) Schematic of the knockout of ZEB1 through CRISPR-Cas9. Critical exon ENSE00003701577 is chosen for CRISPR knockout because it is shared among all isoforms to ensure a complete knockout of all ZEB1-isoforms.

5.2.6 Establishing ZEB1-KO EMT cell lines

Ablation on a protein level is the gold standard for assessing whether the knockout experiment is carried out successfully. ZEB1 is successfully knocked out in RKO, OXCO1, CC20, SW620, Colo320DM, and Colo320HSR. The initial PCR screen narrows down the number of clones, in most cases, from nearly 50 clones to less than 10 clones. These clones are then assessed with western blotting (Figure 5.9). There are two observations that we occasionally discovered in our screens: a genetic knockout accompanied with remaining protein expression or ablation in protein expression with a lack of genetic knockout. In all our cell lines, we have managed to produce ideal clones where both a genetic knockout and the ablation of protein expression coexist and so these are chosen as the knockout model for the corresponding cell line. Nonetheless, possible explanations that could rationalize the two phenomena are included below:

1) Genetic knockout accompanied with remaining protein expression

The knockout is a frameshift that is a multiple of 3n thus a premature stop codon is not introduced. The frameshifted sequence is translated into a truncated protein that contains a deletion of a couple of amino acids; however, still detectable by immunoblotting due to its remaining epitope recognized by the antibody.

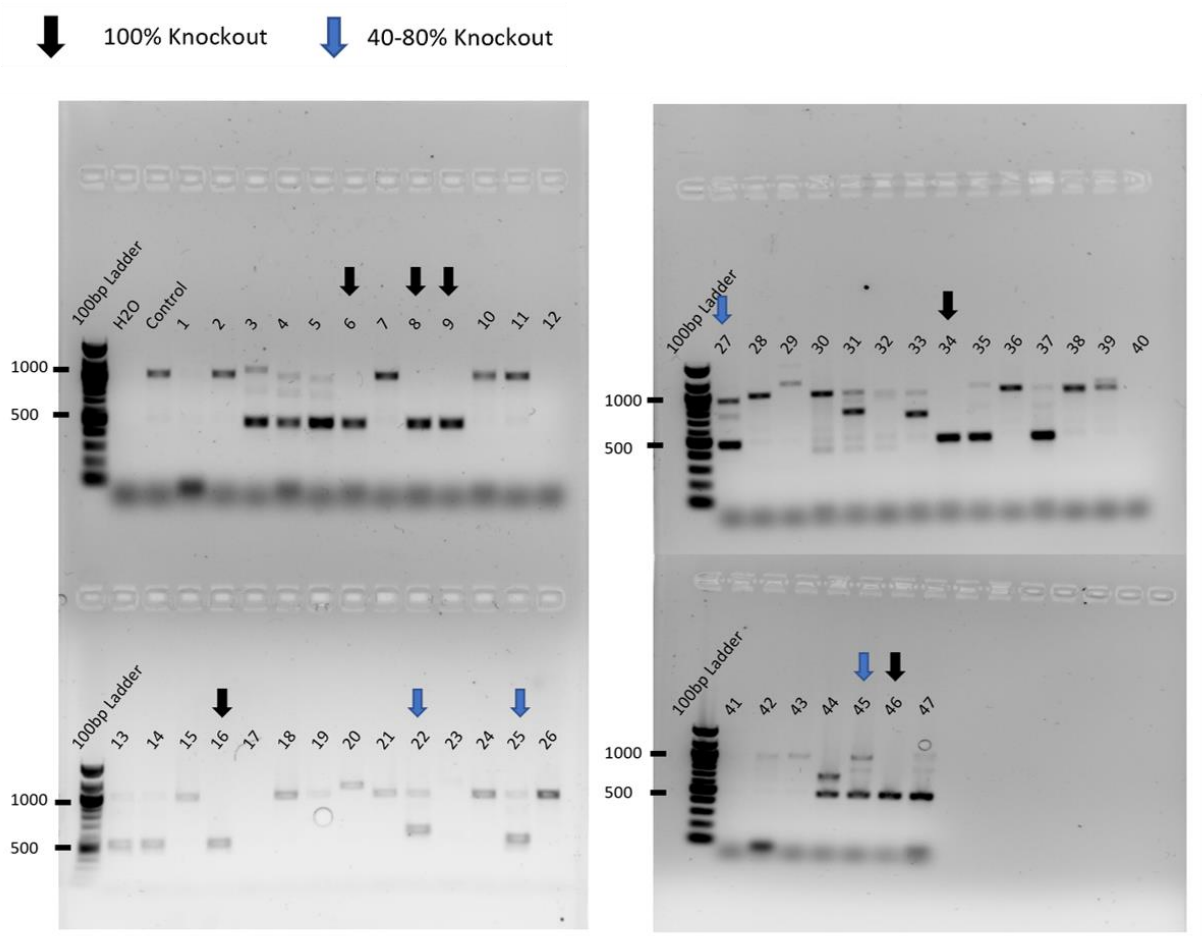
2) Ablation in protein expression with a lack of genetic knockout

The knockout region is inverted, flipped, and reinserted into the knockout region. This would result in a situation where the band size of the PCR product of the knockout region contains the same size as the wild-type control. Nonetheless, it is a bonafide knockout since non-

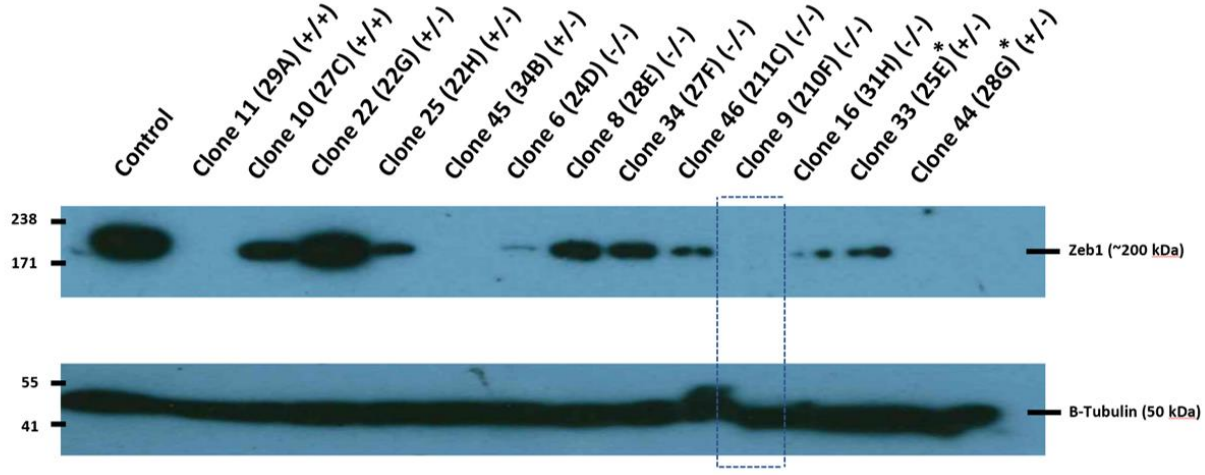
homologous end joining (NHEJ) did introduce a premature stop codon that leads to a nonsense-mediated decay resulting in the ablation of protein expression.

RKO ZEB1-KO

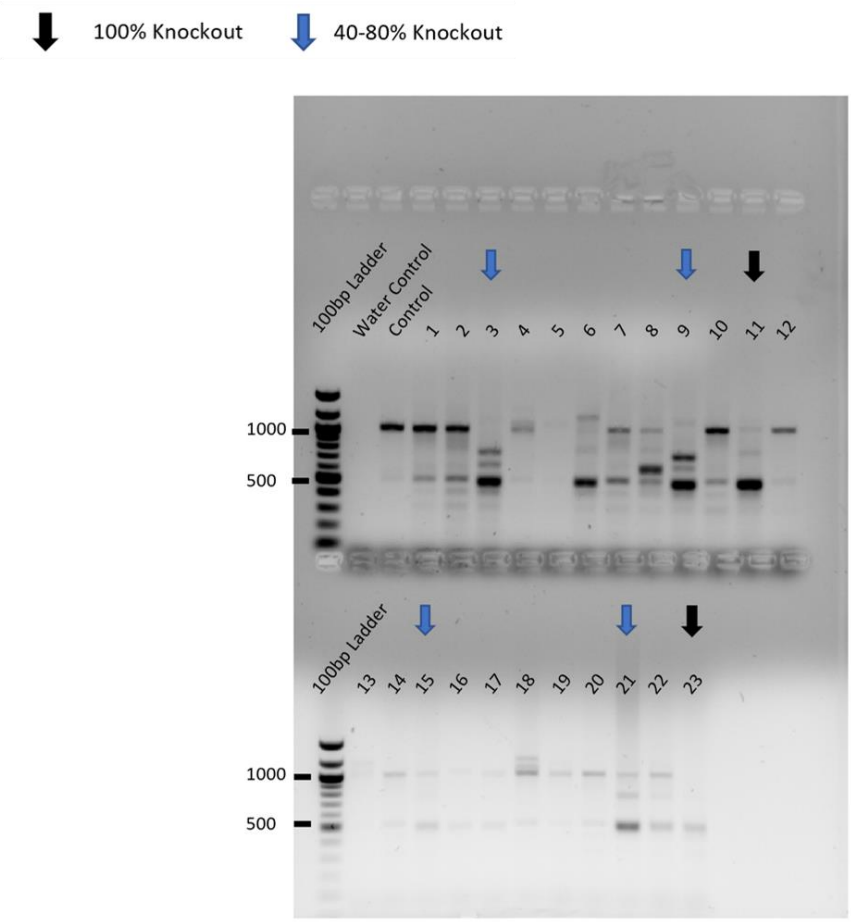
PCR screen for ZEB1-KO



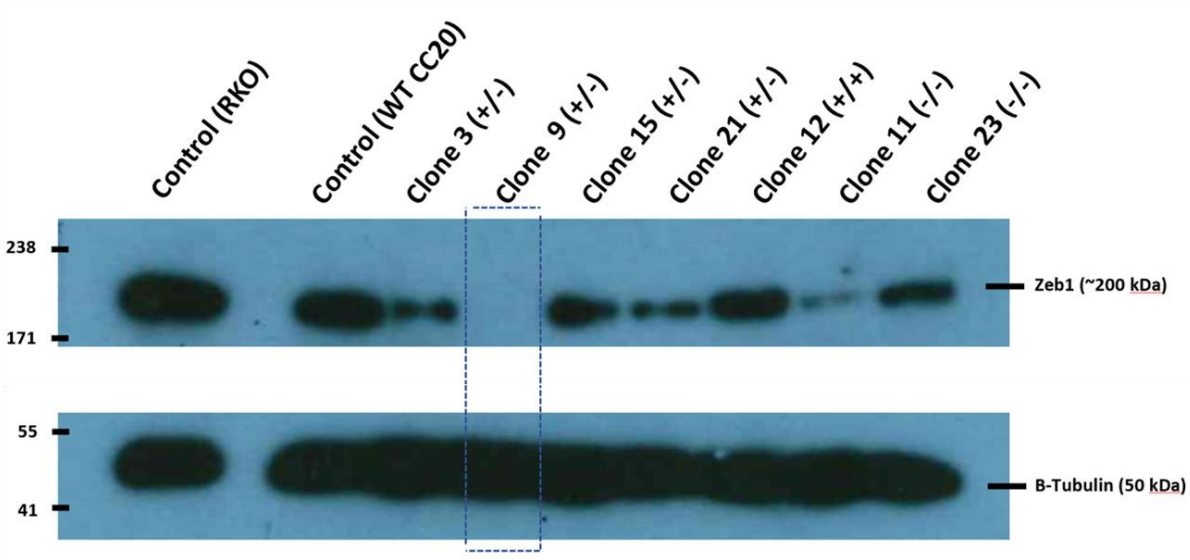
Western blot screen for ZEB1-KO



PCR screen for ZEB1-KO

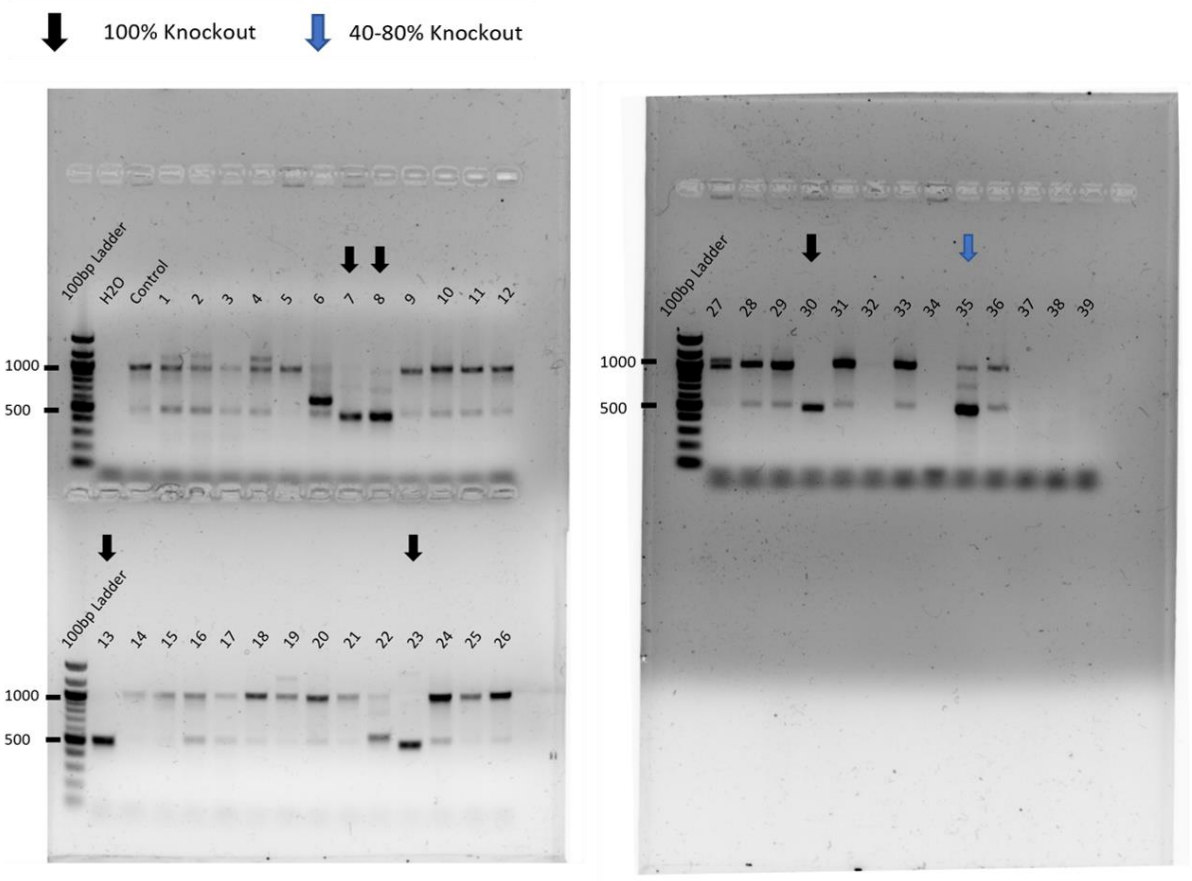


Western blot screen for ZEB1-KO

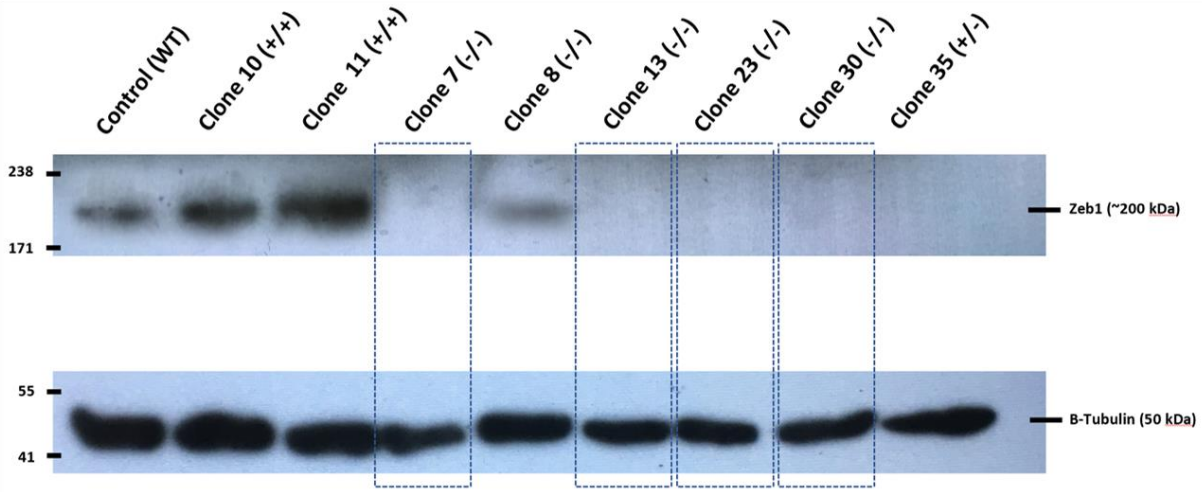


COLO320DM ZEB1-KO

PCR screen for ZEB1-KO

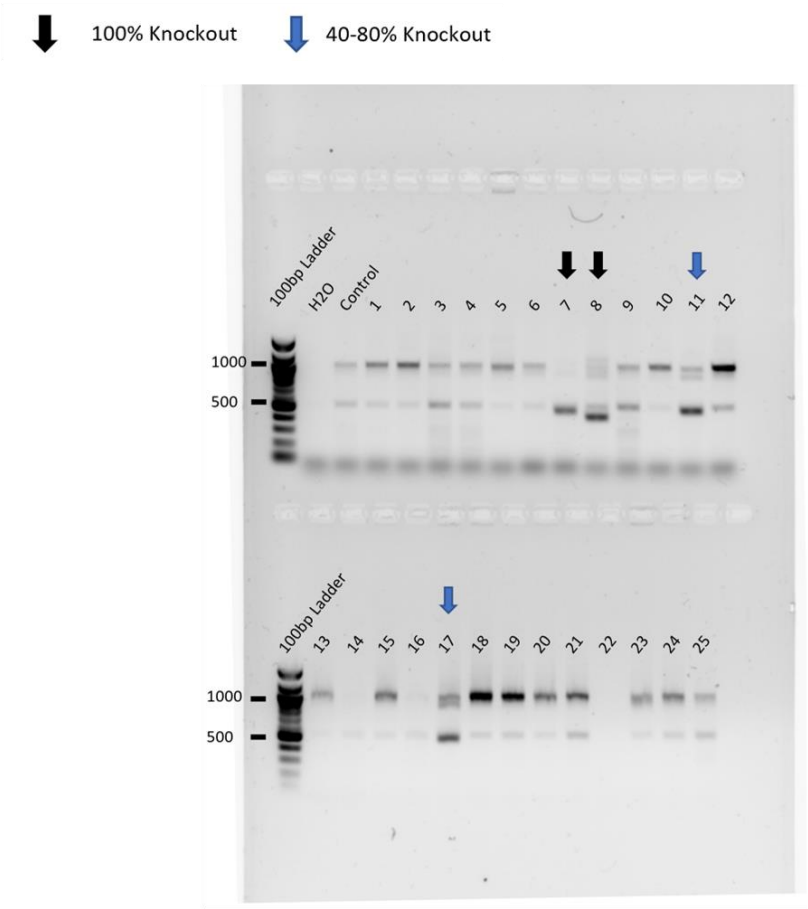


Western blot screen for ZEB1-KO

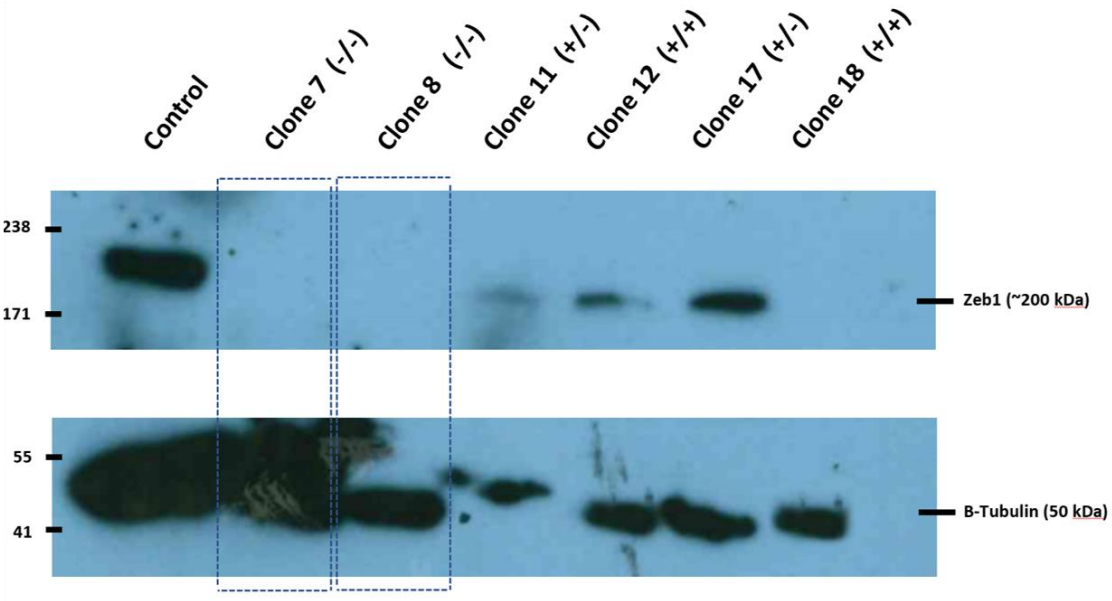


COLO320HSR ZEB1-KO

PCR screen for ZEB1-KO

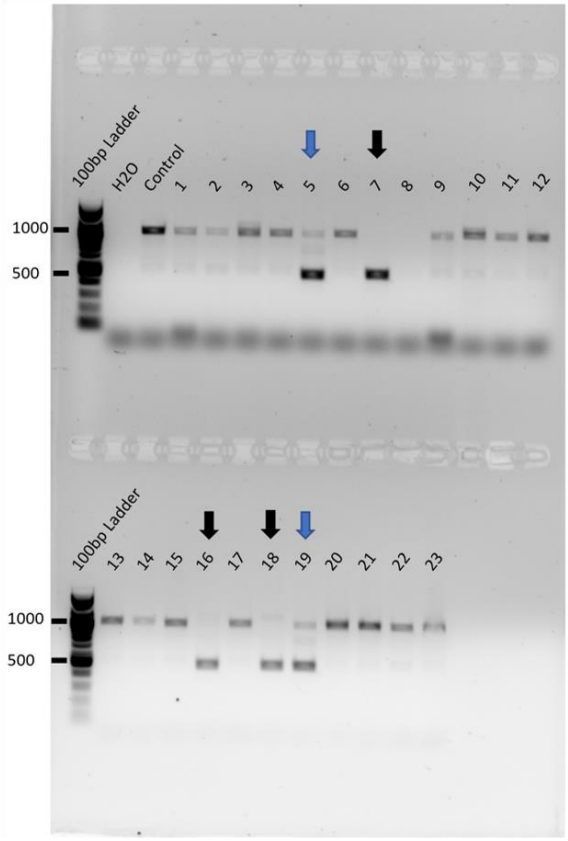


Western blot screen for ZEB1-KO

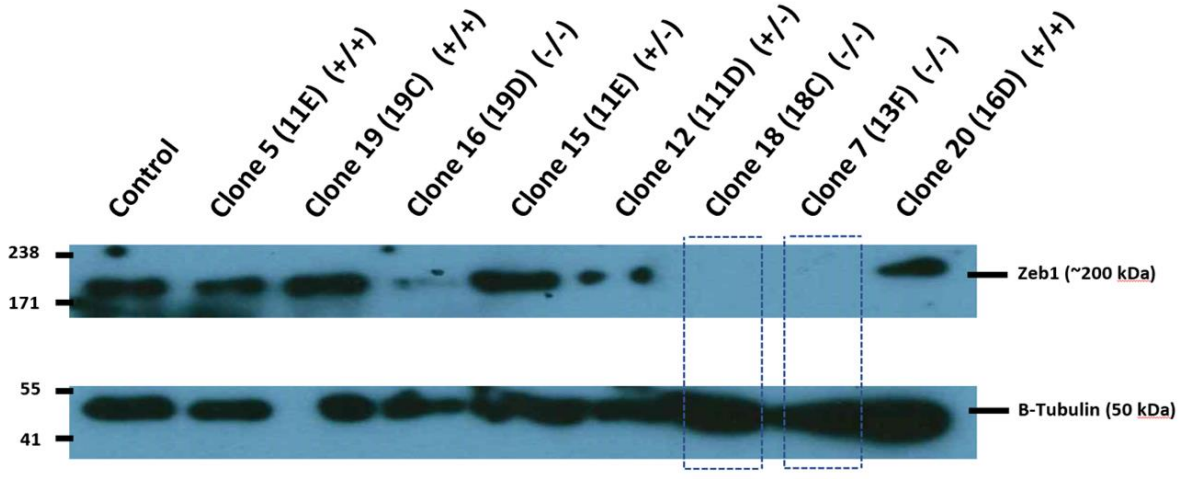


PCR screen for ZEB1-KO





Western blot screen for ZEB1-KO



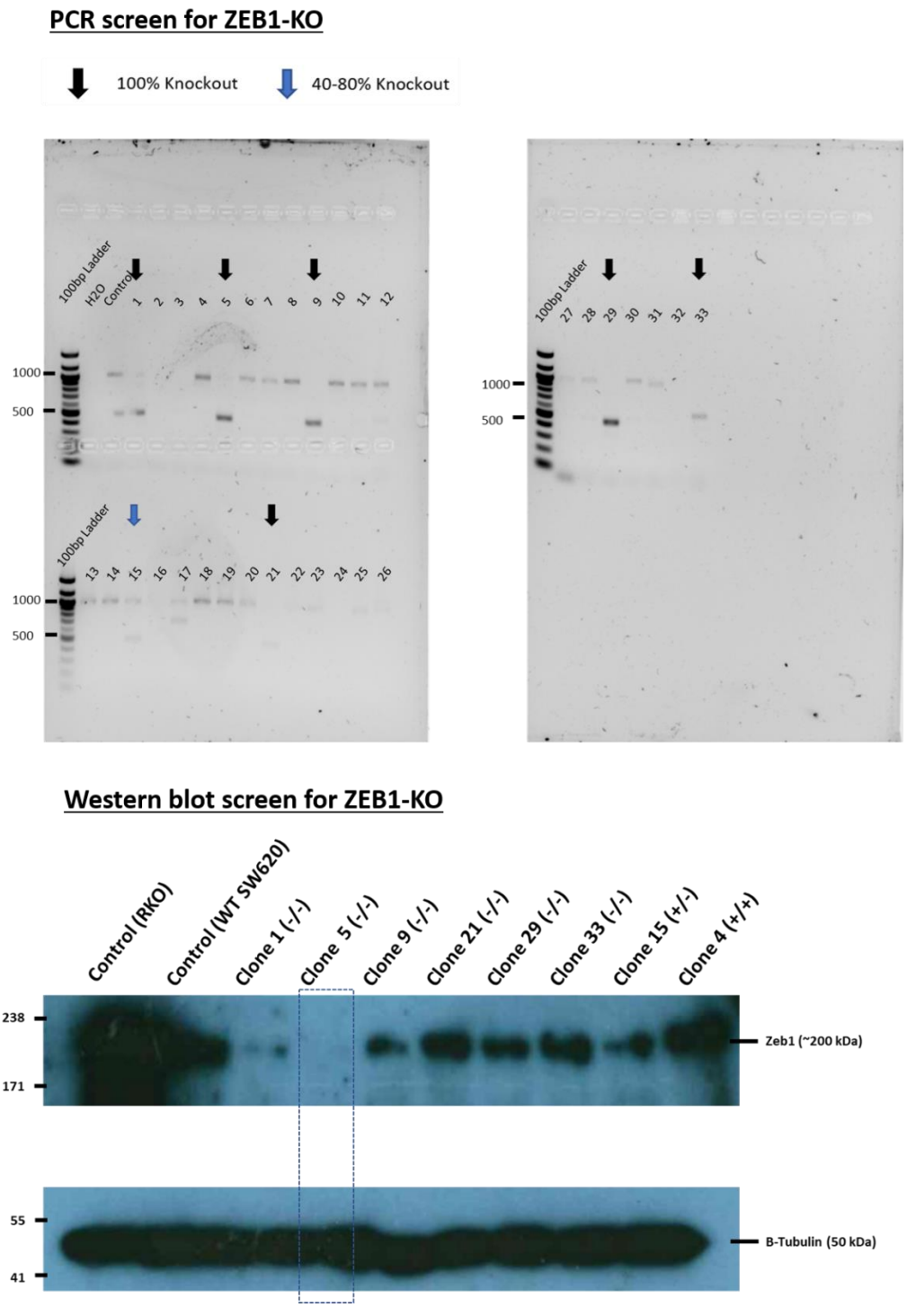


Figure 5.9 PCR and WB of ZEB1-KO EMT CRC Cell Lines Clones: PCR and western blot screen of ZEB1-knockouts in CRC cell lines. Western blot of RKO, OXCO1, CC20, Colo320DM, Colo320HSR, and SW620 clones confirm the complete knockout of ZEB1 expression on both a genetic and protein level.

5.3 ZEB1-KO cell line characterization

Next, we began characterizing ZEB1-KO EMT cell lines. The main characterizations we are interested in are the effect of ZEB1-KO on other known EMT biomarkers including CDH1, VIM, and CTNNB1. In addition, scratch/migration assay to assess the change in cellular plasticity as well as cell doubling time assay to assess the change in proliferation.

When we investigated ZEB1 expression in ZEB1-KO cell lines with western blot (WB) we noticed detectable levels of ZEB1 expression in ZEB1-KO of both RKO and CC20. However, ZEB1 re-expression is only reflected in RKO when we did immunofluorescence (IF) staining (Figure 5.10 and Figure 5.12). Nonetheless, this observation contradicted with the first screen where we selected our clones of ZEB1-KO cell lines prior to subcloning. In the case of RKO, we suspect that there may be a heterogeneous expression of ZEB1 on a single cell level of ZEB1-KO RKO. It is likely that initially, the majority of the clones were RKO ZEB1-KO, however, as the cells doubled and passaged, the RKO ZEB1-wildtype cells outgrew the RKO ZEB1-KO cells. If this assumption is true, this will point to a much higher rate of proliferation in RKO ZEB1-wildtype compared to RKO ZEB1-KO. In the IF, we see that there is no difference between the expression pattern between RKO and RKO ZEB1-KO thus it is reasonable to assume that RKO ZEB1-KO is highly similar to RKO ZEB1-wildtype. We thus focus on those that showed a clear-cut ablation in ZEB1 namely ZEB1-KO cell lines other than RKO.

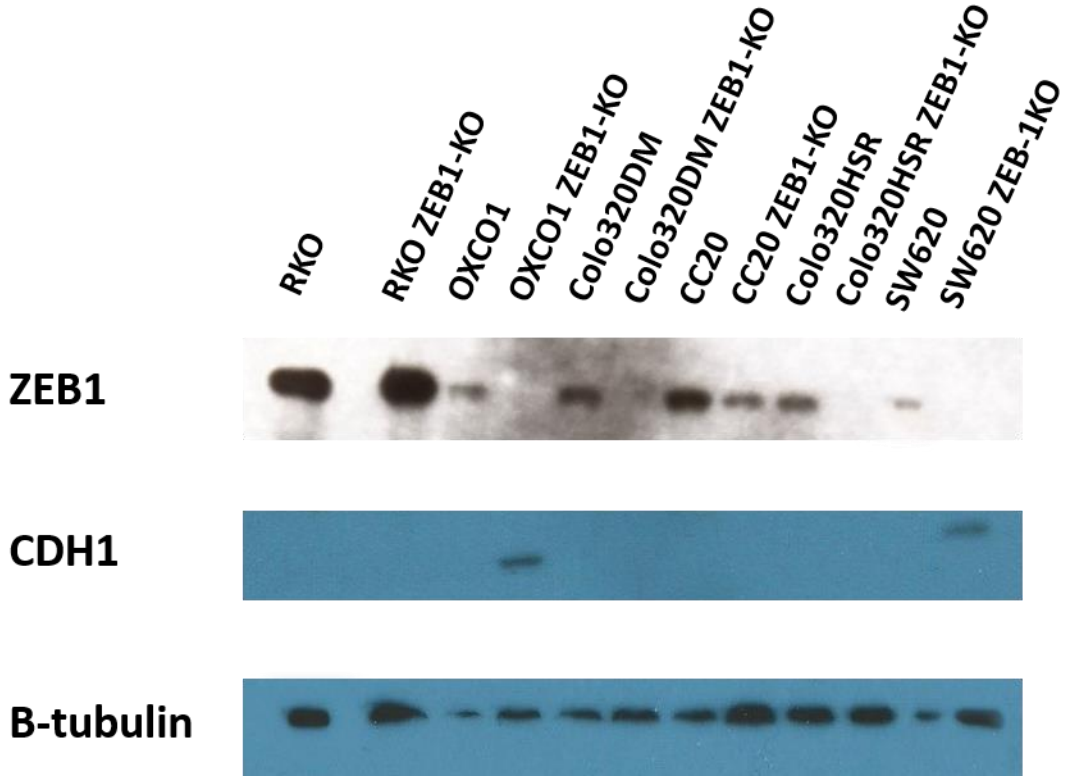


Figure 5.10 WB Detecting ZEB and CDH1 Expression in ZEB1-KO CRC Clones: Western blot of ZEB1, CDH1, and β-tubulin protein expression in our ZEB1-KO cell lines as well as their wild-type counterparts. Here, RKO and CC20 ZEB1-KO had detectable ZEB1 expression comparable to its wildtype counterpart. Also, CDH1 is only detectable in OXCO1 and SW620 ZEB1-KO.

In our IF and WB, we confirmed that ZEB1-KO cell lines other than RKO have no ZEB1 expression compared to their wild-type counterpart. Nonetheless, we observed that only ZEB1-KO in OXCO1 and SW620 led to a re-expression of CDH1. This observation is also seen in our IF. This suggests that ZEB1 is likely not the master regulator of CDH1 and that other factors are affecting CDH1 expression. In addition, we see the downregulation of VIM in OXCO1 ZEB1-KO and SW620 ZEB1-KO. We suspect that the failure of ZEB1 ablation leading to the upregulation of CDH1 and downregulation of VIM in Colo320DM, Colo320HSR, and CC20 suggests that other E-box binding genes can replace the function of ZEB1, which mainly suppresses or upregulates gene

expression via the binding of the E-box domain located in the promoter region of many downstream genes. As shown in Figure 5.11, we found that MITF or melanocyte-inducing transcription factor is a transcription factor that targets and binds onto M-boxes with the target sequence of 5'-TCATGTG-3' and E-boxes (Terragni *et al.*, 2011; Pogenberg *et al.*, 2012; Um *et al.*, 2012). MITF could be a likely candidate that explains why CDH1 and VIM are downregulated and upregulated, respectively, despite the knockout of ZEB1.

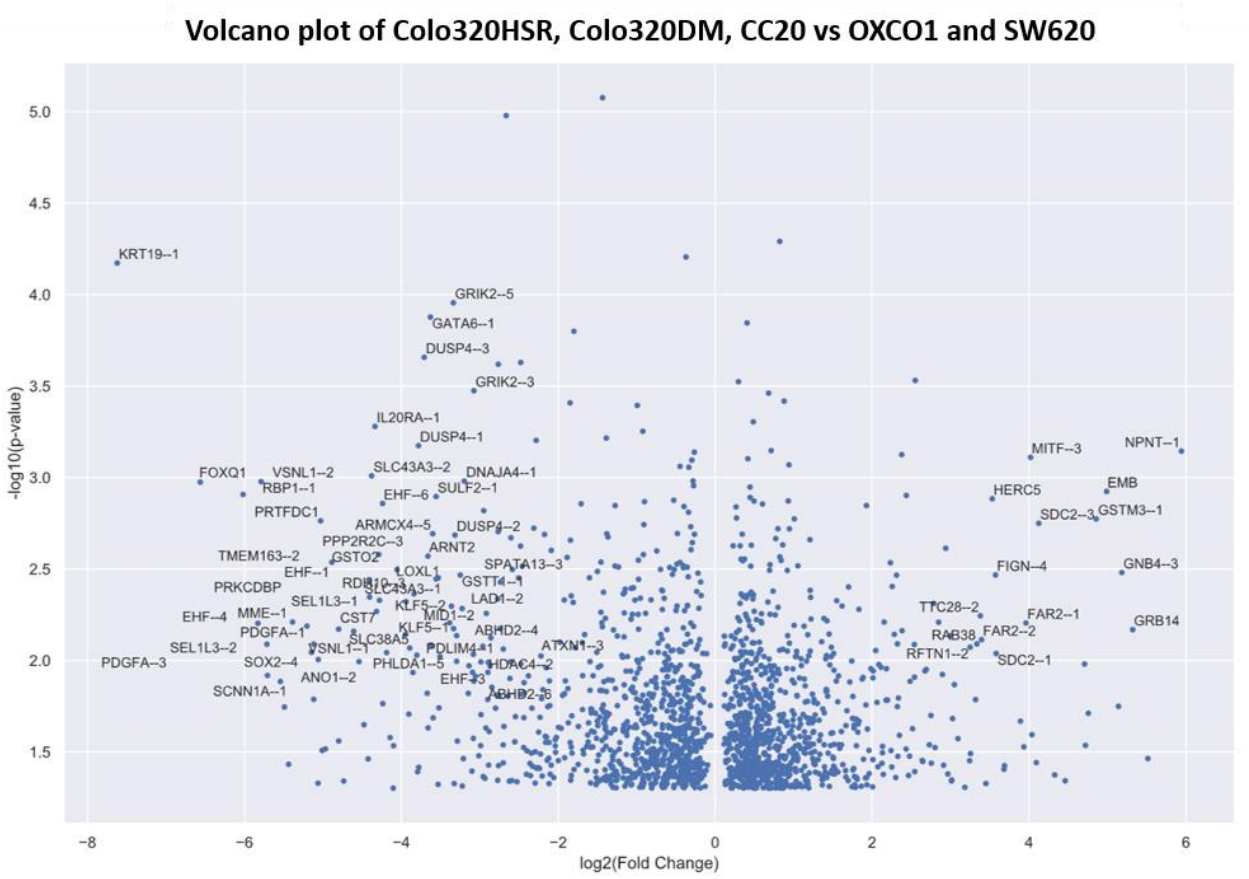
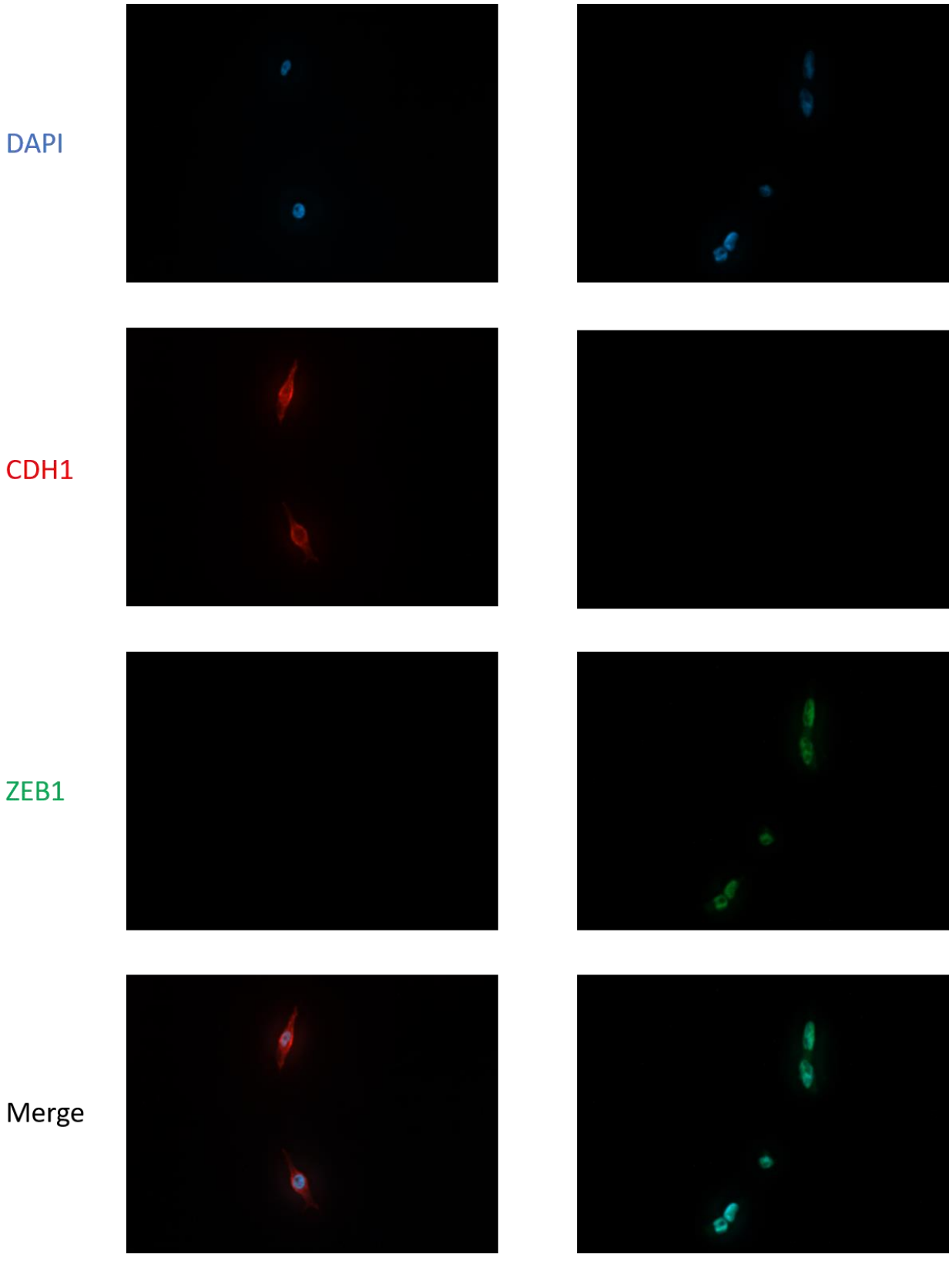
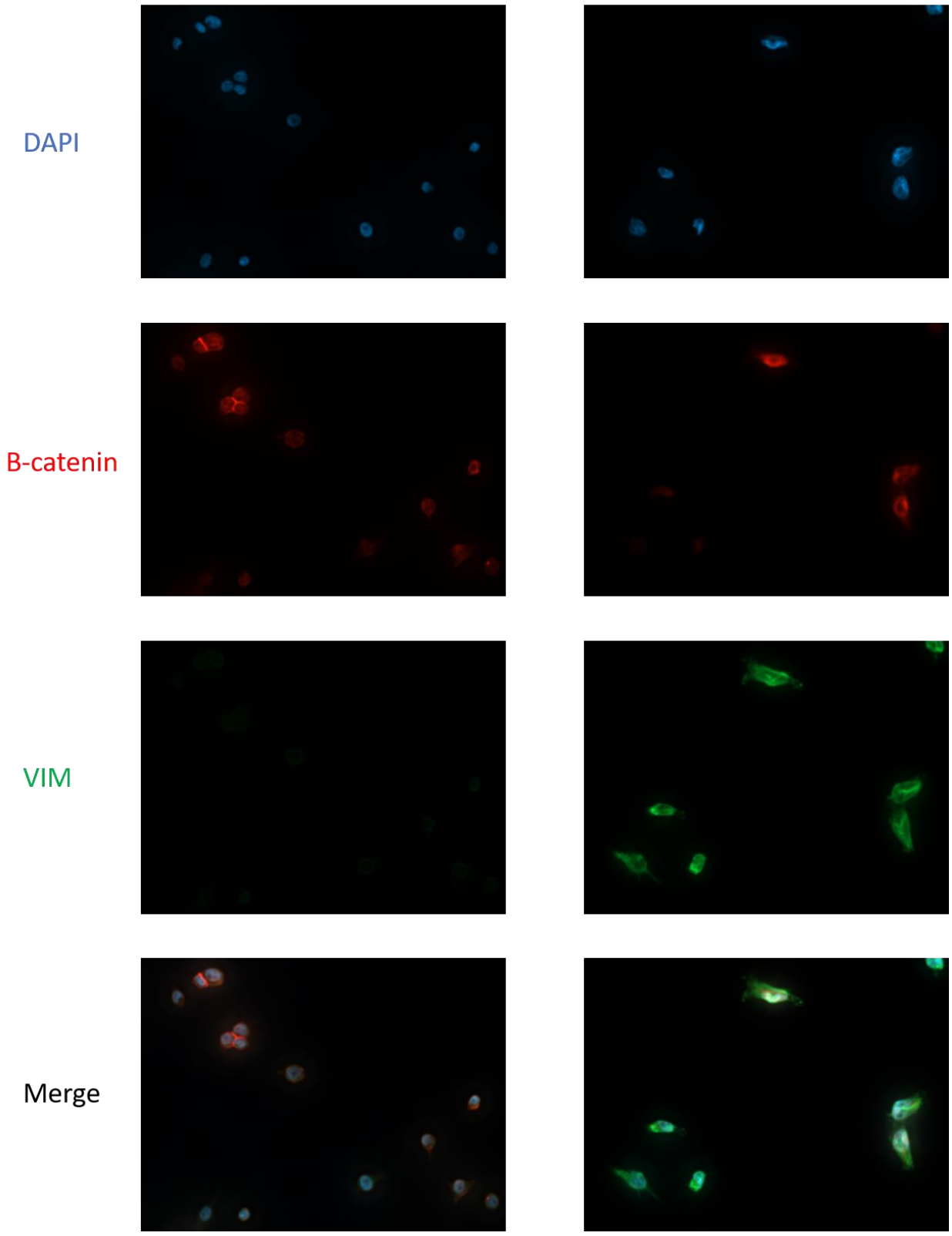
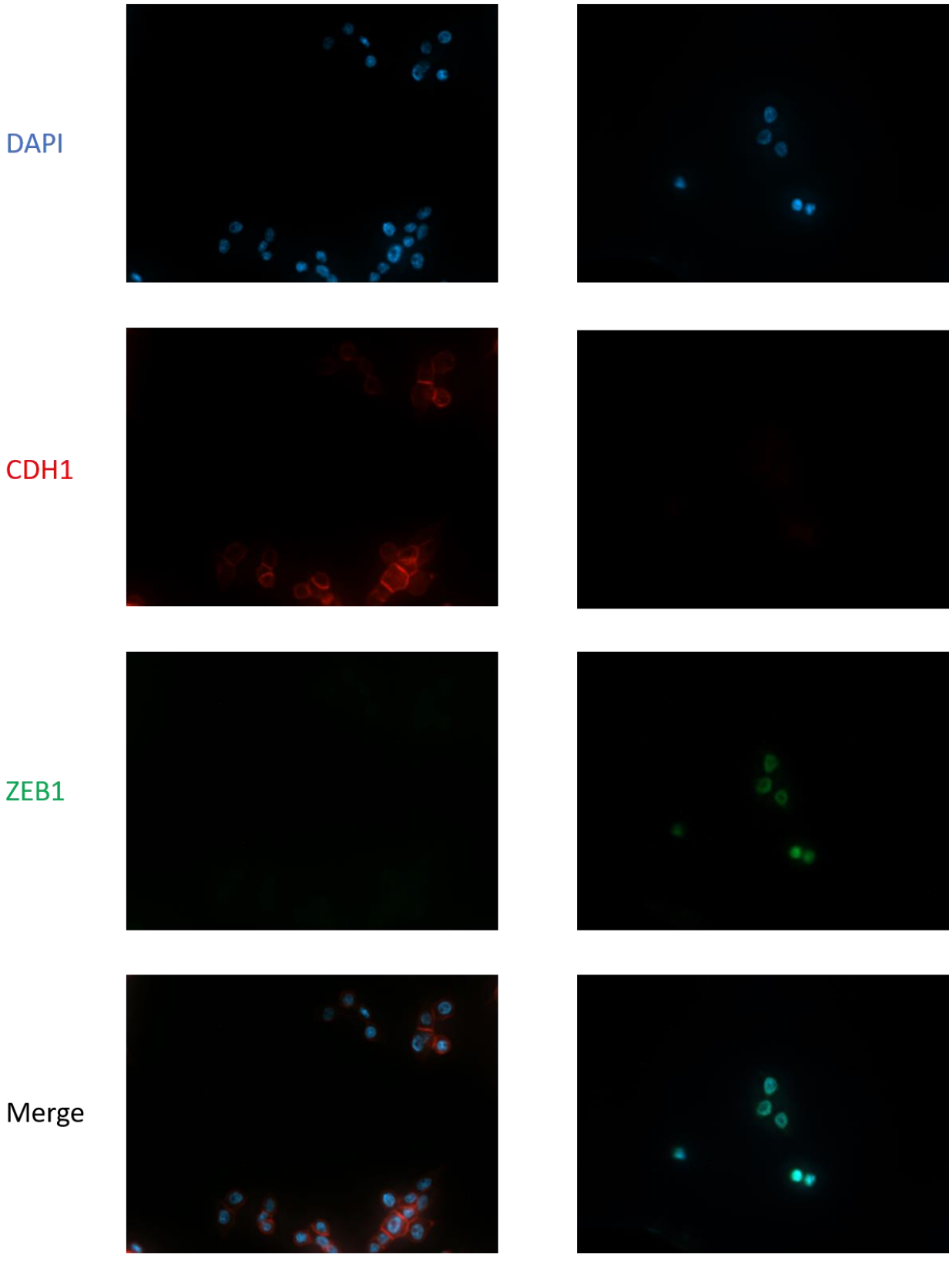


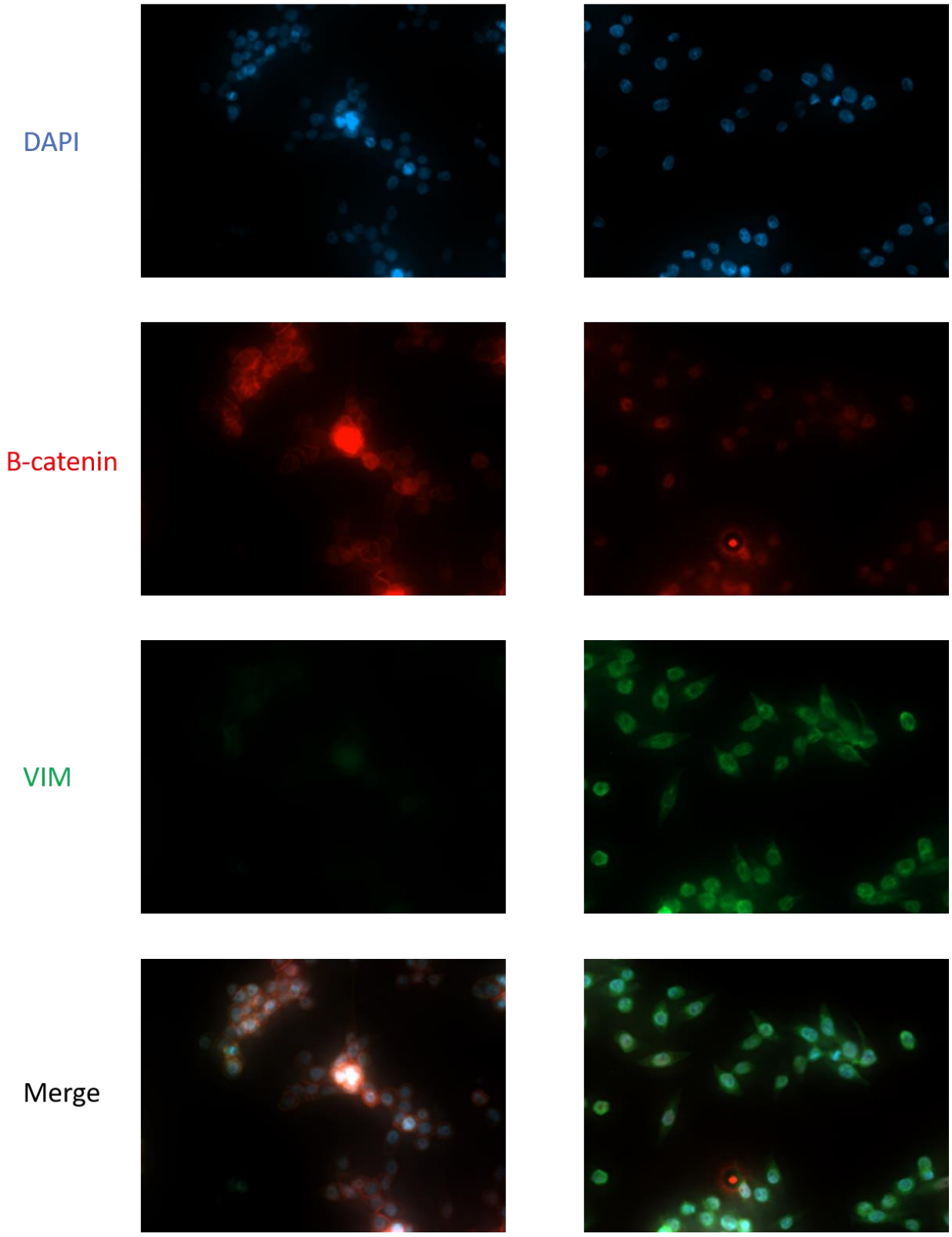
Figure 5.11 Volcano Plot of CDH1-upregulated vs CDH1-downregulated ZEB1-KO CRC Cell Lines: Volcano plot comparing the fold change and p-value of Colo320HSR, Colo320DM, CC20 (cell lines where ZEB1 is ablated but CDH1 and VIM unaffected compared to their wildtype counterpart) and OXCO1, SW620 (cell lines where ZEB1 is ablated along with the upregulation of CDH1 and downregulation of VIM). We show that MITF, which encodes the protein melanocyte inducing transcription factor that regulates its downstream target via the binding of M-boxes and E-boxes, is highly upregulated in the three cell lines where CDH1 and VIM are unaffected and may be responsible for inhibiting CDH1 and upregulating VIM despite the knockout of ZEB1 expression.

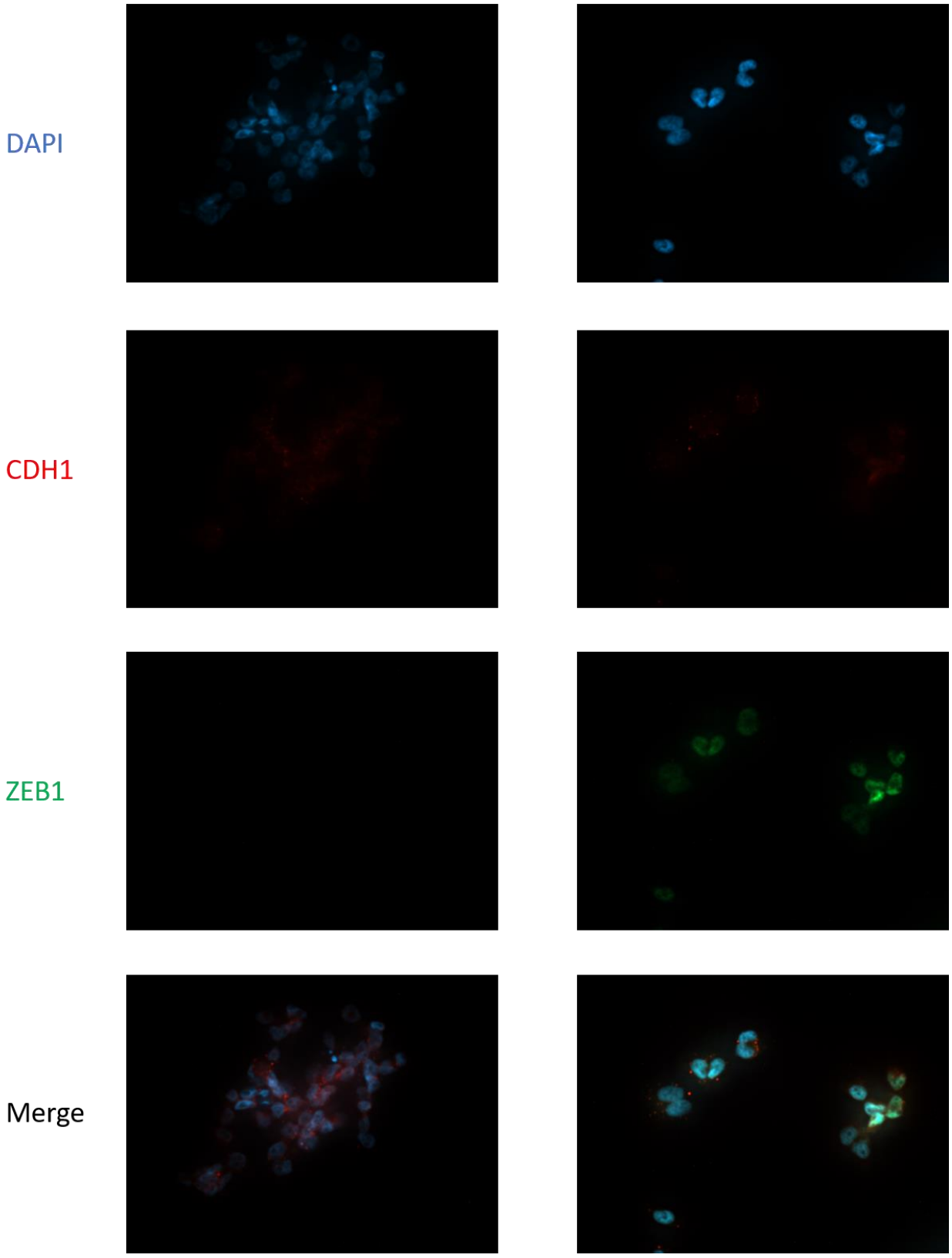
Figure 5.12 ZEB1, CDH1, β-catenin, and VIM Immunofluorescent Staining on ZEB1-KO Cell Lines: Immunofluorescent staining on ZEB1-KO cell lines and their wildtype counterpart. Similar to the result in our western blot, CDH1 is only upregulated in OXCO1 and SW620 ZEB1-KO. We also noticed the membrane expression of β-catenin in OXCO1 and SW620 ZEB1-KO where CDH1 is upregulated. This is in line with previous observations of the membrane localization of β-catenin in CDH1 expressing cell lines. On the other hand, in other cell lines where CDH1 is not expressed, we see nuclear staining of β-catenin. The magnified image is shown in Figure 5.14.

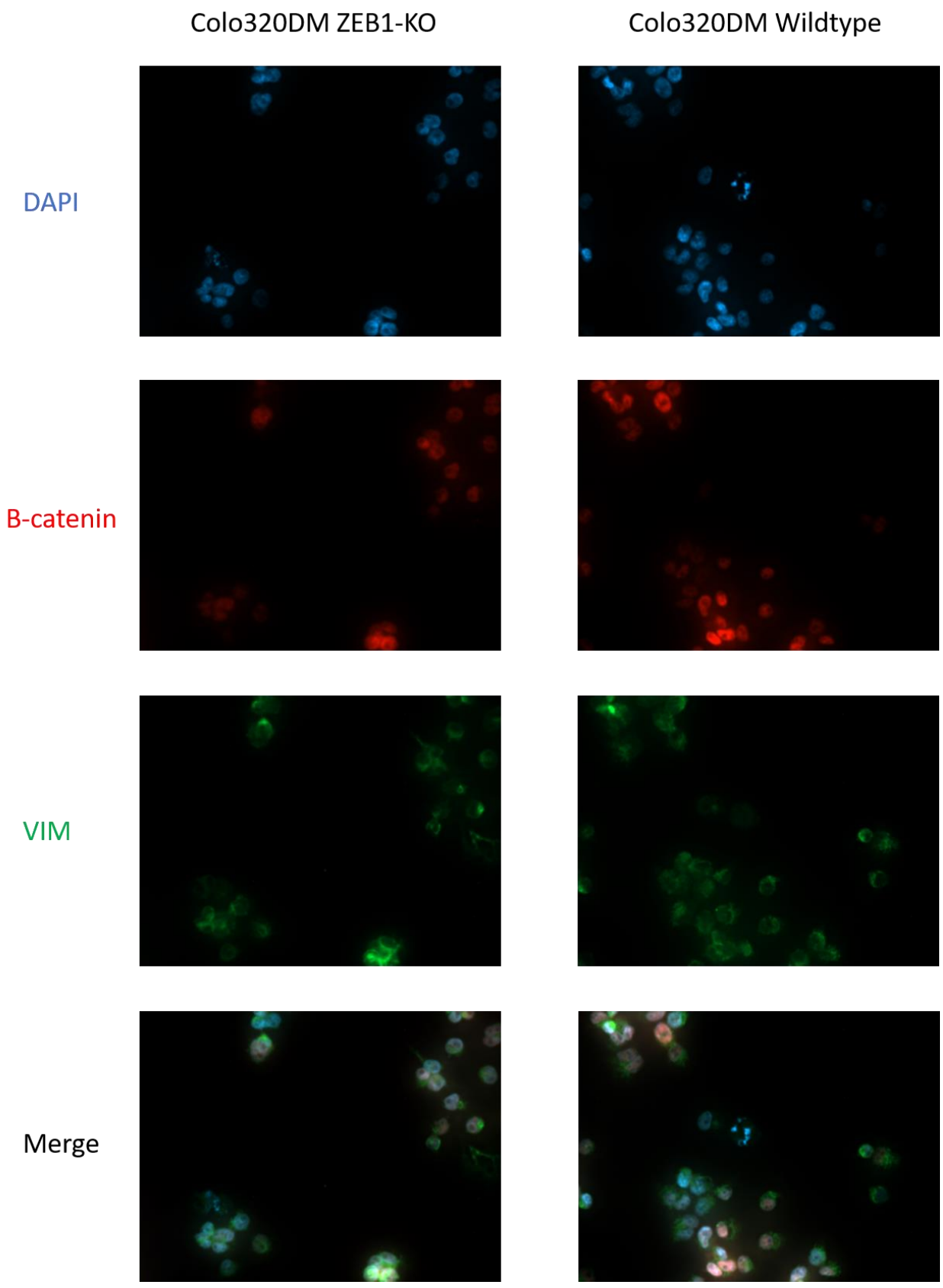

OXCO1 ZEB1-KO
OXCO1 Wildtype
DAPI
CDH1
ZEB1
Merge

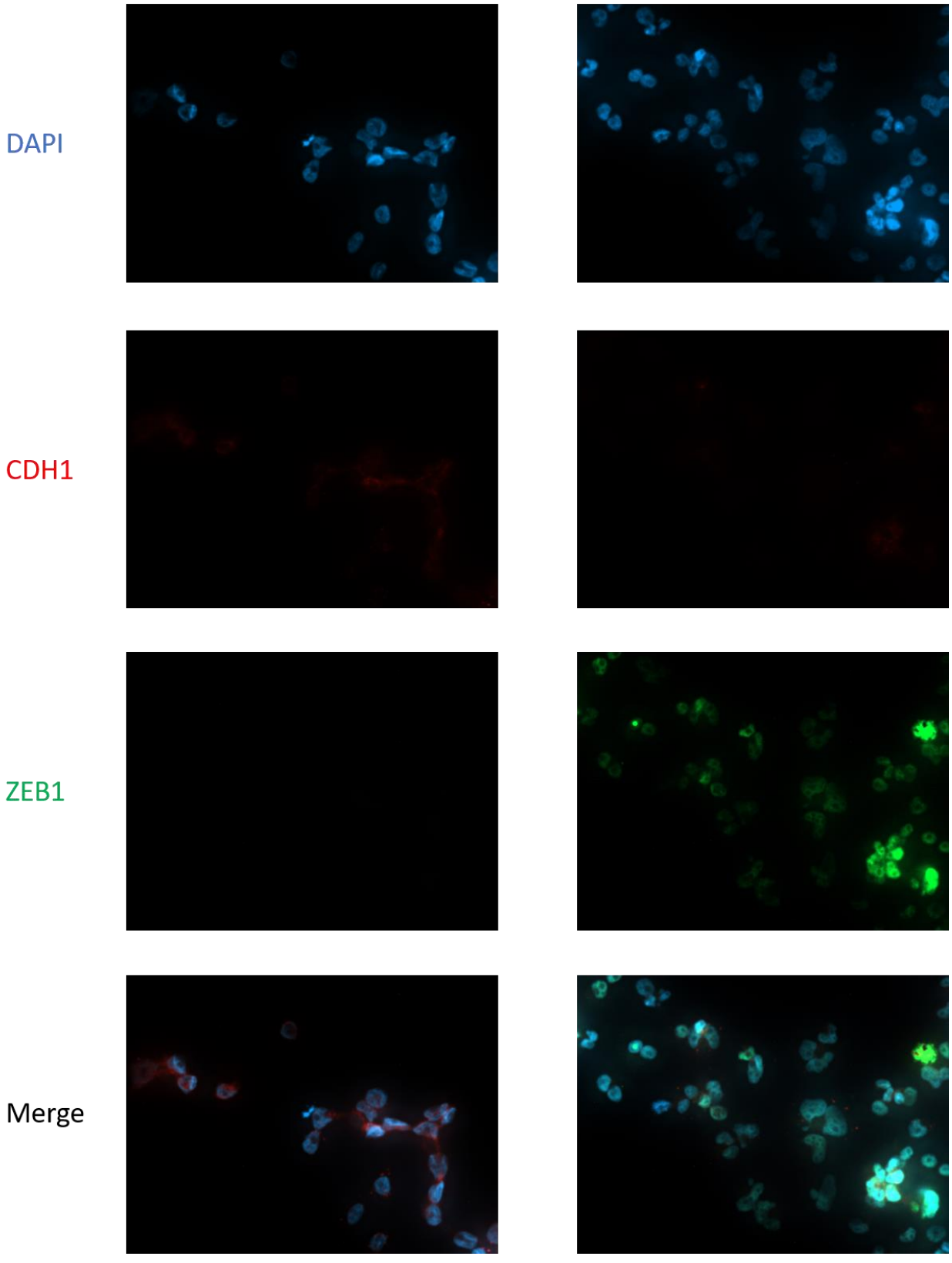

OXCO1 ZEB1-KO
OXCO1 Wildtype
DAPI
B-catenin
VIM
Merge

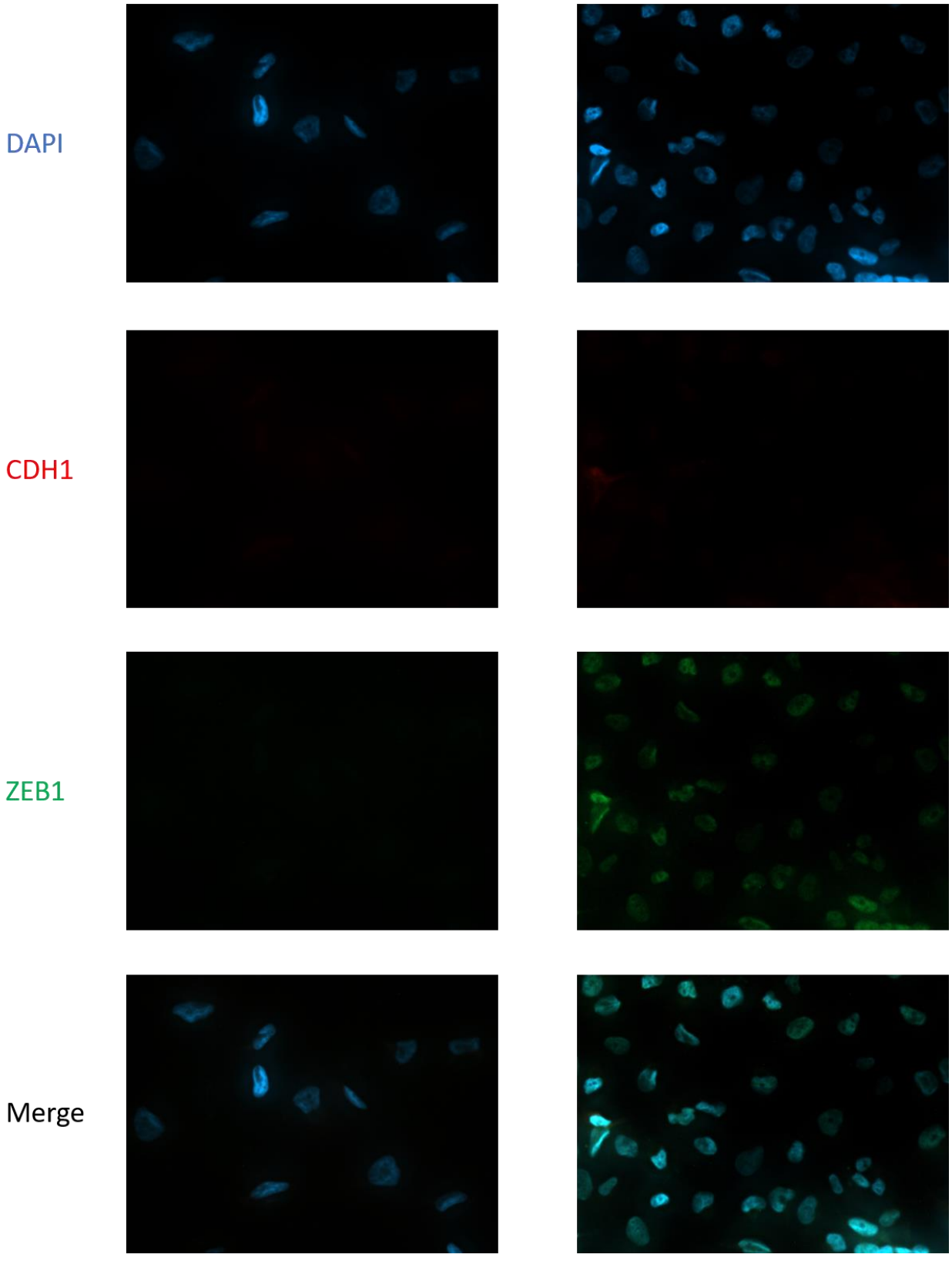

SW620 ZEB1-KO
SW620 Wildtype
DAPI
CDH1
ZEB1
Merge

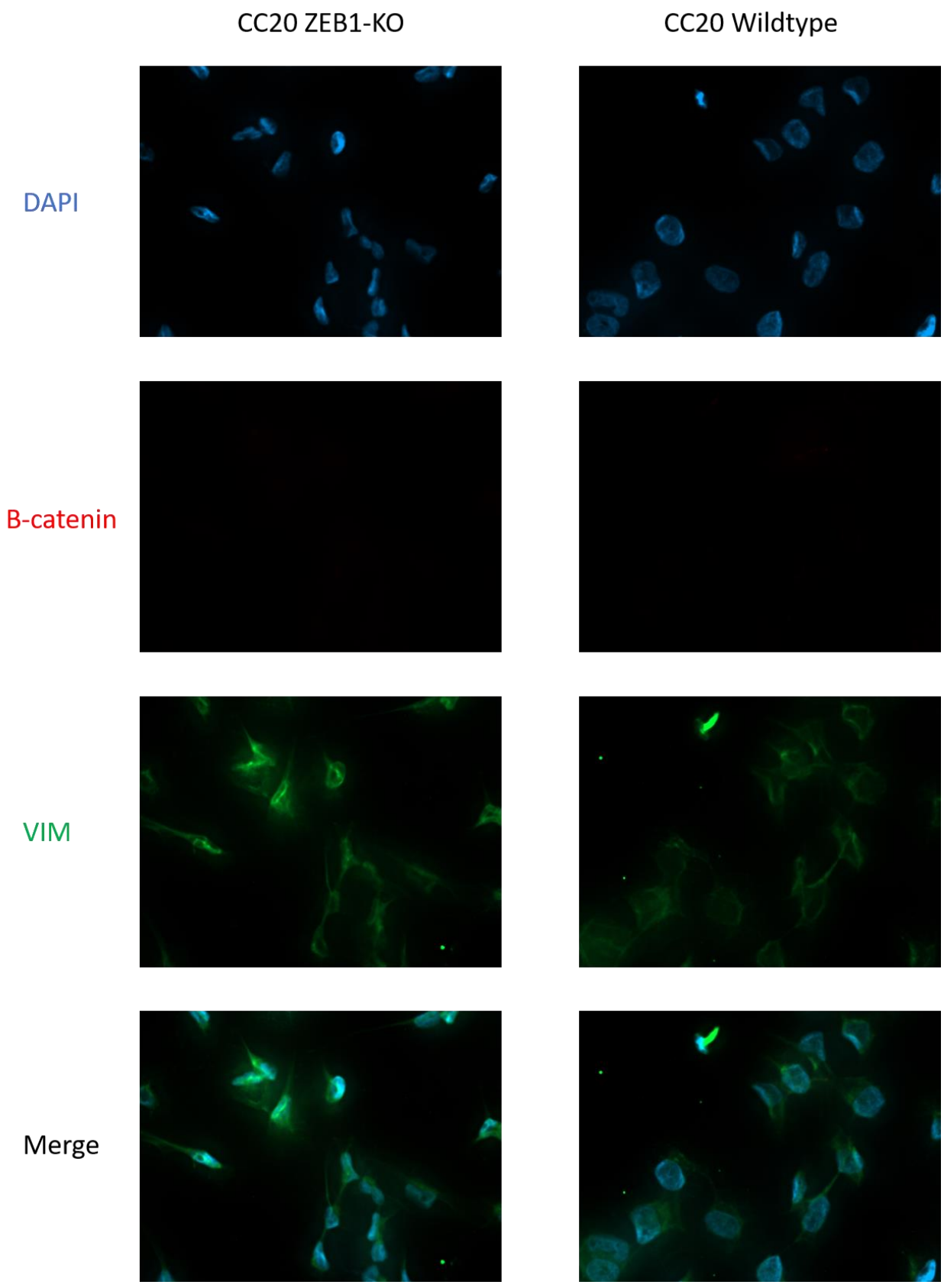

SW620 ZEB1-KO
SW620 Wildtype
DAPI
B-catenin
VIM
Merge

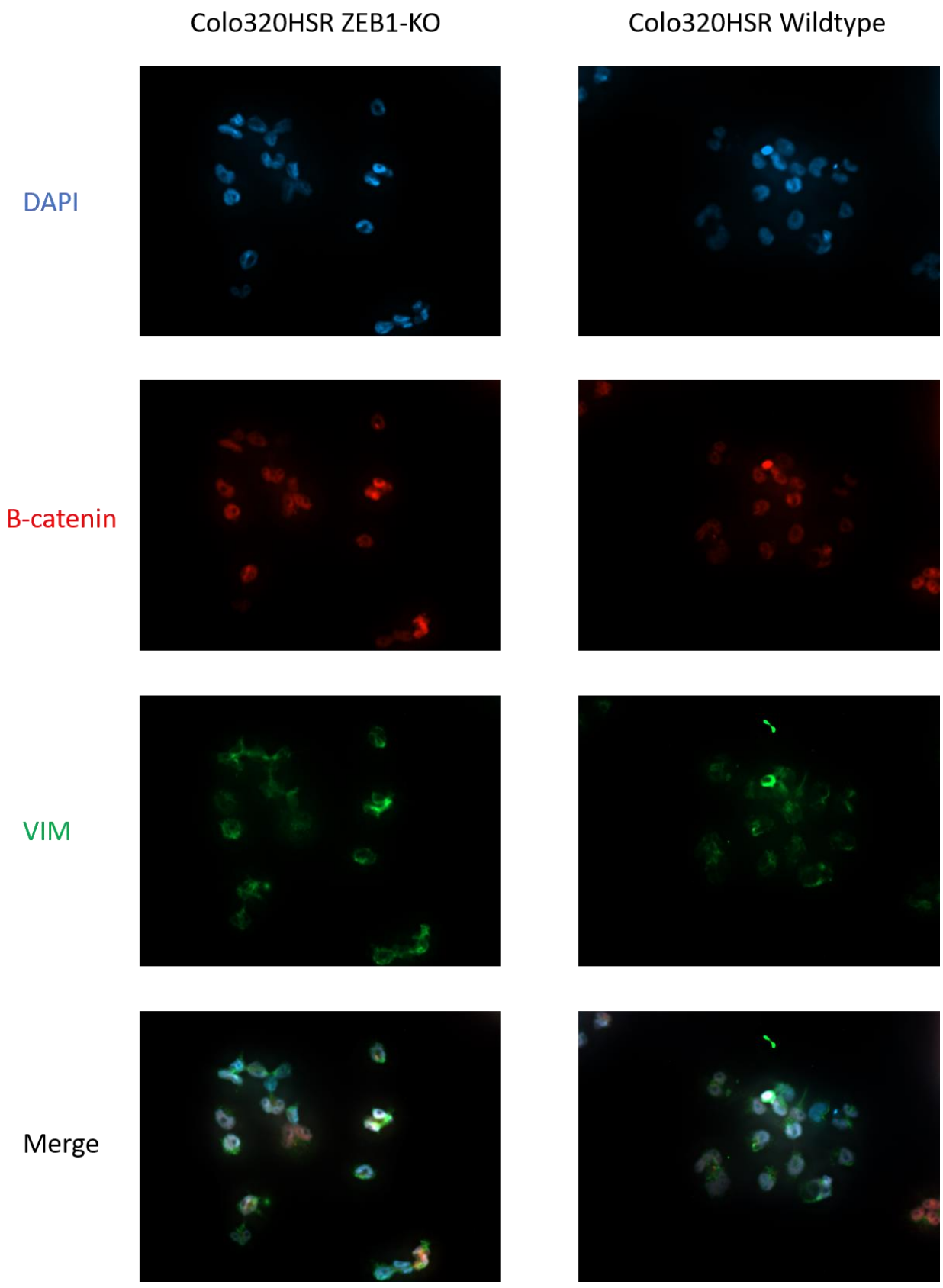

Colo320DM ZEB1-KO
Colo320DM Wildtype
DAPI
CDH1
ZEB1
Merge

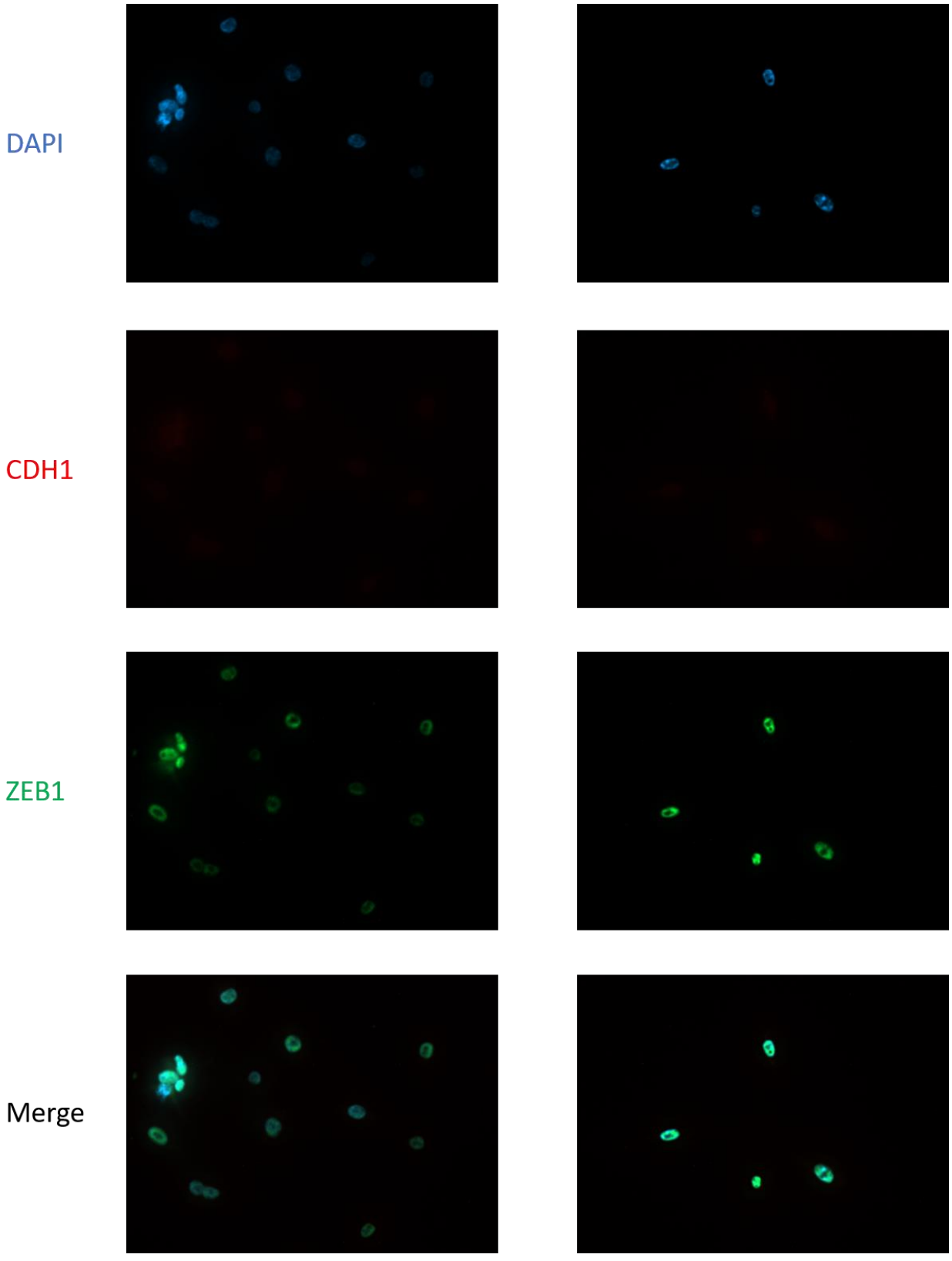

Colo320DM ZEB1-KO
Colo320DM Wildtype
DAPI
B-catenin
VIM
Merge

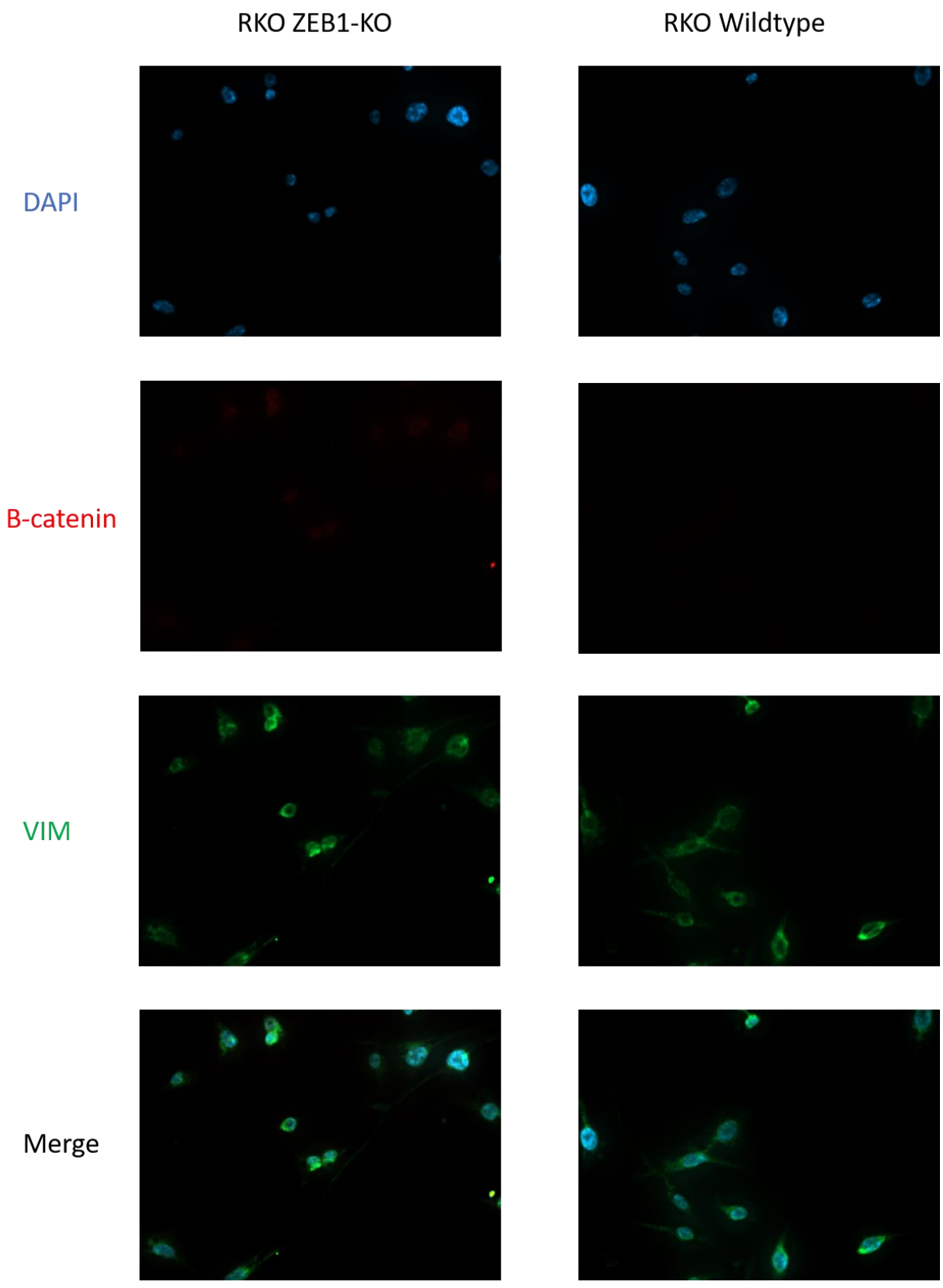

Colo320HSR ZEB1-KO
Colo320HSR Wildtype
DAPI
CDH1
ZEB1
Merge


CC20 ZEB1-KO
CC20 Wildtype
DAPI
CDH1
ZEB1
Merge


CC20 ZEB1-KO
CC20 Wildtype
DAPI
B-catenin
VIM
Merge

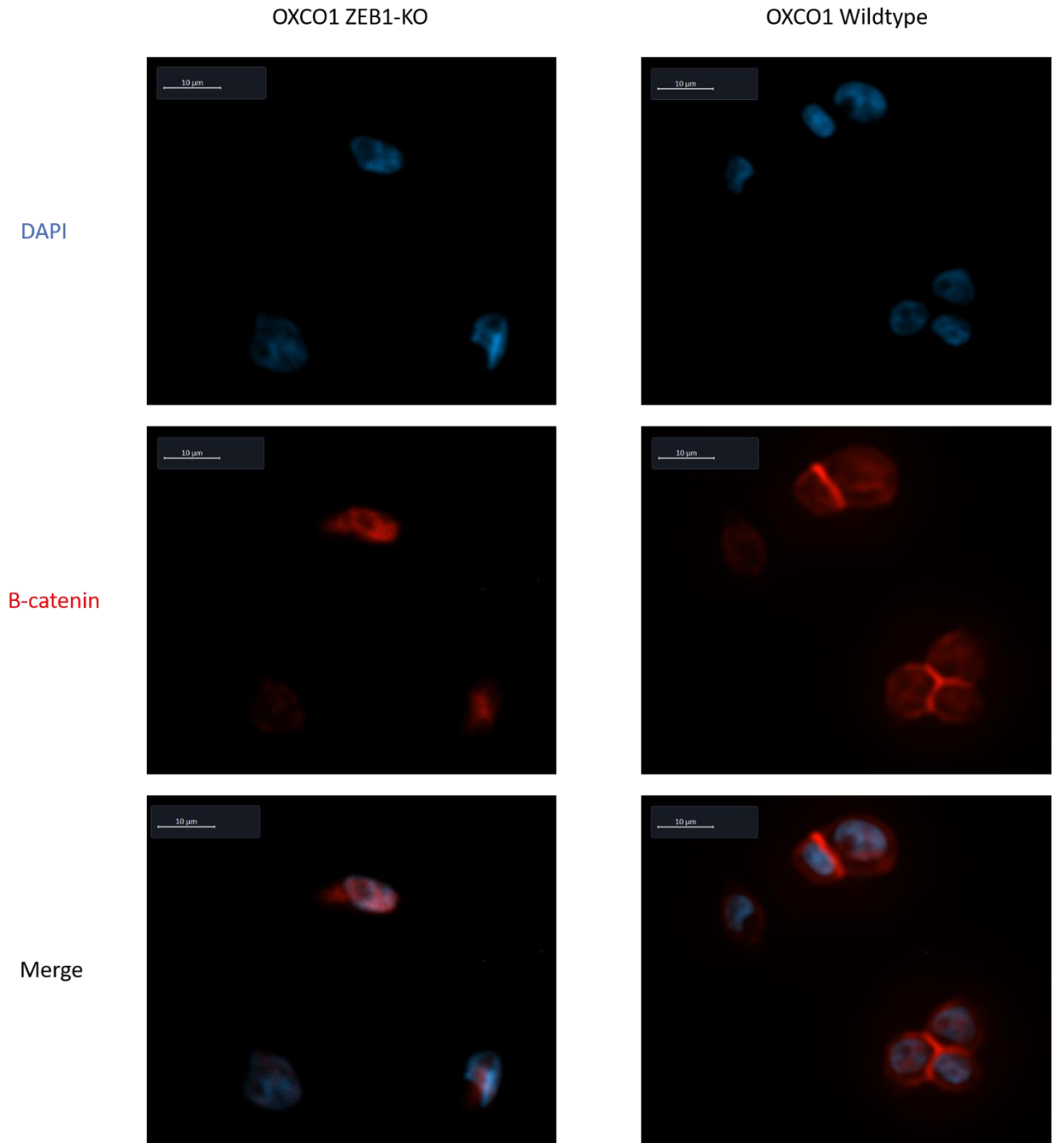

Colo320HSR ZEB1-KO
Colo320HSR Wildtype
DAPI
B-catenin
VIM
Merge

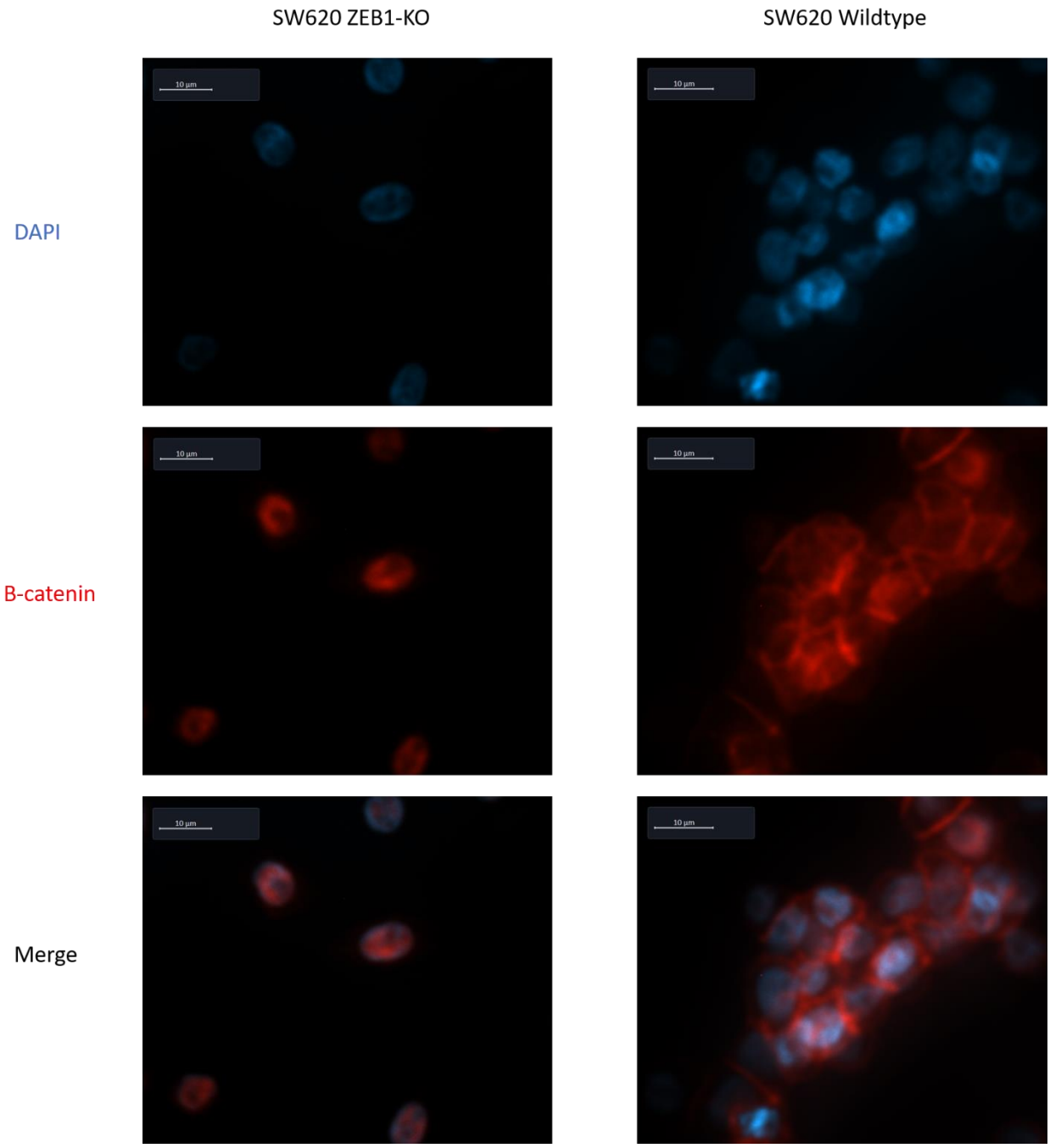

RKO ZEB1-KO
RKO Wildtype
DAPI
CDH1
ZEB1
Merge


RKO ZEB1-KO
RKO Wildtype
DAPI
B-catenin
VIM
Merge

In the literature (Huber and Weis, 2001; Tian *et al.*, 2011) as well as previous research from a DPhil student in our lab, Laura Collings, have shown the Colocalization of β-catenin and E-cadherin when she performed β-catenin IF on a large panel of CRC cell lines. In Figure 5.13, we have strong evidence that reinforces this observation where the upregulation of CDH1 in ZEB1-KO in OXCO1 and SW620 led to the Colocalization and membrane expression of β-catenin. Compared to their wildtype counterpart OXCO1 and SW620 where CDH1 is not expressed, β-catenin is in both of these cases nuclear staining.

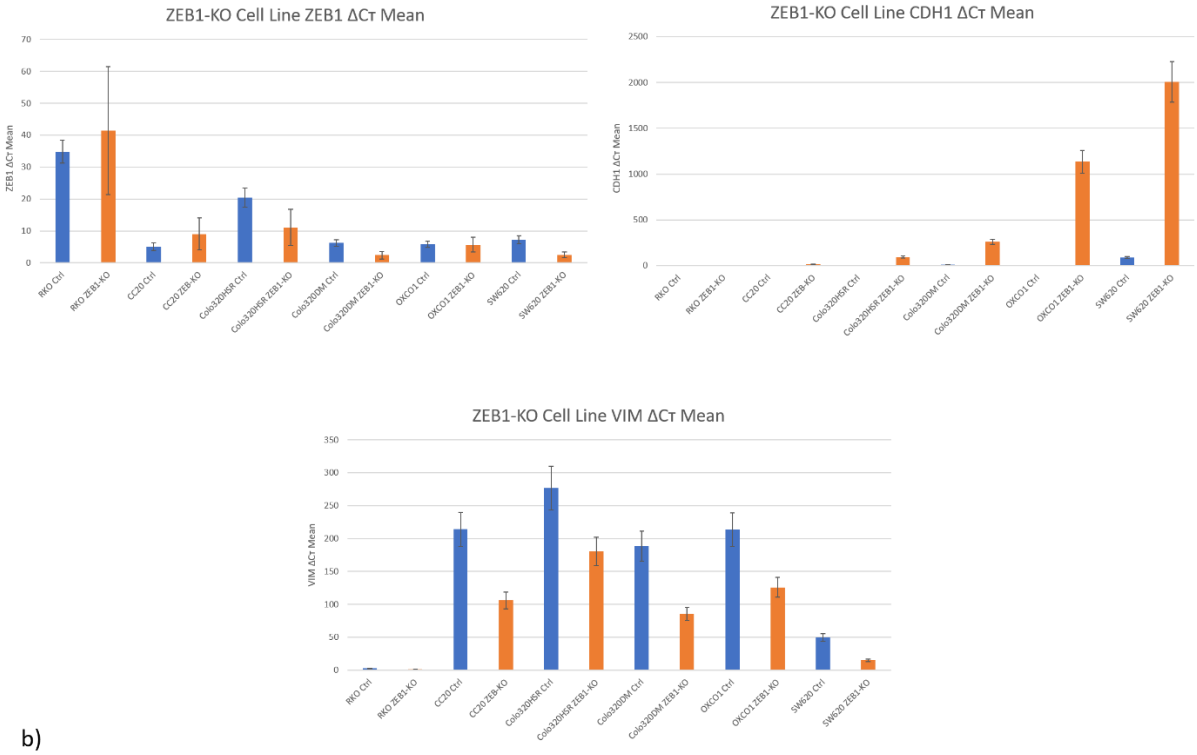
Figure 5.13 Immunofluorescence Staining of β-catenin Reveals Nuclear vs Membrane Staining: Immunofluorescence of β-catenin comparing the nuclear staining vs membrane staining in OXCO1, SW620, and OXCO1 ZEB1-KO, SW620 ZEB1-KO, respectively




SW620 ZEB1-KO
SW620 Wildtype
DAPI
B-catenin
Merge
10 µm

In Figure 5.14a, we show the qRT-PCR results of our ZEB1-KO cell lines. Interestingly, even though the result in western blot reconfirmed an ablation in ZEB1 expression in SW620, OXCO1, Colo320HSR, and Colo320DM, the qPCR result revealed a rather mixed pattern. There is a measurable reduction in Colo320DM, Colo320HSR, and SW620; yet no change in OXCO1. Nonetheless, the reduction in expression level for all three cell lines that revealed a measurable reduction did not indicate an ablation in ZEB1 expression rather only a reduction in expression level. Not to mention no change in OXCO1. We investigated the chromosome location of the target site of the ZEB1 Taqman probe we used in our experiment shown in Figure 5.14b. The target site is located on Chr.10: 31318495 – 31529814 or exon 3-9 depending on the ZEB1 splice variant. Our gRNA that generated a successful ZEB-KO confirmed with protein level ablation targeted ENSE000037015 on Chr.10: 31514603 – 31514708 in which suggested that the premature stop codon that led to nonsense-mediated decay is situated after the target site of our ZEB1 Taqman probe. In other words, the truncated ZEB1 mRNA before the premature stop codon is likely measurable with qRT-PCR resulting in a slight or no reduction in ZEB1 mRNA expression. Nonetheless, we observed a strong increase in CDH1 expression with qRT-PCR across Colo320DM, Colo320HSR, OXCO1, and SW620, confirming the effect of ZEB1 protein ablation. The lack of CDH1 increase in RKO and CC20 also reaffirms the lack of ZEB1 knockout in the two cell lines. Moreover, we observed a reduction of VIM, consistent with expectations with ZEB1 knockdown, in Colo320DM, Colo320HSR, OXCO1, SW620, and, interestingly, CC20.

a)



b)

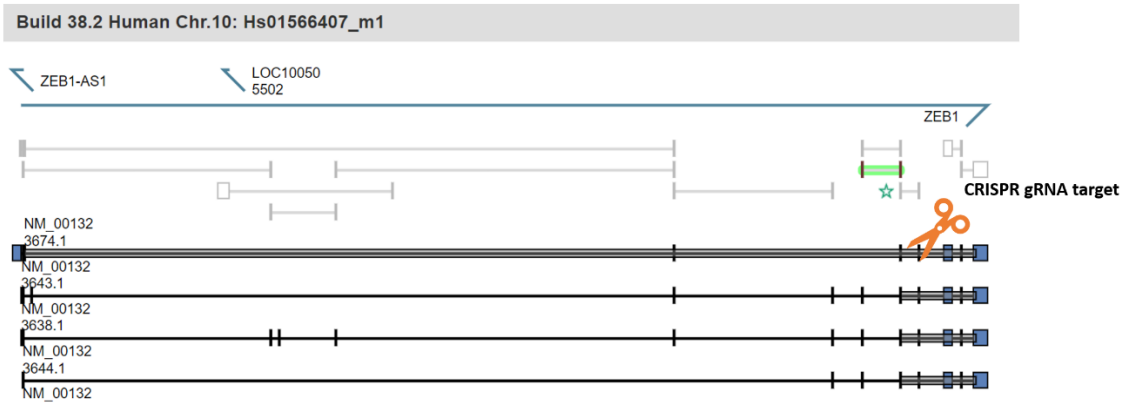


Figure 5.14 qRT-PCR Results of ZEB1, VIM, and CDH1 mRNA Levels in ZEB1-KO Cell Lines: (a) Result of the qRT-PCR on ZEB1-KO and ZEB1 wildtype cell lines. qPCR on ZEB1 reveals a non-uniform decrease in ZEB1 mRNA expression; however, showing a consistent upregulation in CDH1 expression. This is because the TaqMan probe Hs01566407_m1 targets a site that is located before our CRISPR gRNA target knockout site, denoted as an orange scissor in (b). This shows why despite the confirmed knockout of ZEB1 on a protein and genetic level, the epitope situated before our knockout target site is still detectable via qPCR.

Next, we explored whether there is a decrease in migration with a scratch assay. In the scratch assay, we compared OXCO1 and SW620, cell lines confirmed with a ZEB1-KO, as well as RKO and CC20, cell lines without complete ZEB1 ablation. Figure 5.15a shows the percent change in closure area comparing ZEB1-KO with its paired wildtype cell line. As shown, OXCO1 had a 32% and 53% decrease in wound closure at 24 and 48 hours, respectively, compared to wildtype; similarly, SW620 had an 8% and subsequent 58% decrease in wound closure on 24 and 48 hours compared to wildtype, indicating that both ZEB1-KO cell lines had a significantly slower wound closure compared to its wildtype counterpart. Comparing them with RKO and CC20, we only see an 11% decrease and 1.3% increase in wound closure in RKO as well as a 0.9% increase and 10.3% decrease in CC20, showing a minimal change in the percent change in the closure area. In Figure 5.16, we observed that many of the wild-type cell lines such as SW620 and OXCO1 have many more cells in the wound that may suggest migrating cells compared to ZEB1-KO counterparts. To further assess whether the migration is caused by an increase in proliferation or purely due to changes in the migratory ability of the cell lines, we assessed the cell doubling time of all ZEB1-KO cell lines as well as their paired wildtype counterparts. In Figure 5.15b, we showed the cell doubling time of ZEB1-KO cell lines did not significantly change from its wild-type counterpart, indicating that ZEB1-KO does not alter cellular proliferation but only cellular migration. The graph showing the doubling time over 72 hours is also shown in Supplementary Figure S3.

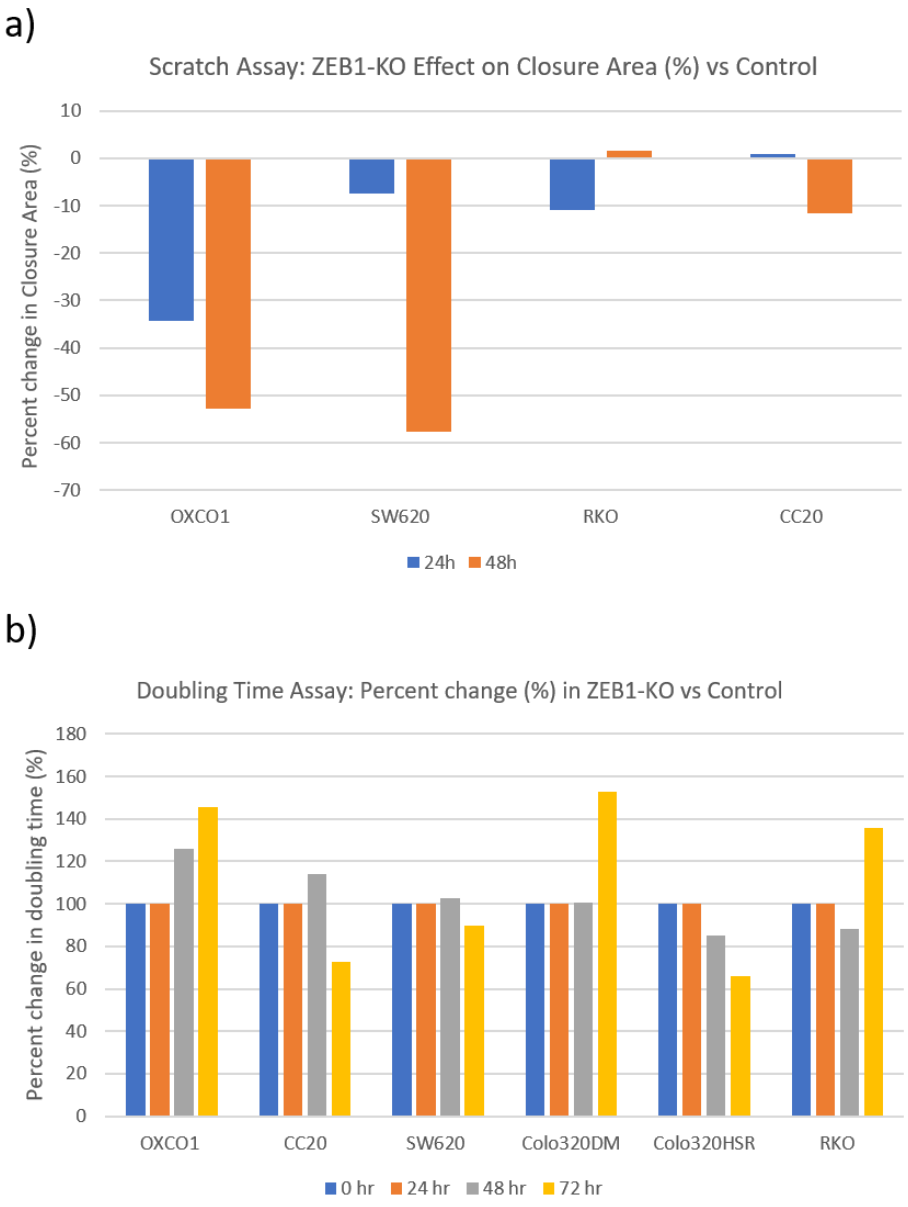
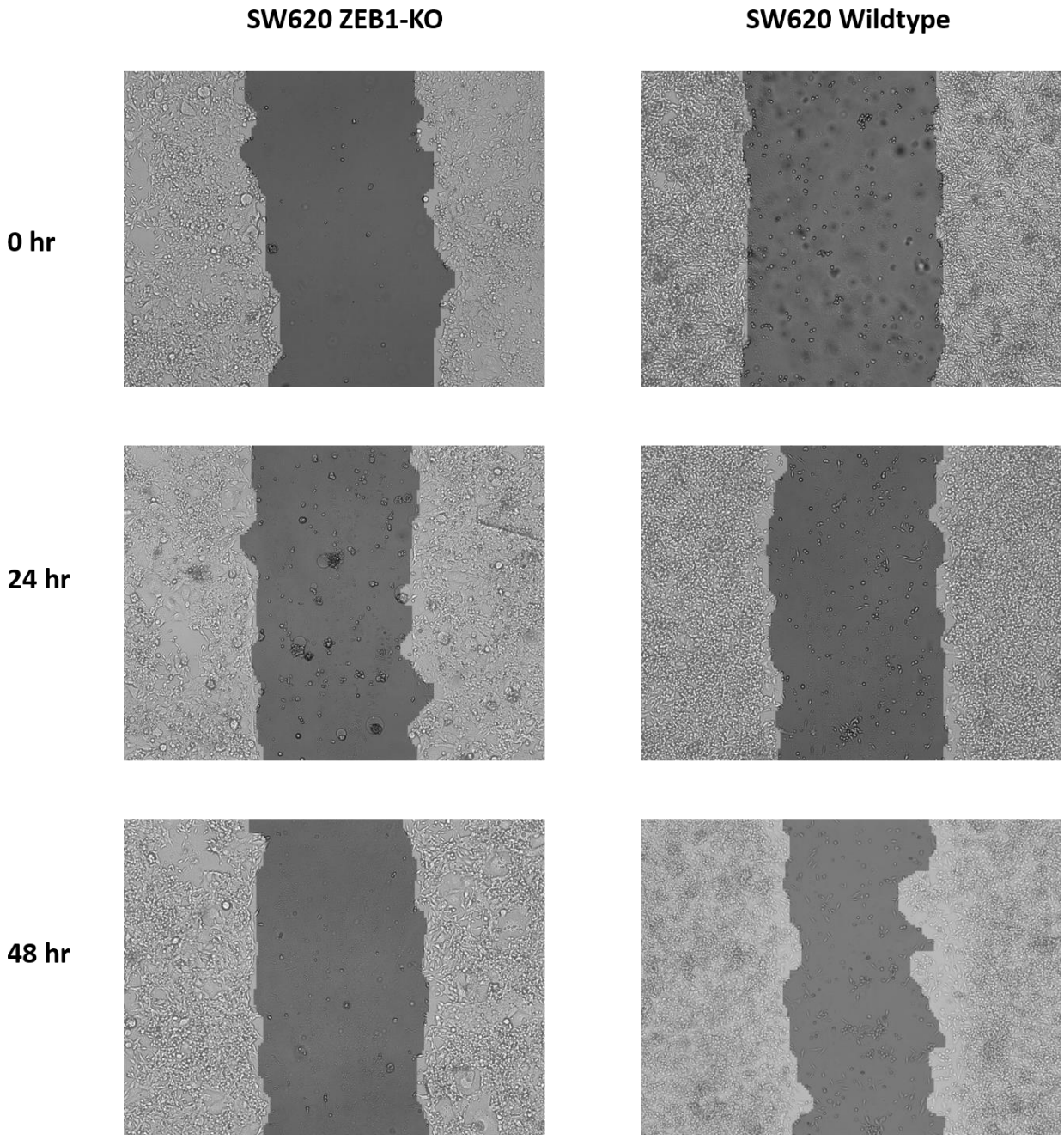
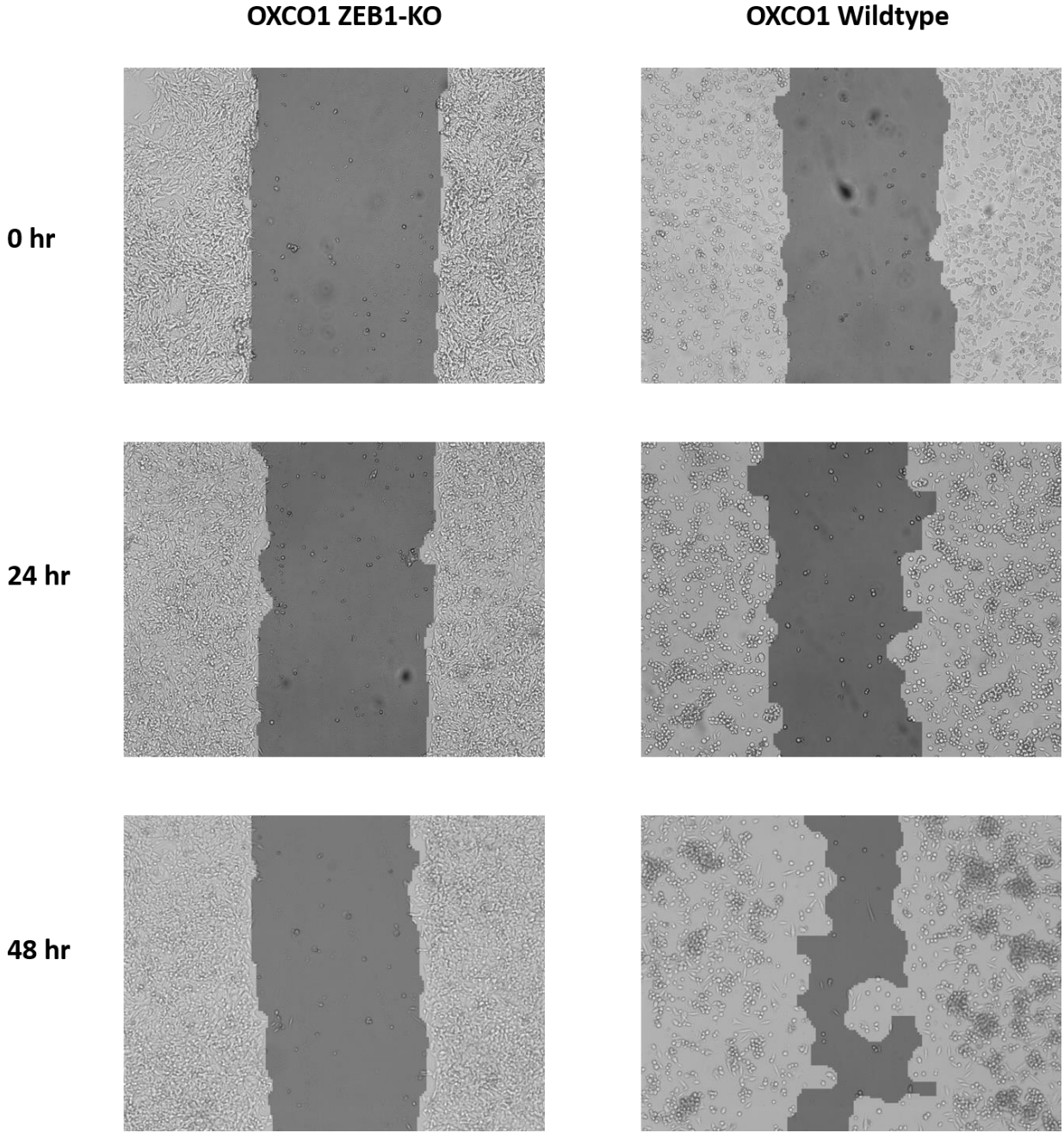
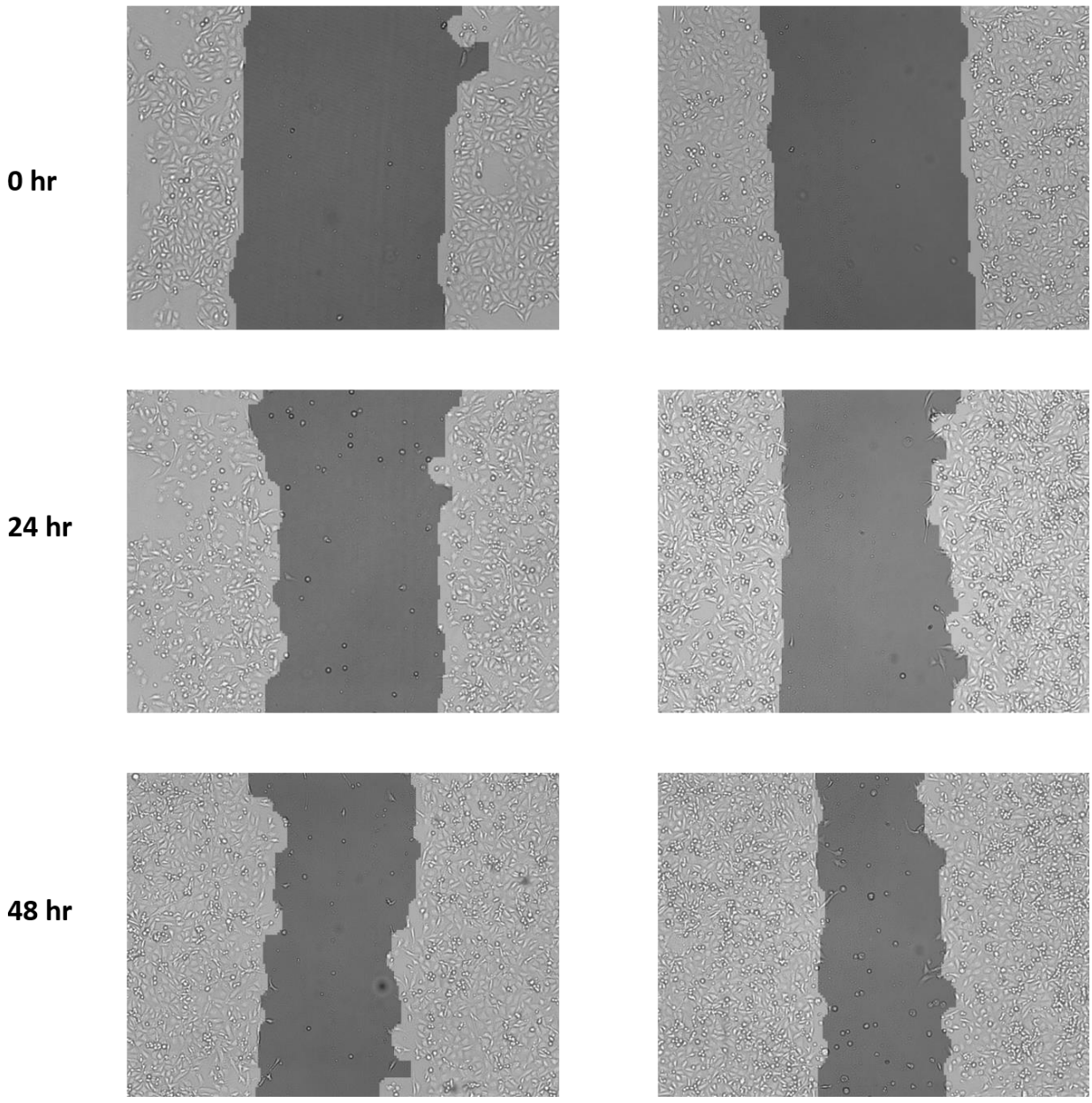


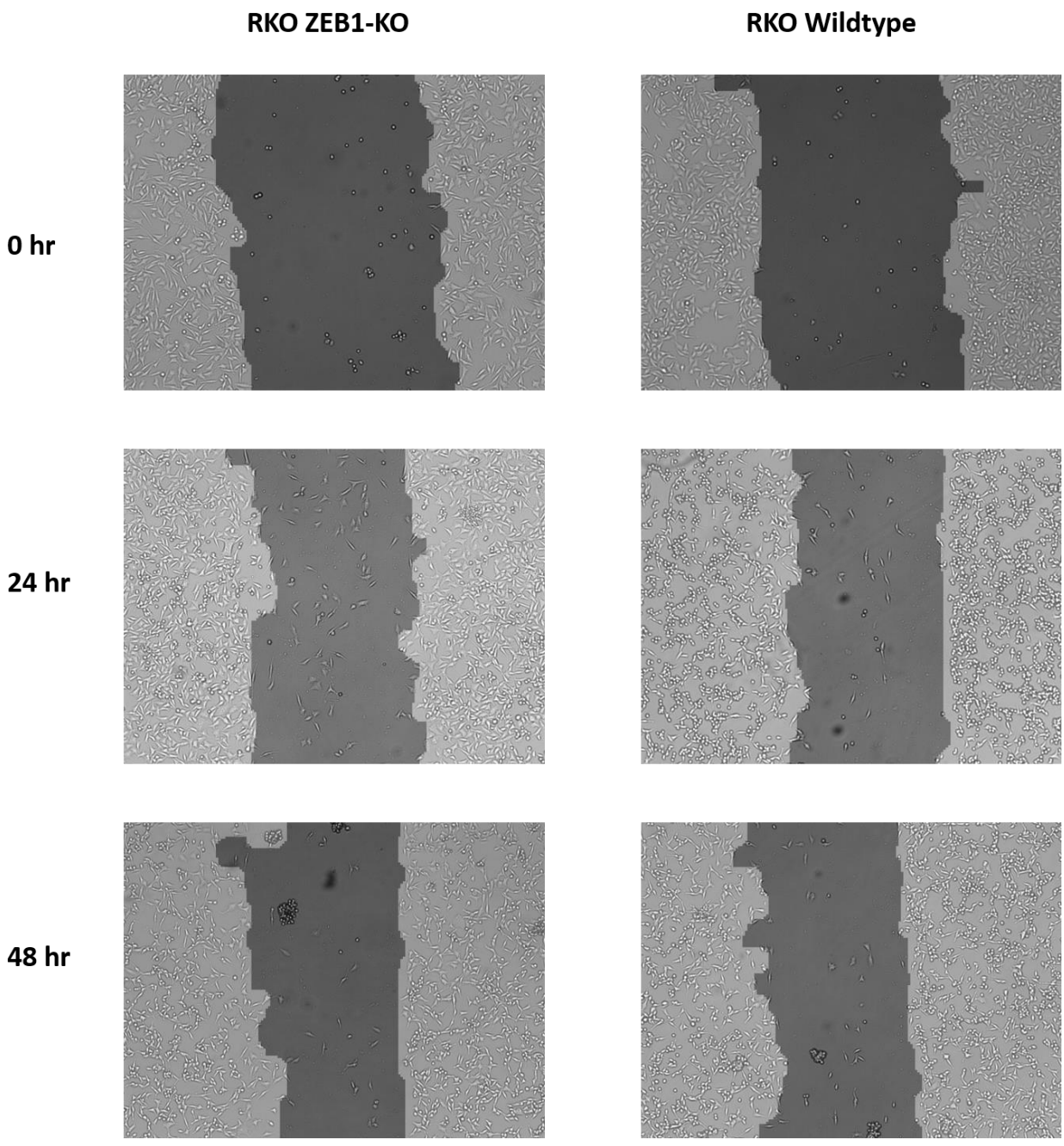
Figure 5.15 Results of Scratch Assay and Doubling Time Assay Comparing ZEB1-KO vs Wildtype Cell Lines: (a) shows the percent change closure area (%) in the scratch assay comparing ZEB1-KO against the wildtype counterpart. OXCO1 and SW620 ZEB1-KO, which underwent a reversal in EMT hallmark biomarkers with the upregulation of CDH1 and downregulation of VIM, have the most striking change in the ability to migrate compared to RKO and CC20. (b) shows the change in doubling time in ZEB1-KO vs wildtype. Here, we show that doubling time is not significantly affected by ZEB1-KO thus suggesting the change in the percent change closure area to be due to cellular migration instead of the change in the rate of cellular proliferation.

Figure 5.16 Scratch Assays of ZEB1-KO vs Wildtype EMT Cell Lines: Examples of the scratch assays of ZEB1-KO vs wildtype in SW620, OXCO1, CC20, and RKO.




OXCO1 ZEB1-KO
OXCO1 Wildtype
0 hr
24 hr
48 hr


CC20 ZEB1-KO
CC20 Wildtype
0 hr
24 hr
48 hr


RKO ZEB1-KO
RKO Wildtype
0 hr
24 hr
48 hr

5.4 Exploring ZEB1 activation in CRC cell lines

After establishing and characterizing ZEB1-knockout clones, we aimed to understand potential upstream regulators of ZEB1 in our panel of EMT CRC cell lines. Many pathways are implicated in the literature to be possible upstream regulators of the ZEB1 pathways including the hypoxic, RTK, and TGFβ pathways. Although TGFβ is one of the most well-known and studied inducers of ZEB1 expression, we have two main reasons why we did not prioritize TGFβ as a driver for ZEB1 expression in our CRC cell lines: Firstly, FBS contains a high level of latent TGFβ, if TGFβ is the primary driver of ZEB1 expression, we should have observed ZEB1 expression in all cell lines. Secondly, we have not observed TGFβ secretion in cell lines during differential expression analysis. Our initial rationale is to look at the hypoxia-inducible factor-1 (HIF-1) pathway based on various indications from the literature, Wang J. *et al.* have shown a direct binding of HIF1A-ARNT complex to the hypoxia-responsive element (HRE) located on the proximal promoter of ZEB1 (Zhang *et al.*, 2015). One important mechanism of the HIF-1 pathway that is important to note is that under hypoxic conditions, mRNA level would not reveal any significant changes, only protein expression level changes. This is worth noting since most of our analysis is done on our microarray data. We did two separate experiments exploring the effect of hypoxia on ZEB1 expression: Firstly, we explored whether hypoxia could induce ZEB1 expression in CCO7 and SW480 which showed low ZEB1 expression in the qPCR/WB screen previously discussed. Secondly, we wanted to see if hypoxia can induce ZEB1 in non-EMT CRC cell lines C80 and C125PM. In Figure 5.17, both results showed that hypoxia is unable to induce any significant ZEB1 upregulation.

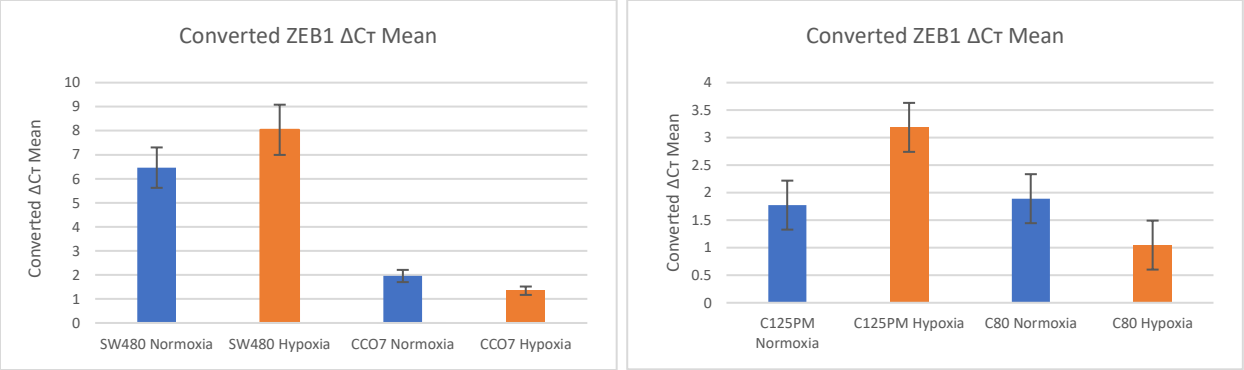


Figure 5.17 ZEB1 qPCR Comparing Normoxic vs Hypoxic Conditions: qRT-PCR on ZEB1 expression of normoxia vs hypoxia. a,b) qRT-PCR on ZEB1 expression of SW480, CCO7, C125PM, and C80 under normoxic (~20% O_2) and hypoxic (~0.1% O_2) revealing no significant upregulation of ZEB1 expression.

Another attempted experiment is knocking down AKT3 to observe whether there is a downregulation in ZEB1 expression. AKT3 is one of the most significantly differentiated genes between the EMT and non-EMT group (Table 5.1), in addition, the AKT pathway is implicated to be a key path for EMT TF induction. However, our results suggest siAKT3 knockdown did not affect ZEB1 expression (Figure 5.18).

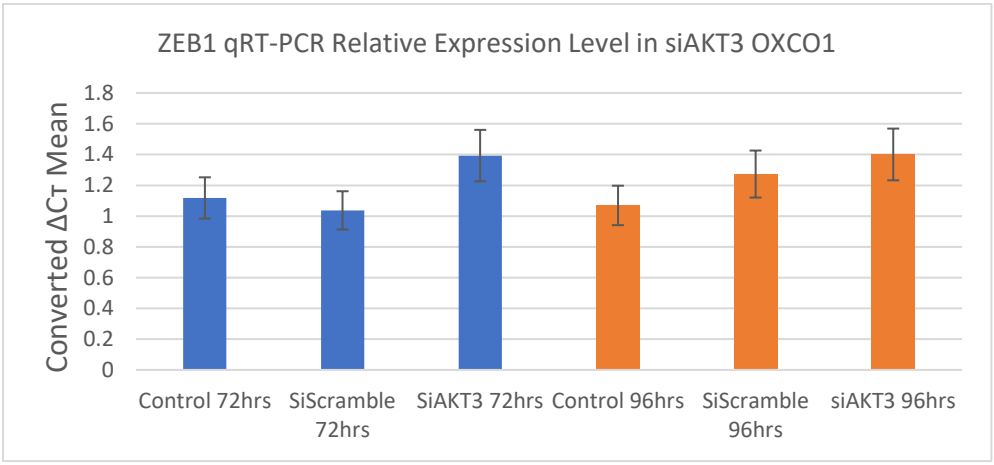


Figure 5.18 siAKT3 Knockdown Experiment Reveals No Change in ZEB1 Expression: qRT-PCR on ZEB1 expression on siAKT3-KD OXCO1. qRT-PCR on ZEB1 expression of siAKT3 OXCO1 did not reveal any significant change in ZEB1 expression.

This urged us to wonder about two important factors: Firstly, a mechanism for a stable mesenchymal cell line. Secondly, what factors could potentially block the interaction of HIF1A-ARNT and the HRE that has been shown to induce ZEB1 expression. We think epigenetic regulation may be the key. We revisited and reanalyzed a set of microarray data containing 11 CRC cell lines comparing demethylated cell lines using 5-azacytidine with control cell lines done in our lab previously by another DPhil student (Figure 5.19a). In Figure 5.19b, a volcano plot of the $\log_2$(fold change) and the $-\log_{10}$(p-value) of the fold change across control and demethylated cell lines are shown. It is striking that ZEB1--2 revealed an extremely specific pattern where ZEB1 expression is downregulated in all three EMT cell lines (CC20, NCIH716, and RKO) under demethylation while eight non-EMT cell lines had nearly remained unchanged with no ZEB1 expression. This suggests two possible scenarios:

1) Gene X capable of inhibiting ZEB1 expression is demethylated.
2) Gene Y capable of inducing gene X that can inhibit ZEB1 expression is demethylated.

Nonetheless, when we again dived deeper into the table in Figure 5.19a where we listed out all EMT TFs and the CDH1 and VIM fold change comparing each cell line and its 5-azacytidine treated counterpart, we noticed that not only ZEB1--2 had a significant pattern associated with VIM and CDH1 but ZEB1--4 as well. Surprisingly, ZEB1--4 had an inverse relationship with ZEB1--2 in these cell lines, which suggested that demethylation increased ZEB1 expression and, interestingly, upregulated VIM and downregulated CDH1 expression concurrently. This striking relationship allowed us to hypothesize that ZEB1--4 maybe the more likely scenario if ZEB1 is to be controlled by methylation. Nonetheless, we proceeded with 5-aza treatment experiments to validate either scenario.

a)

	Non-EMT								EMT		
Gene	DLD-1	HCA7	HCT116	LS174T	LS411	PC-JW	SW48	Vaco5	CC20	NCI-H716	RKO
SNAI1	0.65	1.94	2.03	0.7	0.14	-1.43	0.67	0.69	0.7	0.23	1.87
SNAI2	-0.17	0.03	2.83	-0.04	0	-0.09	0.11	1.99	1.98	-0.68	-0.1
SNAI3	-0.06	-0.03	-0.07	-0.01	0	0	0.39	-0.14	-0.01	-0.01	-0.02
TWIST1	6.23	0.19	4.4	-0.09	0	0.52	0.13	0.02	3.55	-0.05	-0.05
ZEB1--2	-0.08	0.01	0.16	-0.03	0	-0.01	0.07	-0.01	-1.89	-3.11	-2.25
ZEB1--4	4.86	2.28	3.32	-0.02	2.29	-0.02	2.22	4.55	0.68	0.43	-0.76
VIM	5.06	4.75	8.08	0.45	3	0.14	2.97	6.79	-0.3	-0.34	5.22
CDH1	-2.05	-1.82	-1.97	-1.52	-0.82	-1.94	-1.45	-0.78	-0.05	-0.02	-0.74

b)

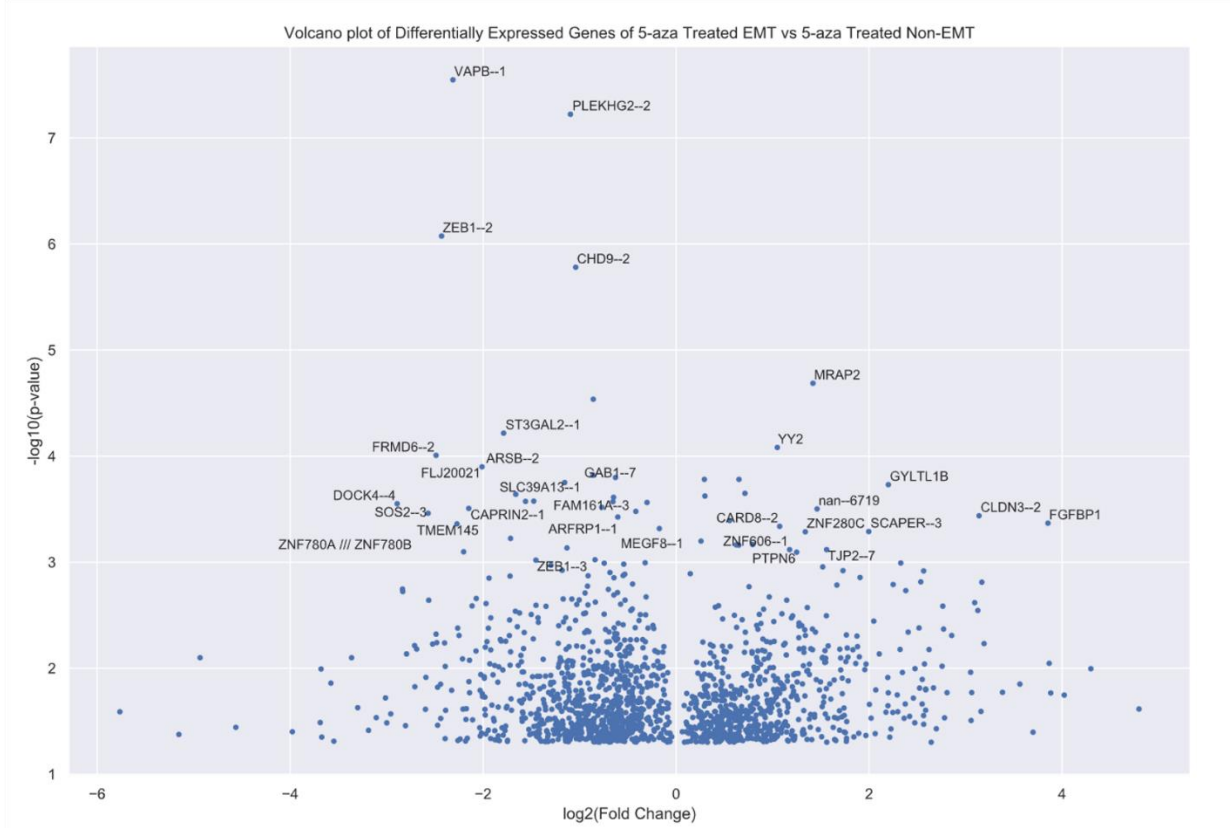


Figure 5.19 5-azacytidine Treatment Upregulated ZEB1 Expression in Non-EMT Cell Lines: (a) Table showing the fold change in the fold change of Control vs 5-aza treated cell lines (b) Volcano plot of control vs demethylation in 11 CRC cell lines. Volcano plot of $\log_2$(fold change) vs -$\log_{10}$(p-value) of the fold change of cell lines comparing control and demethylation with 5-azacytidine. The plot shows a significant downregulation in ZEB1 after demethylation suggesting genes that are turned on that regulate ZEB1.

5.5 5-aza treatment in non-EMT cell lines

We explored both scenarios of ZEB1--2 and ZEB1--4 by repeating the 5-aza treatment experiment on 3 EMT cell lines (CC20, NCIH716, and RKO) vs 3 non-EMT cell lines (DLD1, VACO, and HCT116) to investigate whether we can replicate the trend observed in ZEB1--2 or ZEB1--4 listed in the Figure 5.19a. In the qPCR result shown in Figure 5.19b, it is evident that ZEB1 is highly upregulated when non-EMT cell lines are treated with 5-azacytidine, but the effect is muted in EMT cell lines. This result reflects the trend of ZEB1--4 thus we consider ZEB1--4 to be the likely scenario. Nonetheless, in the CDH1 qPCR result, we see an upregulation of CDH1 consistently throughout the non-EMT cell lines despite the upregulation of ZEB1. We suspect this may be due to 5-azacytidine acting as a pan-demethylation agent that may activate a complex network of methylated genes in which resulted in this observation. It is intriguing as to why our result of CDH1 expression differs from the previous experiment done and shown in Figure 5.20d.

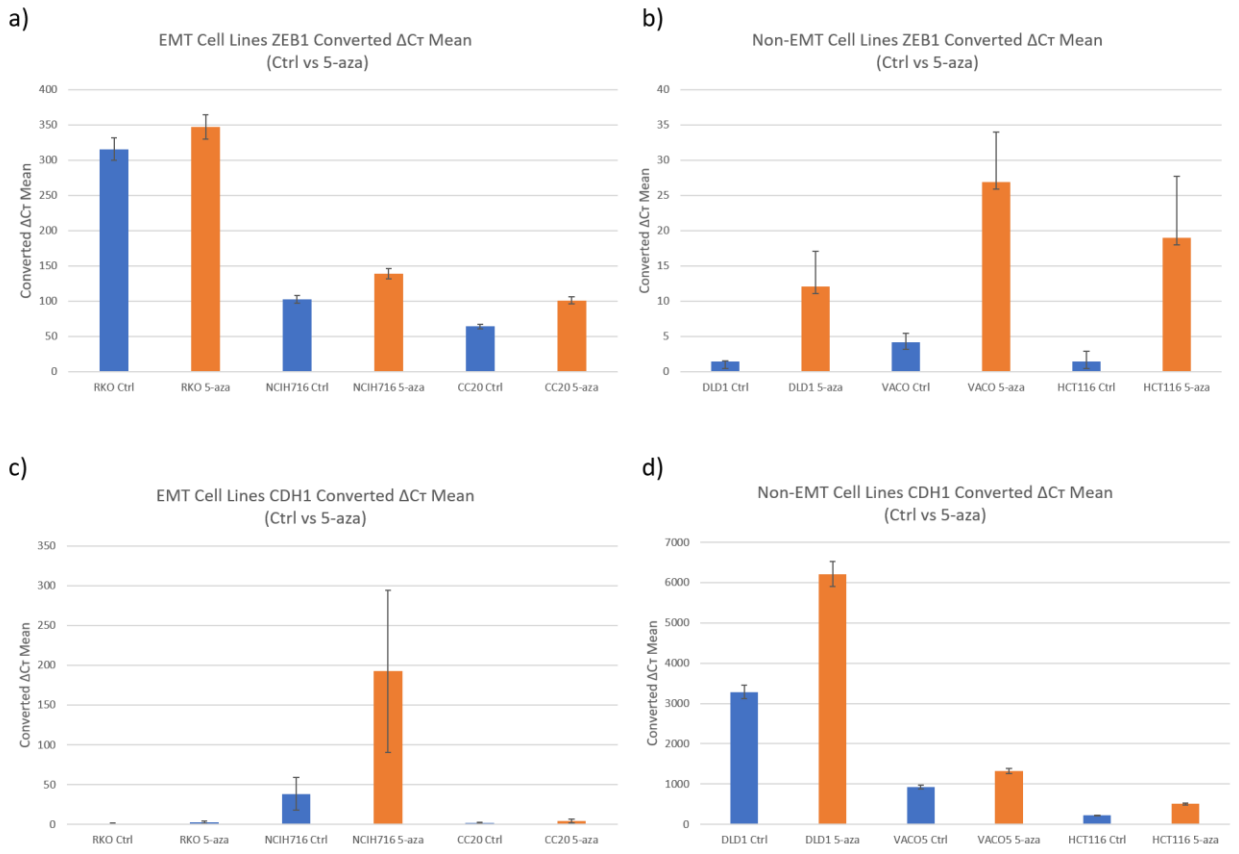


Figure 5.20 ZEB1 qRT-PCR on 5-aza-treated Non-EMT Cell Lines: In (a) and (b), the result of qRT-PCR on ZEB1 comparing control vs 5-azacytidine treated cell lines of EMT (RKO, NCIH716, CC20) and non-EMT (DLD1, VACO, HCT116) cell lines, respectively. As expected, ZEB1 expression had very little change after 5-azacytidine treatment but highly upregulated in our panel of non-EMT cell lines. (c) and (d) shows the qRT-PCR on CDH1 on the same panel of cell lines. Here, we observe CDH1 unexpectedly upregulated after 5-azacytidine treatment despite the upregulation of ZEB1, suggesting other possible mechanisms of CDH1 regulation via pan-demethylation which may activate other methylated genes that upregulates CDH1.

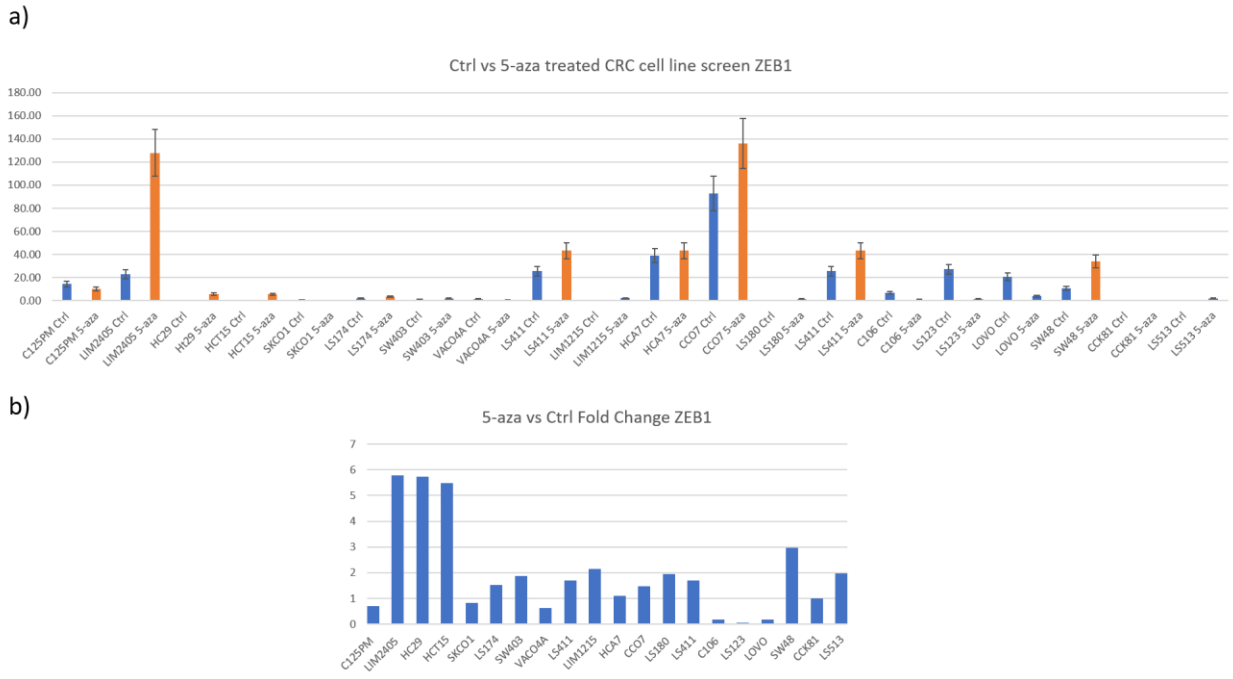


Figure 5.21 ZEB1 Upregulation Observed in 5-aza-treatment on a Panel of 20 CRC Non-EMT Cell Lines: (a) Shows the qRT-PCR result on 5-azacytidine treatment on a panel of non-EMT CRC cell lines (n=20) and (b) shows the fold change of 5-azacytidine treated vs control in each cell line.

5.6 Methylation of ZEB1 in large datasets

To assess the role of methylation controlling ZEB1 expression in non-expressing cell lines, we performed a cell line-wide 5-azacytidine screen (n = 20). In Figure 5.21a and 5.21b, we see that most cell lines have a general upregulation of ZEB1 across all 20 cell lines. Cell lines including LIM2405, HT29, HCT15, and SW48 had over 5-fold change after 5-aza treatment. This provides a piece of strong evidence that reinforces methylation as the upstream regulator of ZEB1.

Next, we explored available datasets on methylation such as the CCLE methylation data discussed in Chapter 4 and methylation data downloaded from TCGA. We began with the analysis of ZEB1 methylation in TCGA methylation data. In Figure 5.22, we compared the beta value, representing the level of methylation ranging from 0 to 1 with 1 being fully methylated and 0

being unmethylated. We explored CpG islands located upstream of the ZEB1 promoter in high ZEB1-expressing vs low ZEB1-expressing tumor samples. ZEB1 expression level in the TCGA RNA-seq data was categorized as unimodal via GMMchi shown in Figure 5.22a. As shown in Figure 5.22a, the way we defined high and low ZEB1-expressing tumor samples is by defining tumor samples with ZEB1 expression levels below the background as low-expressing samples while only taking the 95^{th} percentile as the highest ZEB1-expressing tumor samples to compare them. The purpose is to look at non-expressing samples vs the highest expressing samples to explore the biggest difference in methylation. We showed that in Figure 5.22b there is a noticeable difference in methylation levels on specific locations such as 31320962, 31320990, 31320967, 31321031, and 31320976 which are likely candidates for ZEB1 methylation.

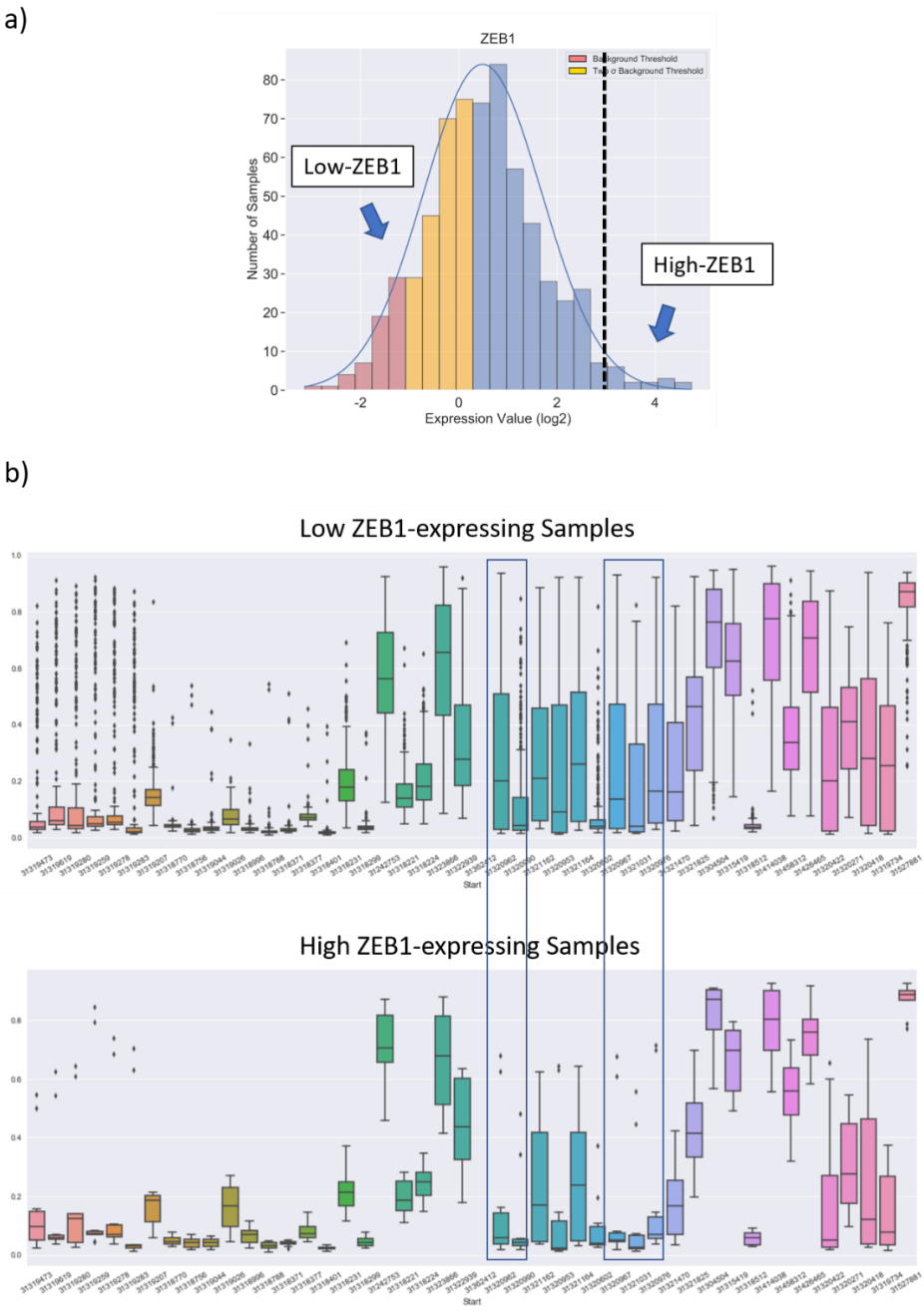


Figure 5.22 ZEB1 is Methylated in Low ZEB1-Expressing Samples in The TCGA Data: (a) shows the distribution of ZEB1 expression level in the TCGA RNA-seq data. ZEB1 exhibits a unimodal pattern in our GMMchi analysis. To understand the difference between non-expressing vs high ZEB1-expressing patient samples in methylation, we divided patient tumors into two groups by grouping patient tumors with ZEB1 expression level below background level vs the top 95th percentile. The result is shown as a box plot in (b) of the distribution of β value in each CpG island located near the ZEB1 promoter. We observe locations delineated in the blue box as sites with noticeable differences in methylation levels, revealing likely candidates for ZEB1 methylation.

In Chapter 4, we also found potential candidates for ZEB1 methylation sites using CCLE methylation data and our in-house microarray data with ZEB1 probe set 4 and location 31606675, 31606797, and 31606814 of chromosome 10. This is an important observation because ZEB1 probe set 4 is the probe set we have previously identified as the ZEB1 probe set that is shown to correlate with the upregulation of VIM and downregulation of CDH1 through experimental work with 5-azacytidine treatment. Although only 2 EMT cell lines are found in the CCLE methylation data, this, nonetheless, still shows a striking correlation and significance that reinforces the experimental work. Thus, this would suggest a highly likely mechanism of ZEB1 methylation and its control of EMT in CRC cell lines. The 2x2 table of ZEB1 probe set 4 and ZEB1 methylation site is shown in Figure 5.23. Putting it all together, these observations suggest that ZEB1 is controlled by methylation in Colorectal cancer tumor development, revealing a novel mechanism where ZEB1 is methylated in normal Colonic tissue while demethylated in Colorectal cancer which promotes CRC to undergo EMT. This may further provide insight into understanding the initiation of metastatic CRC and the mechanisms of ZEB1-mediated EMT in CRC.

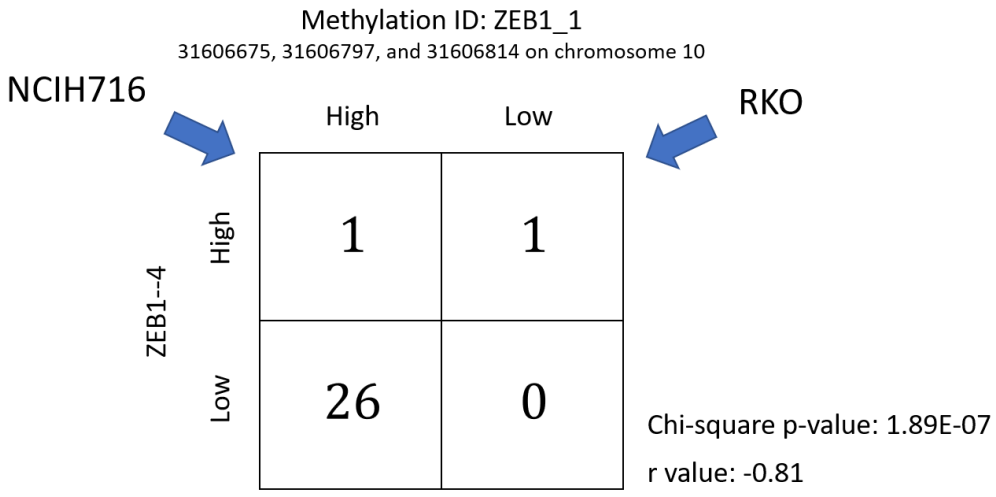


Figure 5.23 ZEB1 Expression is Negatively Correlated with ZEB1 Methylation in the CCLE Database: 2x2 table analysis of ZEB1--4 and ZEB1_1, which is the methylation ID used in the CCLE methylation dataset that represents CpG islands located 31606675, 31606797, and 31606814 on chromosome 10. Here, we show that there is a significant negative correlation between ZEB1 expression and methylation, further reinforcing the finding in our experimental work with 5-azacitydine treatment on non-EMT cell lines.

Chapter 6 Myofibroblast CRC Interaction and PDGFC

6.1 Introduction

The microenvironment of the colonic crypt is surrounded by an extracellular matrix (ECM) that maintains a three-dimensional structure with different cell types that support the colonic stem cell niche. The stem cell niche has a rapid turnover with differentiating cells that move up along the crypt axis (Fre, 2015). Myofibroblasts that line the colonic crypt secrete collagen type IV, the main collagen component of the basement membrane that makes up the thin layer of ECM separating them from the epithelial cells, as well as signaling molecules such as BMP and WNT ligands to interact and maintain the homeostasis of the crypt (Klingberg, Hinz and White, 2013; David *et al.*, 2020). This adjacent structure provides neighboring crosstalk between the two cell types that we will further explore in this chapter.

It has been suggested that the recruitment of cancer-associated fibroblasts (CAFs) or tumor-associated fibroblasts to the tumor niche is mainly initiated with cytokines secreted by cancer cells and infiltrating leukocytes (Sahai *et al.*, 2020). In solid cancers, tumor with myofibroblast invasion is known to exhibit a more severe form of cancer. Moreover, myofibroblasts are found within the tumor setting in varying amounts correlated with the stage of the tumor (Tsujino *et al.*, 2007). Nonetheless, myofibroblasts are found in the normal physiological setting in tissues including lymph nodes, blood vessels, lung septa … etc, suggesting that there is likely a family of different types of myofibroblasts (Bochaton-Piallat, Gabbiani and Hinz, 2016). We have previously shown in our lab that CAFs not only express myofibroblast biomarker vimentin (VIM) and α-smooth muscle actin (ACTA2) but also a specific set of markers AOC3+, NKX2-3+, SHOX2- shown to distinguish myofibroblasts from skin fibroblasts (Hsia *et al.*, 2016). In wound healing, transforming growth factor β1 (TGFβ1) and platelet-derived growth

factor (PDGF) have been shown to activate myofibroblasts in the initial inflammatory phase of healing, which promotes it to remodel the ECM and secrete cytokines and chemokines that further drive the inflammatory response (Pakyari *et al.*, 2013). In a tumor setting, CAFS, similar to wound healing, is known to secrete important signaling molecules that have been shown to promote tumor proliferation and invasion (Sun *et al.*, 2019).

The bidirectional crosstalk between myofibroblasts and cancer cells is an important biological observation that is related to fostering cancer growth and potentially contributing to tumor malignancy. In this chapter, we will utilize a coculture system using a panel of CRC, intestinal myofibroblast cell lines, and skin-derived fibroblast to study the recruitment of myofibroblast into CRC cell Matrigel blobs previously established by our MPhil student Peter Kalugin (Peter Kalugin, 2017). Based on previous results from our lab, we have narrowed down PDGFC and PDGFA as important secreted ligands that contribute to the bidirectional talk between CRC cell lines and myofibroblasts. Here, we will attempt to establish PDGFC-KO CRC cell lines in order to study the effect of PDGFC on myofibroblast recruitment.

6.2 Results

6.2.1 Myofibroblast Recruiting CRC Cell Lines

The interaction between epithelial cells and myofibroblasts is of great interest due to the implication of a worse prognosis with fibroblast-infiltrating tumors as well as a major player in the tumor microenvironment (TME). Thus, previous experiments were done on coculturing myofibroblasts and CRC cell lines to determine whether a cell line recruits surrounding myofibroblasts into the Matrigel blob CRC cell lines are cultured in. The goal of the analysis is thus exploring possible upregulated or downregulated genes that suggest cell-cell communication between the two cell types.

To provide some context, in Figure 6.1, we show a schematic of how the recruitment assay is carried out. At the end of the assay, the myofibroblast recruitment phenotype was scored on a scale of 1 to 5 with 1 being no recruitment and 5 being highly recruited. The images were scored by 3 independent reviewers and the average score is listed as the final score for the myofibroblast recruitment phenotype.

Next, we show the genes that were previously found to be differentially expressed in myofibroblast recruiting cell lines using differential gene expression analysis shown in the volcano plot in Figure 6.2. In brief, CXCR4, the gene that encodes CXC chemokine receptor specific for stromal-derived factor-4, and PDGFC as well as PDGFA, genes encoding platelet-derived growth factor C and A, are found to be the most promising candidates responsible for the cell-cell interaction in the recruitment assay (Figure 6.2). These results are produced by our previous student Peter Kalugin (Peter Kalugin, 2017). In Figure 6.3, we show the GMMchi analysis on both candidates PDGFC as well as CXCR4. Both genes exhibit a bimodal distribution suggesting two

distinct groups of cell lines in either case where most of the cell lines express little or no mRNAs from PDGFC or CXCR4. Only 3 cell lines expressed high levels of CXCR4 and PDGFC. This may suggest that a cell line only needs either gene to achieve myofibroblast recruitment. Moreover, our lab has previously shown that Imatinib, a tyrosine kinase inhibitor, inhibits the recruitment of myofibroblasts into the myofibroblast-recruiting CRC cell lines. The treatment with Imatinib targets the PDGF tyrosine kinase receptor thus shows that either PDGFC or PDGFA expressed in CRC cell lines and PDGFRA expressed in myofibroblast cell lines is responsible for the observed myofibroblast recruitment.

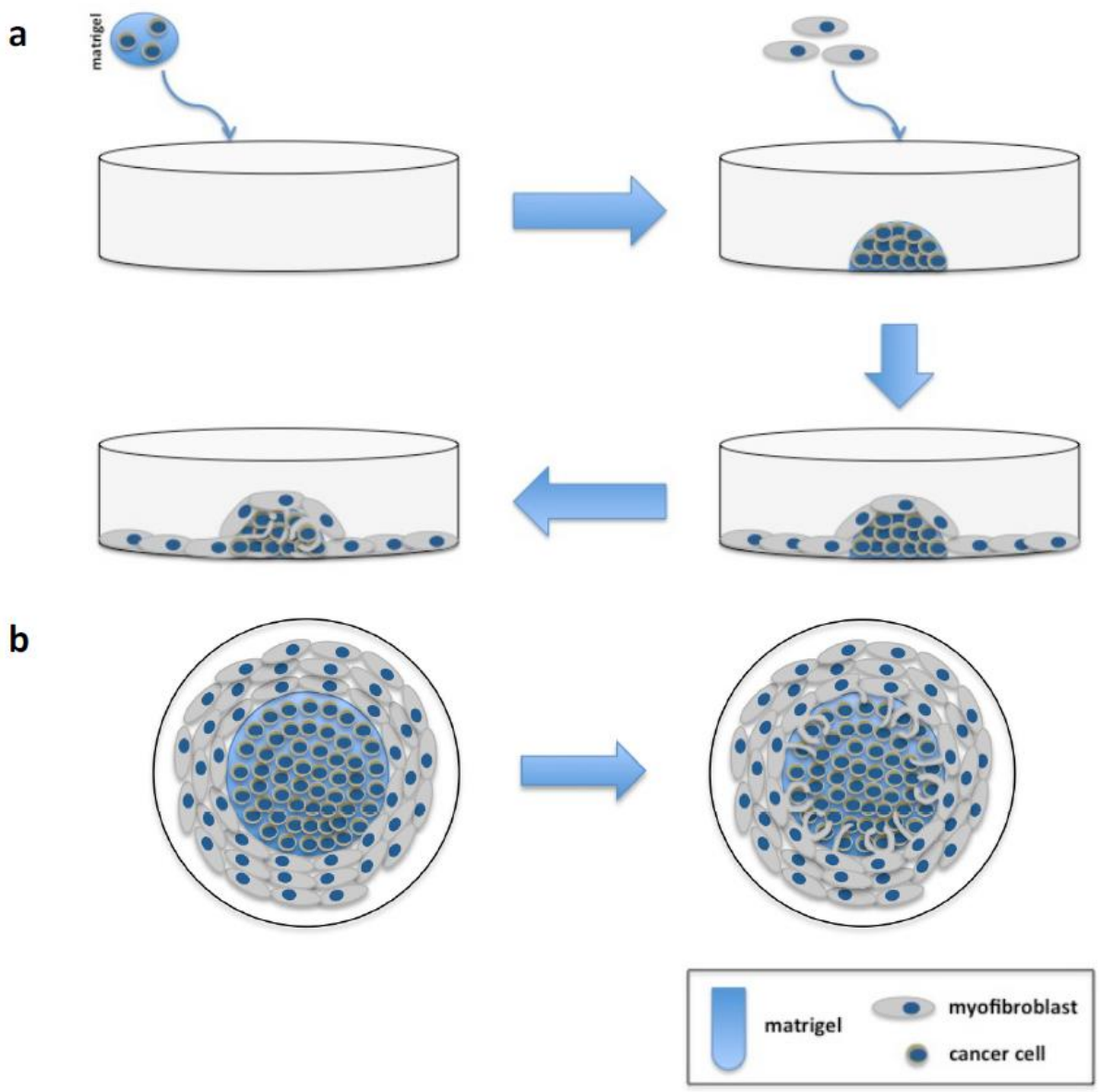


Figure 6.1 Schematic of Coculture Assay: (graph retrieved from Peter Kalugin's Thesis) (Peter Kalugin, 2017) (a) Cancer cells are mixed in 50:50 Matrigel: DMEM solution and placed in the middle of the culture well. Once the drop has hardened, myofibroblasts are added on top of the drop in the culture medium. The coculture is grown for a period of time before fixed and stained to observe whether myofibroblast has invaded into the cancer cell blob. (b) Top view of the culture well.

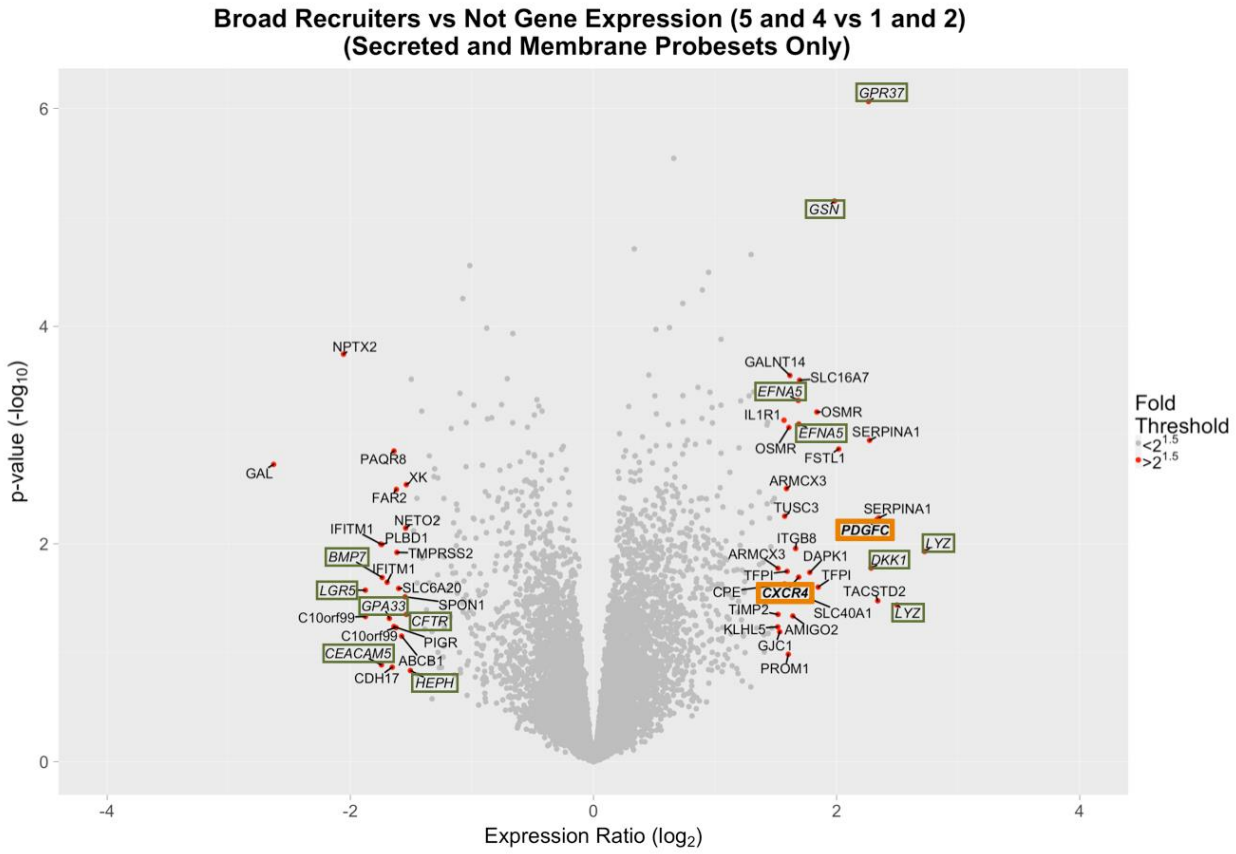


Figure 6.2 Volcano Plot Reveals PDGFC and CXCR4 as Potential Candidates that Contribute to Myofibroblast Recruitment: (graph retrieved from Peter Kalugin's Thesis) (Peter Kalugin, 2017) This is a volcano plot that compared differentially expressed genes across MF recruiting cell lines (5 and 4) vs non-MF recruiting cell lines (1 and 2). Here, only genes that are identified as secreted and membrane probes are shown. We can see a differential upregulation of PDGFC and CXCR4. They are the only probes boxed in orange because of previous inhibitor experiments that show the inhibition of PDGFC and CXCR4 successfully blocked myofibroblast recruitment in coculture assays.

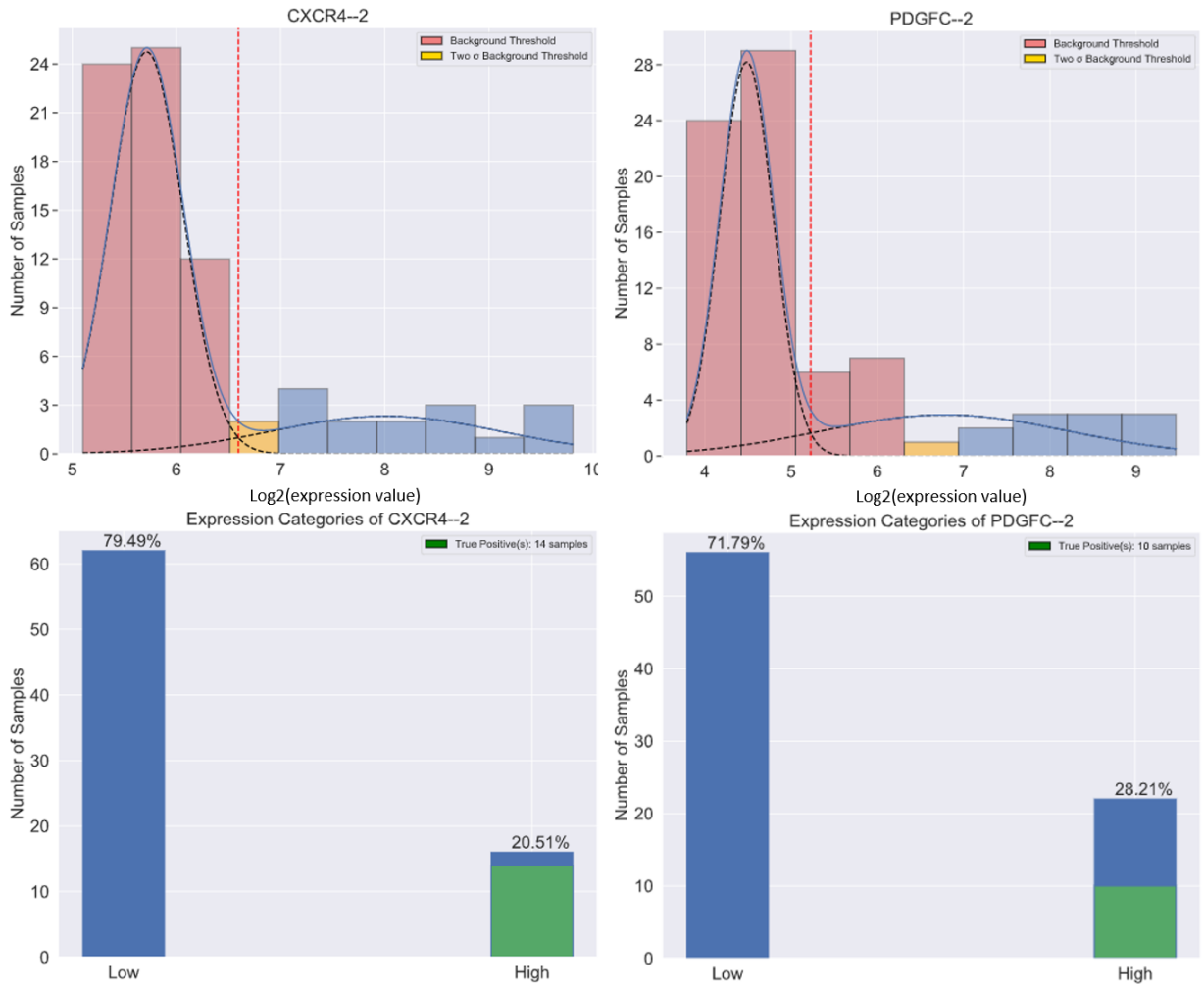


Figure 6.3 GMMchi Analysis on CXCR4 and PDGFC: GMMchi analysis on CXCR4 and PDFC reveals a similar pattern exhibiting a bimodal distribution, suggesting two distinct groups of cell lines. In the bar chart showing each expression category, blue represents putative negatives, and green shows the number of clear-cut positives.

6.2.2 Validation on MF Recruitment Phenotype using Different Myofibroblast Cell Lines

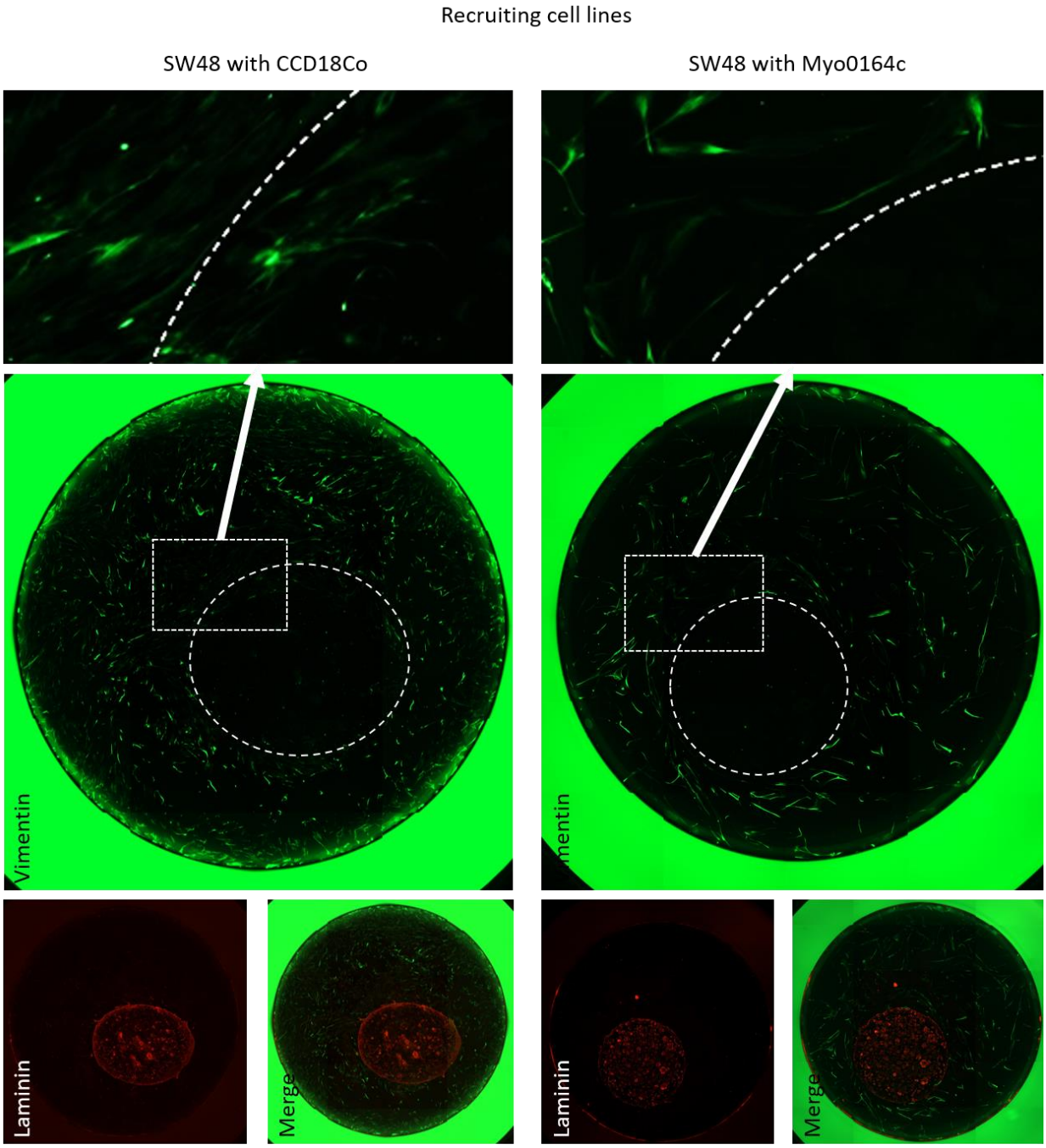
Previous recruitment assays were carried out using a single myofibroblast cell line CCD18Co. Our goal is to validate whether the cell lines classified as myofibroblast recruiting or myofibroblast non-recruiting based on the previous experiment show the same characteristics when cocultured with myofibroblast cell lines other than CCD18Co. We thus carried out the recruitment assay on two groups of CRC cell lines: myofibroblast-recruiting cell lines including

SW48 and OXCO1 as well as myofibroblast-non-recruiting cell lines including LS180 and SW620. All four CRC cell lines are then cocultured with myofibroblast cell lines including CCD18Co (control), Myo0164c, and Myo1998 using the recruitment assay. These myofibroblast cell lines were harvested in-house from patient samples by Marahaini Musa. The result is shown in Figure 6.4. Samples are stained with vimentin (green) as a myofibroblast marker and laminin (red) to delineate the Matrigel blob. One immediate observation was the difference in the level of density in Myo1998. This is mainly due to cellular senescence observed in myofibroblast cell lines where they usually dramatically slow down in terms of doubling rate after multiple passages. Many of the myofibroblast cell lines stop dividing at passage 10-15. Myo1998 is one of the examples where most of our stock has been exhausted and the vial we seeded had a passage number of 7. Thus, the doubling time compared to other myofibroblast cell lines was dramatically lower despite leaving it for 3 weeks to double. Nonetheless, the experiment was still carried out because even with the lower cellular density we can still observe whether myofibroblasts were able to invade into the Matrigel blob. Also, notice that EMT cell lines such as OXCO1 and SW620 were stained positive for Vimentin. The Matrigel blob is distinguished with Laminin IF staining.

We observe that myofibroblast-non-recruiting CRC cell lines exhibit a similar behavior regardless of the myofibroblast cocultured, with no recruitment of myofibroblasts into the Matrigel CRC blob delineated by the white dashed circle. On the other hand, although we observed recruitment into the Matrigel blob of myofibroblast-recruiting cell lines in both cases of CCD18Co and Myo1998, Myo0164c failed to exhibit any sign of recruitment or invasion into the Matrigel blob. We revisited the microarray data we have on a limited number of myofibroblast lines including CCD18Co and Myo1998 and found that both cell lines express high

levels of PDGFRA, which is the receptor for PDGFC and PDGFA, with both values greater than 2^{10}. Unfortunately, we do not have expression data on Myo0164c. We hypothesized that it may be possible that Myo0164c lacked the expression in PDGFRA, which may explain the lack of recruitment. A sensible experiment to continue to explore this observation is a simple qRT-PCR or western blot to detect the level of PDGFRA in Myo0164c. However, due to the pandemic, I did not have time to carry out this experiment.

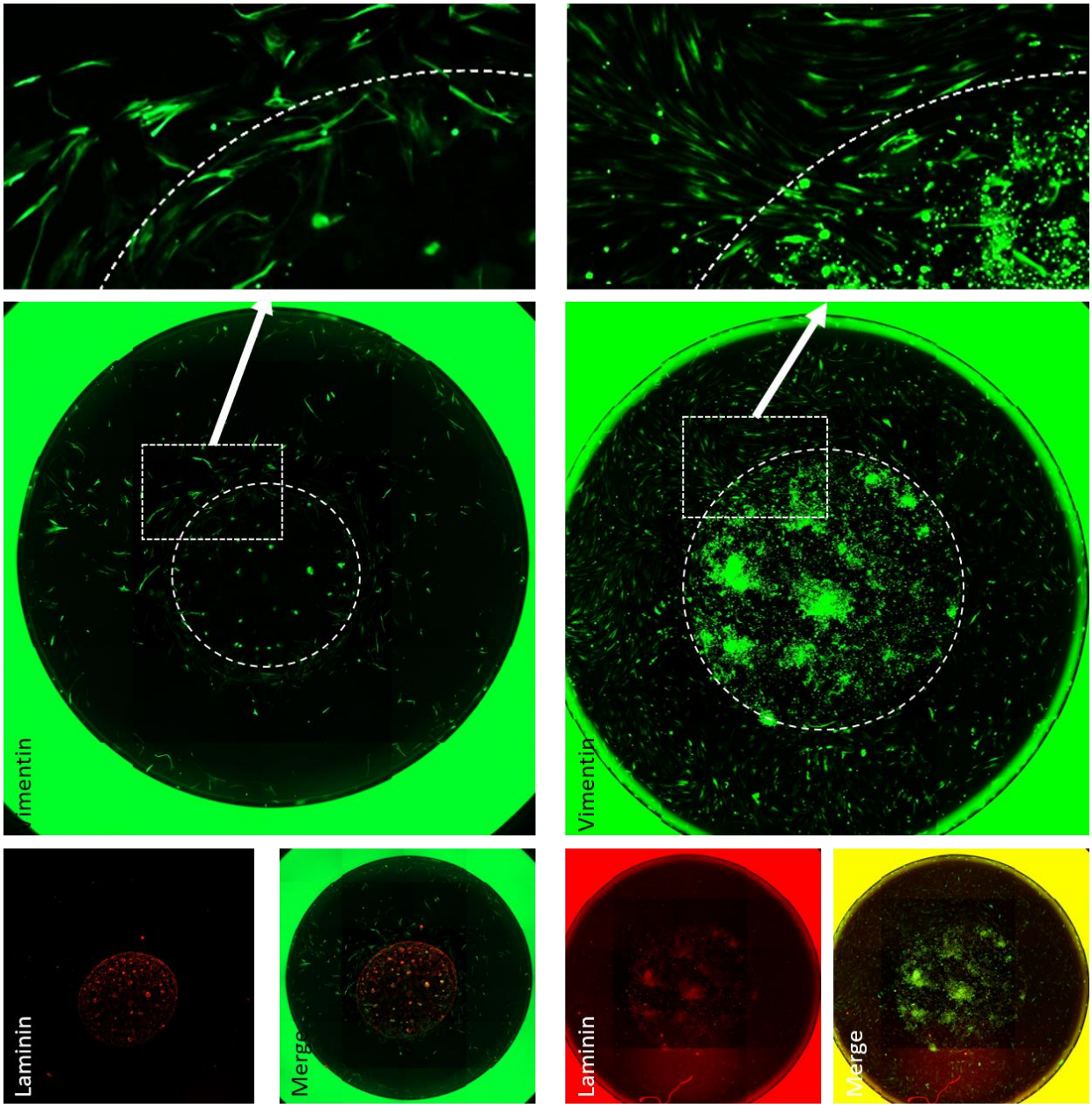
Figure 6.4 Recruitment Assay of CRC Cell Lines Cocultured with Other Myofibroblast Cell Lines Reconfirms Previous Observations: IF staining of recruitment assay with myofibroblast-recruiting cell lines including SW48 and OXCO1 as well as myofibroblast-non-recruiting cell lines including LS180 and SW620. All four CRC cell lines are cocultured with myofibroblast cell lines including CCD18Co (control), Myo0164c, and Myo1998. Green = vimentin (a biomarker for myofibroblast) and red = laminin (marker to delineate the Matrigel blob).



Recruiting cell lines

SW48 with Myo1998

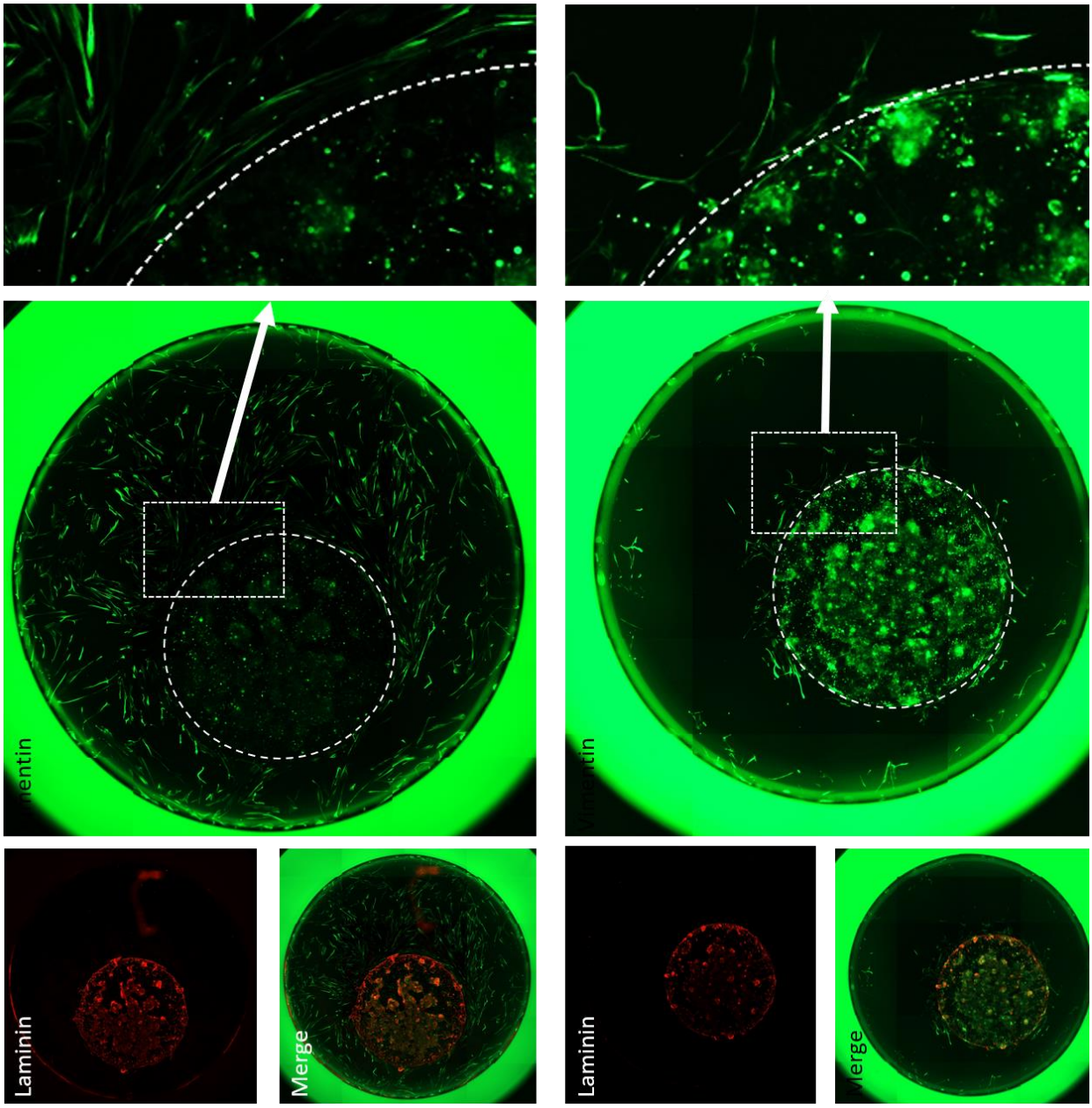
OXCO1 with CCD18Co



Recruiting cell lines

OXCO1 with Myo0164c

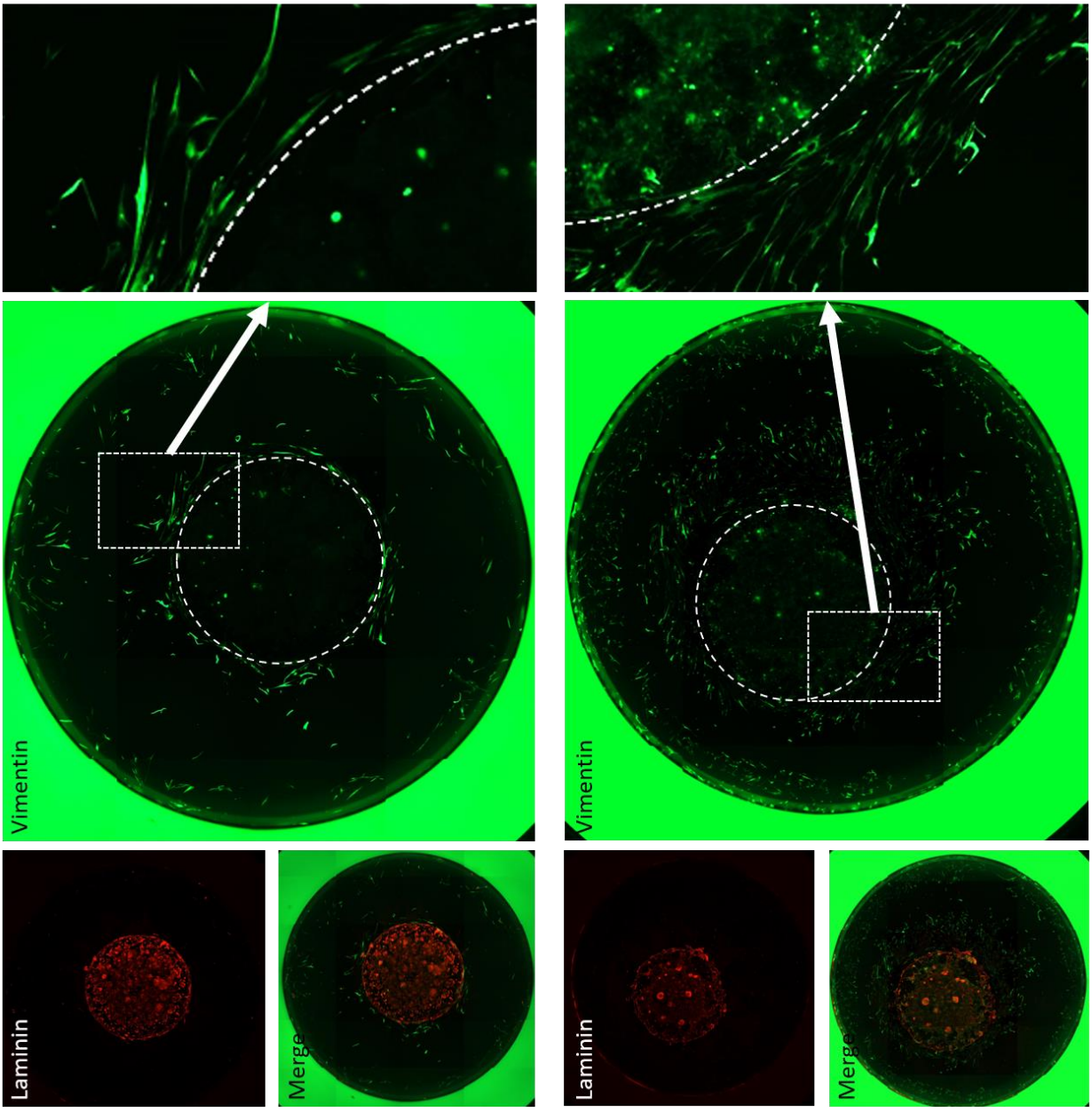
OXCO1 with Myo1998



Non-recruiting cell lines

LS180 with Myo1998

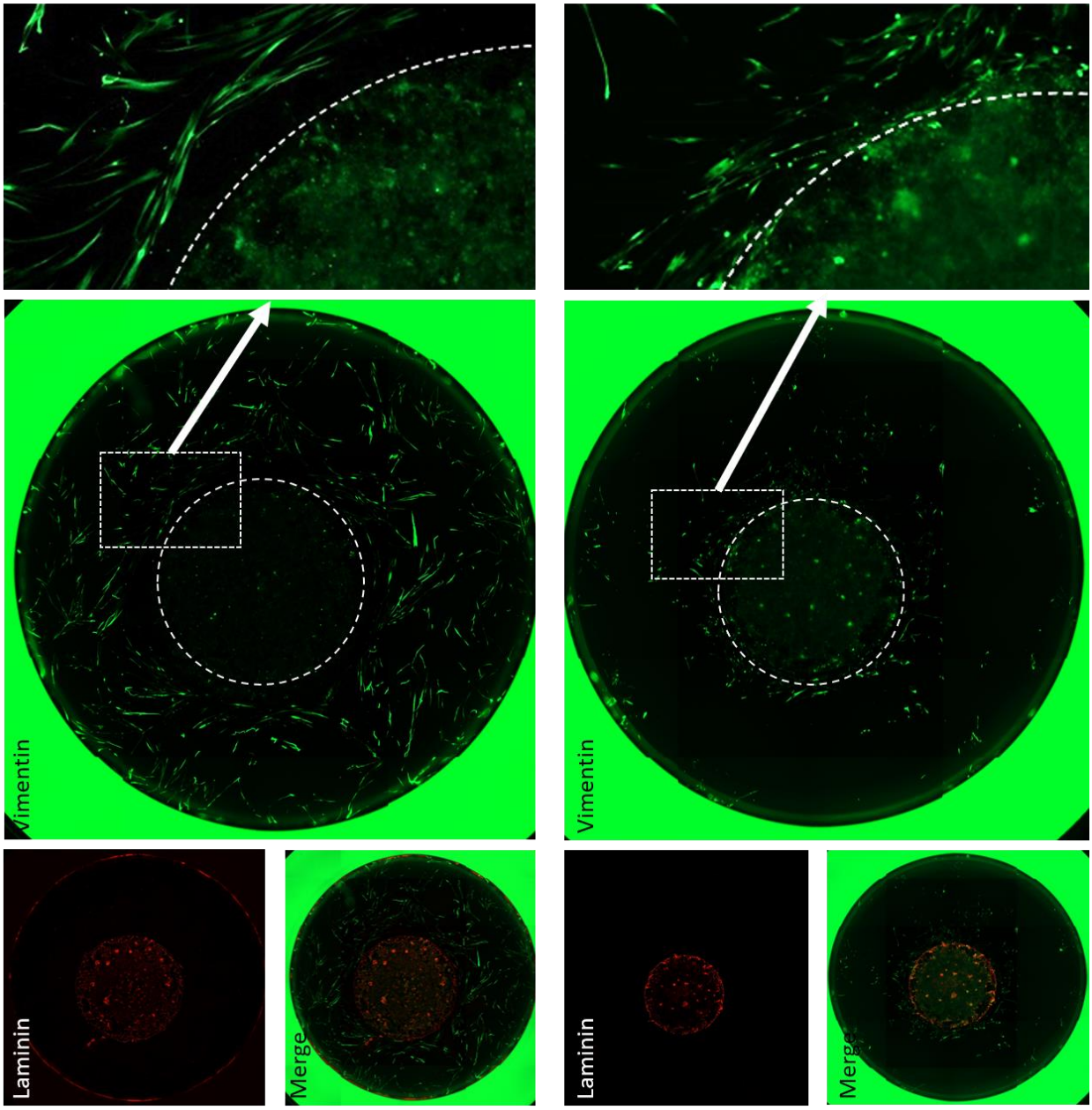
SW620 with CCD18Co



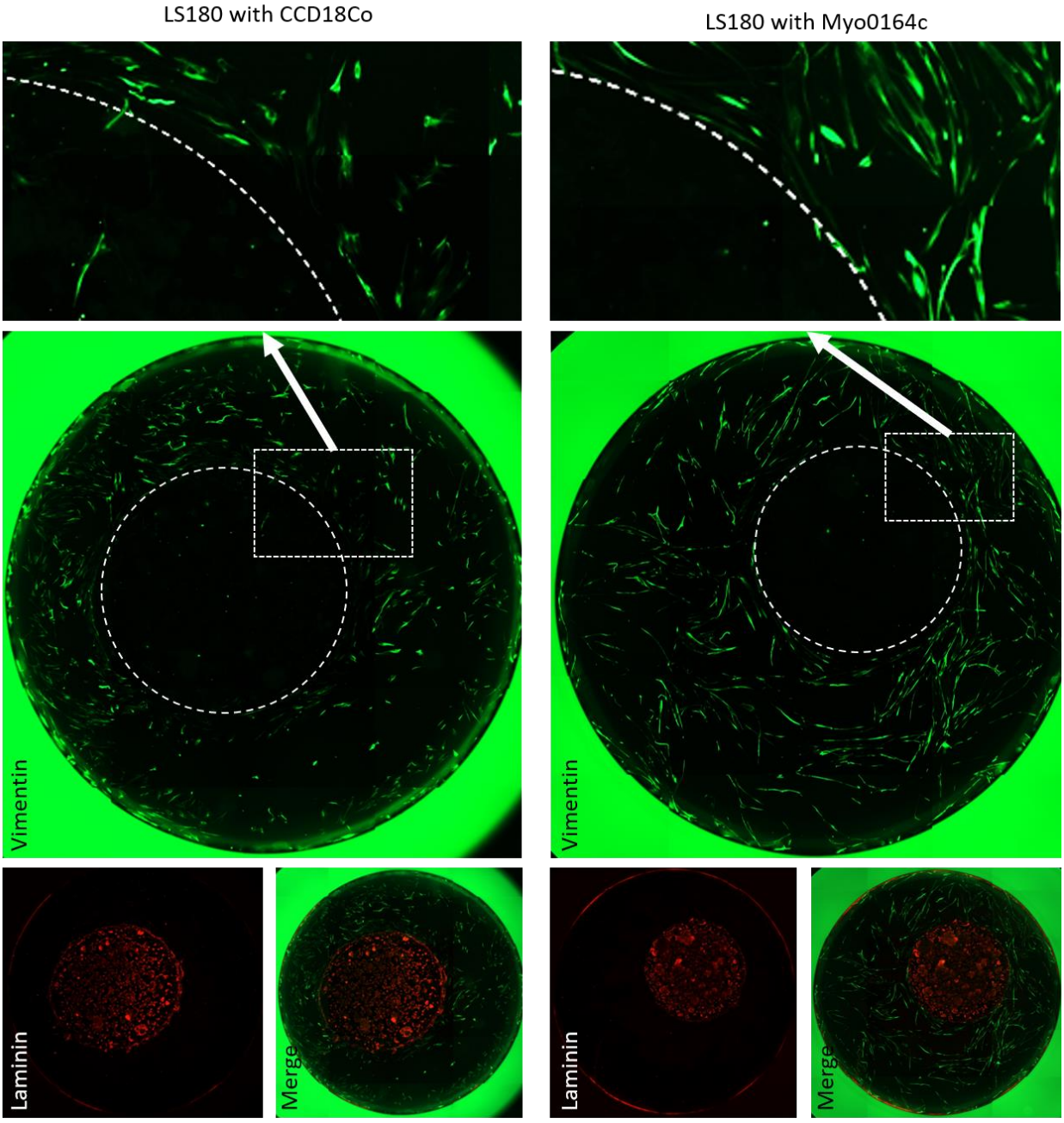
Non-recruiting cell lines

SW620 with Myo0164c

SW620 with Myo1998



Non-recruiting cell lines



Another experiment we carried out was to look at the effect of ZEB1-KO on OXCO1's ability to recruit myofibroblasts. The ZEB1-KO OXCO1 was established and discussed in Chapter 5. Surprisingly, we found that ZEB1-KO in OXCO1 led to a loss in the myofibroblast recruitment phenotype in wildtype OXCO1. As seen in Figure 6.5, none of the myofibroblast cell lines invaded into the Matrigel blob of ZEB1-KO OXCO1. This prompts the possibility of whether ZEB1-KO in OXCO1 in any way affected the level of PDGFC and PDGFA expression. A sensible experiment to continue to explore this observation is a simple qRT-PCR or western blot to detect the level of PDGFC in ZEB1-KO OXCO1. However, due to the pandemic, I did not have time to carry out this experiment.

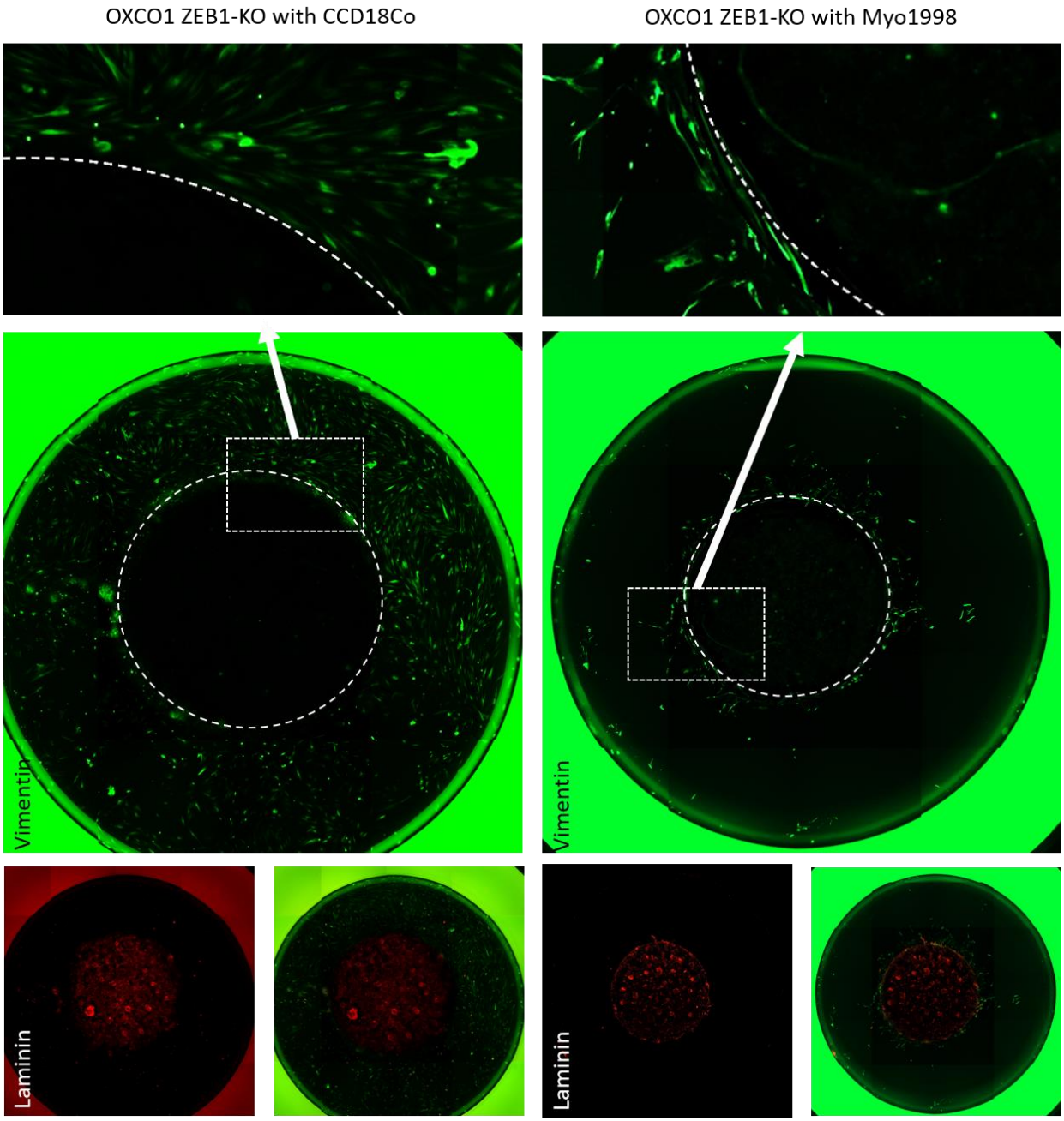


Figure 6.5 ZEB1-KO Blocks Myofibroblast Recruitment in OXCO1: IF staining of ZEB1-KO OXCO1 showing a loss of myofibroblast recruitment compared to OXCO1 wildtype shown in Figure 6.4.

6.2.3 Establishing PDGFC-KO clones

After confirming that previous observations on myofibroblast recruitment are repeated across different myofibroblast cell lines, the next step is to establish PDGFC-KO clones to confirm PDGFC as the main signaling molecule responsible for this phenotype. We have at this stage designed the gRNA targeting PDGFC, established the CRISPR vector, sequenced to confirm the inserted gRNA, transfected vector into 3 different myofibroblast-recruiting cell lines, and are currently waiting to confirm the knockout of PDGFC via PCR/gel electrophoresis and western blot. Unfortunately, due to the pandemic, we were not able to finish this experiment which will hopefully be carried out by another DPhil student in the future.

Chapter 7: Single Cell Sequencing Pipeline for Rare Genes

7.1 Introduction

In recent years, single-cell sequencing is quickly becoming an important technology powerful in identifying molecular profiles of individual cells from a bulk population of the input sample. Single-cell RNA-sequencing (scRNA-seq) enables researchers to begin understanding transcription profiles on a cellular level rather than on a bulk level. This further allows researchers to parse out the heterogeneity in a bulk sample, resulting in the understanding of the makeup and interaction between different cell types (Haque *et al.*, 2017; Tang *et al.*, 2019). The opportunities of single-cell sequencing in understanding cancer biology are huge. In bulk RNA-sequencing data, the expression levels of the variety of cell types in a cancer setting result in a single output consisting of averaged signals of each gene across the entire sample. Therefore, in a bulk RNA-seq data, researchers assume the output data represents a homogenous population of cells. Yet, through techniques such as immunofluorescence or FISH, it is evident that in a tumor microenvironment or tumor tissue sample, there exists a heterogeneous population of cells, each expressing varying levels of any given gene. Thus, scRNA-seq becomes a powerful tool in achieving a better understanding of different cell types in a cancer setting.

Nonetheless, due to current limitations in cDNA library preparation and machine sensitivity, scRNA-seq data is noisier and much more complex to post-process compared to RNA-seq data (Ran *et al.*, 2019). Many methods have been developed to deal with different aspects of postprocessing scRNA including identifying doublets and empty cells, dealing with RNA soup, compensating for high dropout rates as well as selecting for high-quality barcodes used in the downstream analysis (Young and Behjati, 2018; Qiu, 2020; Slyper *et al.*, 2020). In practice, doublets or empty droplets can skew downstream analysis by over-representing or under-

representing transcription profile if not properly removed. Another important observation is of RNA soup, introduced by *Sam Behjati et al.*, where cell-free RNA can potentially confound the interpretation of the transcriptome of individual cells in scRNA-seq (Young and Behjati, 2018). Furthermore, Peng Qiu (Qiu, 2020) has shown that the dropout rate of scRNA-seq varies across different platforms but remains high with an average dropout rate of 89% on the SMART-seq2 platform and a 93% dropout rate on the droplet-based 10X Genomics platform. Thus, postprocessing of scRNA-seq is an important field of active research critical to extracting the maximum amount of useful data from a scRNA-seq output.

Most of the research today using scRNA-seq has been performed to understand diverse populations of cell types from a tissue sample or biopsy. Inherently, the more diverse the cell population within a sample, the less complex and tedious it is to discern discrete cell types. This naturally makes tissue sampling a strong candidate to approach due to the inherent heterogeneity. The result is usually mapped onto a dimensionally reduced plot using t-distributed stochastic neighbor embedding (TSNE) or uniform manifold approximation and projection (UMAP) to reveal distinct local clusters of different cell types. Thus, current pipelines are efficient in postprocessing, extracting, and generating a reasonably meaningful output. Nonetheless, very little research has been done on distinguishing mixed cell lines of the same tissue origin which is the focus of this chapter. This presents a very different challenge where the input sample represents a population of cells that are very similar in the majority of their expression profile. We believe by establishing a pipeline capable of distinguishing clusters from cell types of the same tissue origin, we can thereby improve the separation of scRNA-seq datasets studying cell types of similar origin i.e. different types of differentiated epithelial cells in the colon.

In this chapter, the goal is to establish and validate a GMMchi-based single-cell sequencing pipeline that improves the ability to retrieve high-quality barcodes (cells) and so results in a better separation among samples with mixed similar cell types. This method delves into a lesser discussed area of scRNA-seq that can be powerful in improving our understanding of cell line models and the heterogeneity in different tissue-derived cell lines and the implications for cell line-based experiments.

7.2 Results

7.2.1 GMMchi-based Single Cell Sequencing Pipelines

In this chapter, barcode(s) is the equivalent of single cell(s) and Unique Molecular Identifiers or UMI(s) is the equivalent of a gene(s). In our lab, another DPhil student Gulnar Abdullayeva, whose research focus is on understanding goblet cell differentiation using CRC cell line models, conducted a scRNA-seq experiment to understand cell lines present and absent of MUC2, the gene encoding goblet cell marker mucin 2. The input sample for the scRNA-seq consisted of a 50:50 mix of CRC cell lines LS180 and Colo678, representing a MUC2+ and a MUC2- cell line, respectively. Neither cell line was barcoded prior to the mix thus we cannot identify which cell line individual cells belonged to just by looking at the barcode. The returned output is a matrix of 50,000 possible barcodes and 25,345 UMIs, which represents the expression level of each gene in counts. Based on previous analysis of LS180 and Colo378 using our microarray analysis, we are aware that LS180 expressed MUC2 but not ALPP, the gene encoding placental alkaline phosphatase, while Colo378 expressed ALPP but not MUC2, giving us a set of biomarkers to discern whether the postprocessing pipeline worked efficiently in separating the two cell lines. Thus, we label these two genes in all clusters to discern the two cell lines.

Current 10X Genomics Pipeline and Disadvantages

We initially post-processed our scRNA-seq data using standard pipelines such as ***scprep*** provided by 10X Genomics and ***scanpy***, a python implementation for large-scale single-cell gene expression analysis developed by Fabian J. Theis *et al* (Krishnaswamy *et al.*, 2014; Wolf, Angerer and Theis, 2018). The documentation for scprep is found in the link: https://pypi.org/project/scprep/.

The documentation for scanpy is found in the link: https://scanpy.readthedocs.io/en/stable/.

In Figure 7.1, we show the results of processing our scRNA-seq data with the ***scprep*** developed by 10X Genomics. ***scprep*** which resulted in 3270 cells. Here, we only label cells with at least 1 count in either ALPP or MUC2 with orange and green.

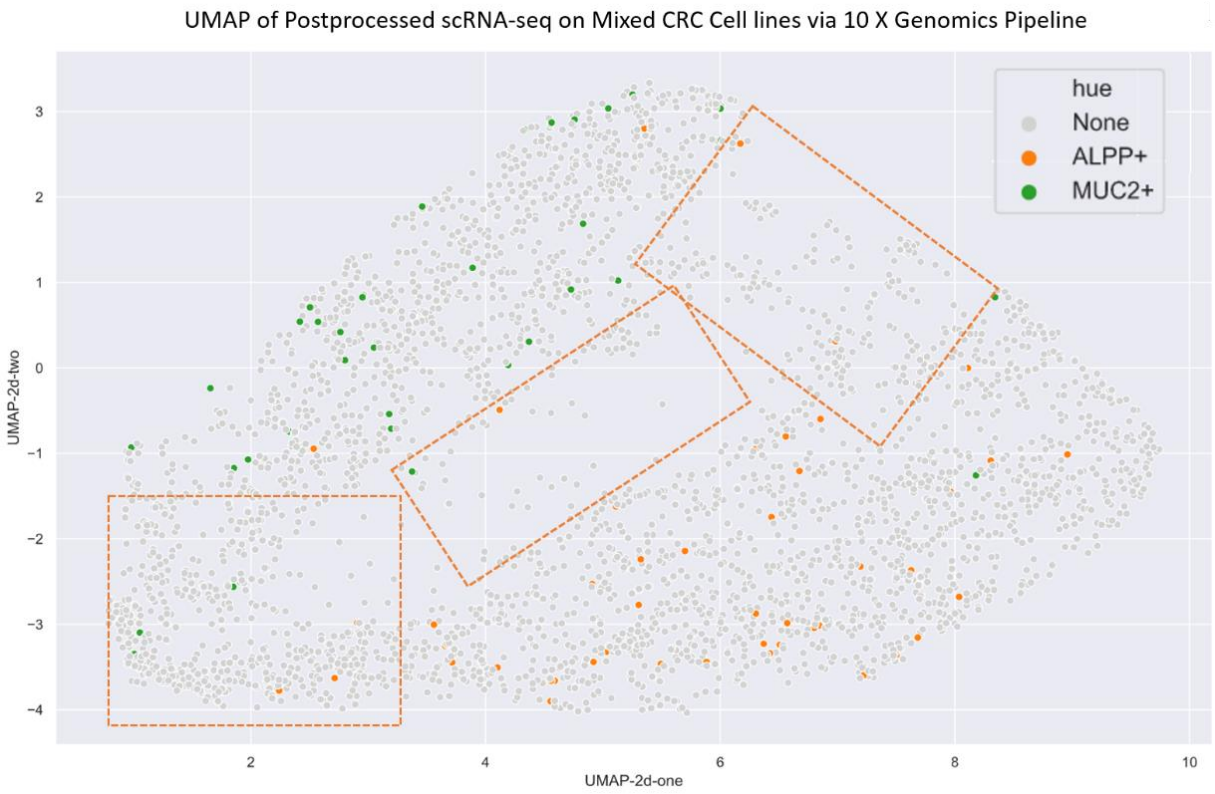


Figure 7.1 UMAP of scRNA-seq Data Postprocessed with scprep: Shows the UMAP of post-processed scRNA-seq data of mixing 50:50 CRC cell lines Colo678 and LS180 using the *scprep* pipeline from 10X Genomics. MUC2+ cells are denoted in green while ALPP+ cells are denoted in orange. MUC2 and ALPP are exclusively expressed in LS180 and Colo678, respectively, according to our cell line microarray data and immunofluorescence analysis. Here, we show that a significant proportion of the cells have poorly defined cluster memberships, as shown by the orange boxes, while the remaining cells form two major clusters as indicated broadly by the largely separate distributions of MUC2 and ALPP expressing cells.

The 10X Genomics pipeline ***scprep*** consists of four steps 1) filtering cells by library size (Figure 7.2a), 2) filtering out lowly expressed genes (Figure 7.2b), 3) library size normalization,

and 4) removing cells with high mitochondrial gene expression (Figure 7.2c). We applied the ***scprep*** pipeline onto our mixed cell line scRNA-seq data and showed the result of each postprocessing step in Figure 7.2. The resulting post-processed data is visualized using the dimensionality-reduction technique UMAP (Figure 7.1).

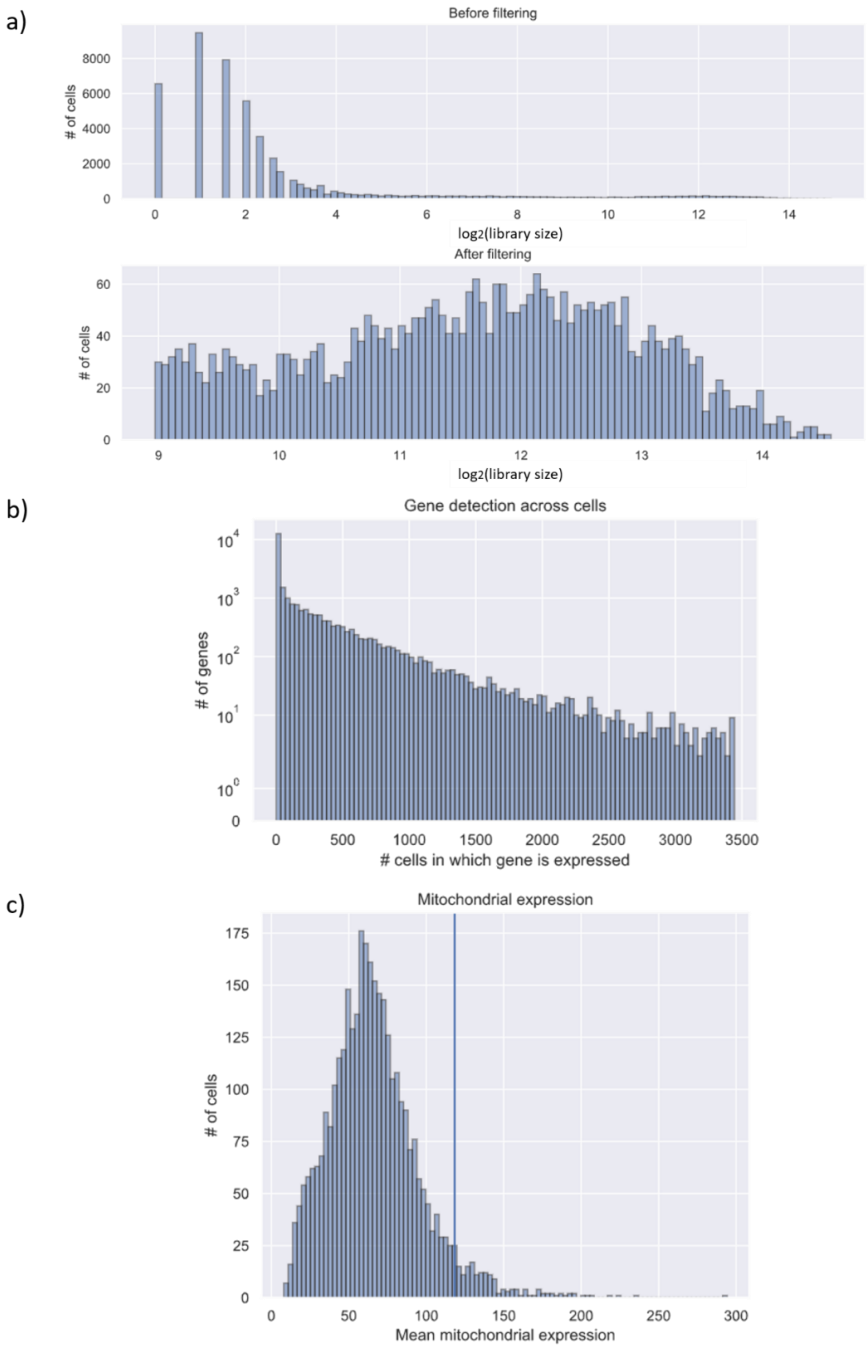


Figure 7.2 Outputs of each Step in Postprocessing scRNA-seq Data with scprep: (a), (b), and (c) are snapshots of steps involved in postprocessing scRNA-seq data using the library *scprep* by 10X Genomics. (a) Graph showing the library size of each cell before and after filtering library size > 25,000 and < 500. (b) shows the total number of counts a gene is expressed across all cells and removing genes with less than 3 counts. (c) shows the mean mitochondrial expression of each barcode/cell after the normalization step. Cells that are found beyond the top 95th percentile (blue line) are removed.

In Figure 7.1, we recognize that ***scprep*** did perform reasonably by identifying the two cell lines delineated roughly by the expression of MUC2 and ALPP colored respectively in green and orange. Nonetheless, we observe a substantial proportion of cells that has vague cluster membership to either of the two major clusters. Based on current practice with most applications of scRNA-seq done on tissue samples which consist of a diverse makeup of cell types that expresses very different sets of gene markers, it is not surprising that the standard pipeline results in a poorer differentiation between the two epithelial CRC cell lines which are very similar in many aspects if compared to, for instance, lymphocytes vs epithelial cells.

Furthermore, when we were examining published literature on single-cell sequencing data, we found that some of the gene markers of identified clusters were oddly enough housekeeping genes or highly expressed genes such as those coding for mitochondrial or ribosomal proteins. We suspect that this is due to the variance of housekeeping genes, such as GAPDH, being much larger compared to rare genes, such as MUC2. Visualization techniques such as TSNE or UMAP can thus fail to characterize and cluster together cells based on a more general representation of all genes, but instead, cluster biased towards highly expressed genes that results in a gradient over highly expressed genes. We thus went back to our data and mapped the expression level of highly expressed genes including GAPDH, RPL5, RPS6, MT-CO2…etc... These are genes that either encode known housekeeping genes, mitochondrial genes, or ribosomal proteins. In Figure 7.3, we show that post-processed scRNA-seq data can be heavily biased towards highly expressed genes. Most of the time we observe a gradient. In other words, cells with low expressing genes associated in the same cells with highly expressed housekeeping and mitochondrial genes will be clustered together dominantly by the expression of the highly

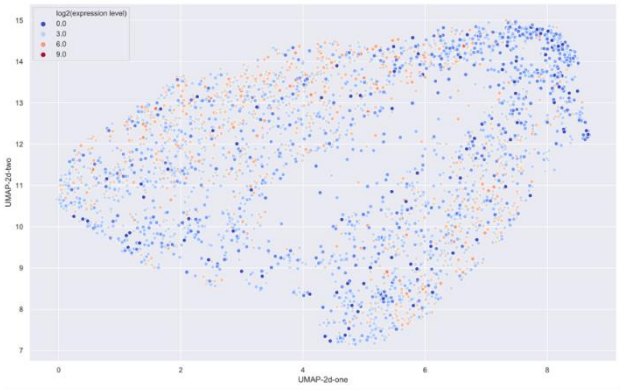
expressing housekeeping and mitochondrial genes. This minimizes the contributions of the lower expressing genes, notably, for example, transcription factors, which play a much more important role in the specification of cell types than the highly expressing house-keeping genes. Ideally, these highly expressed genes should be uniformly expressed in all cells, so they do not affect the result of visualization or clustering. In Figure 7.4, we further show the striking correlation between all these highly expressed genes. In the same figure, transcription factors (i.e. CDX1) and gene markers such as MUC2, LYZ, ALPP, and VIM revealed little to no correlation with the set of highly expressed genes delineated by a dashed black box. This strongly suggests that these highly expressed genes may present as a set of highly confounding factors during visualization and clustering. Ultimately, this may lead to the identification of clusters where the most differentially expressed genes are highly expressed genes including housekeeping genes, mitochondrial, and ribosomal proteins.

Based on these two very important observations, our goal is to improve the current pipeline to achieve:

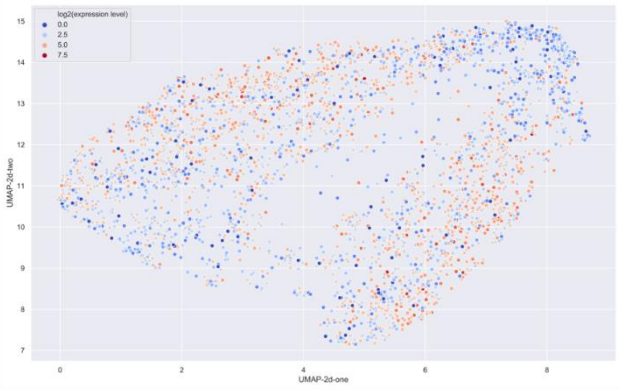
1) Better separation of cell types of similar origin.
2) Removal of the confounding factor due to highly expressed genes (housekeeping genes, mitochondrial genes, and ribosomal protein…etc.) in visualization and clustering.

Figure 7.3 Gradient in Highly-expressed Genes Found in Postprocessed scRNA-seq Data with scprep: UMAP graph of our scRNA-seq data post-processed with *scprep*. Here, we show the distributions of $\log_2$(expression values) of housekeeping genes, mitochondrial (MT) genes, and genes encoding ribosome proteins in the figure. All three categories of genes are considered highly expressed genes that can bias the visualization and clustering of scRNA-seq data. Genes such as GAPDH, ACTG1, ACTB, MALAT1, PABPC1, HSP90AB1, and RPL5 exhibit a relatively uniform expression inter- and intra-cluster. Nonetheless, we see a significant proportion of cells that expresses near-zero expression (blue cells) which disagrees with studies (Roy, McElhaney and Verschoor, 2020) showing that housekeeping genes are stably expressed at relatively high levels in all cell types. On the other hand, we observe a reciprocal cell line difference involving RPS6 and HSP90AA1 as well as a high level of MT-gene expression in the two major clusters and a near-zero expression of MT-genes at the tail that connects the two clusters.

GAPDH



ACTG1



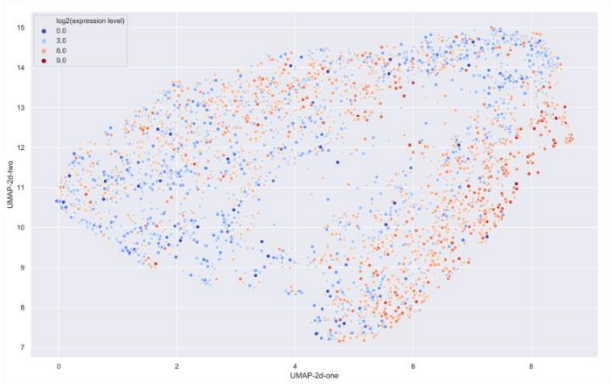
RPL5



HSP90AA1



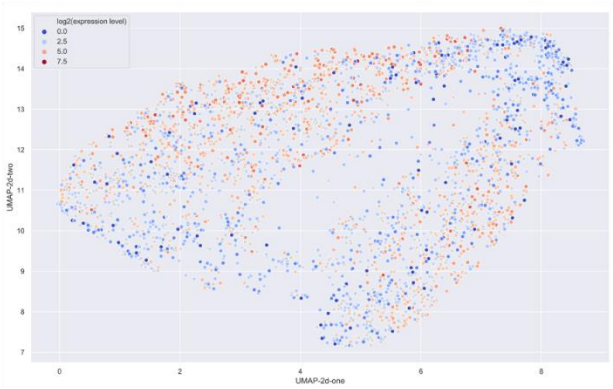
MALAT1



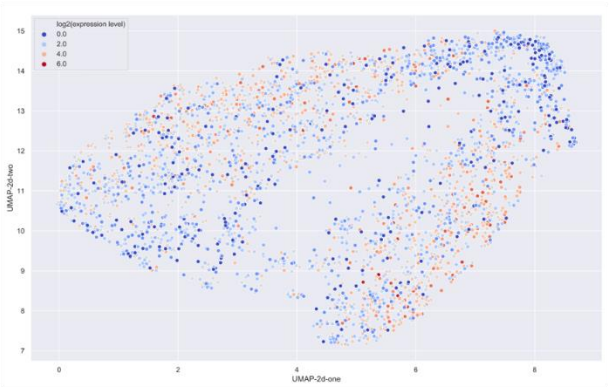
PABPC1



HSP90AB1



ACTB



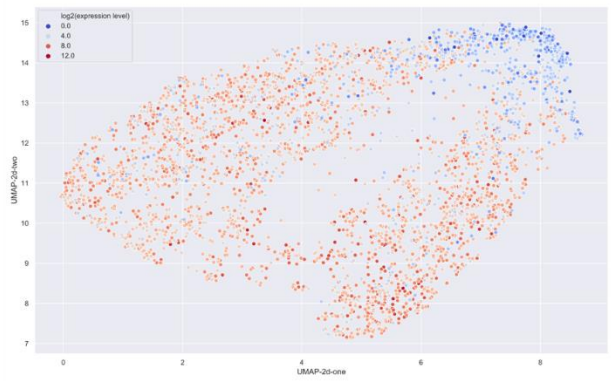
RPS6



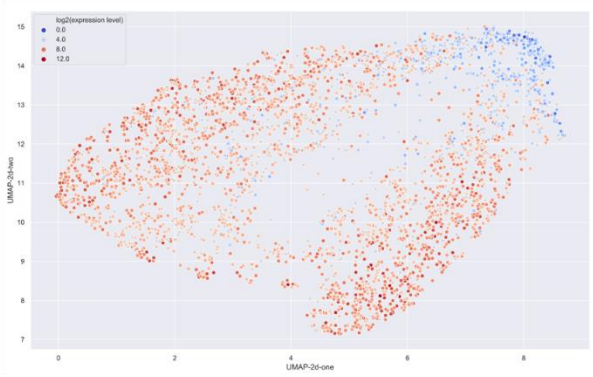
MT-ND4



MT-CO3



MT-ATP6



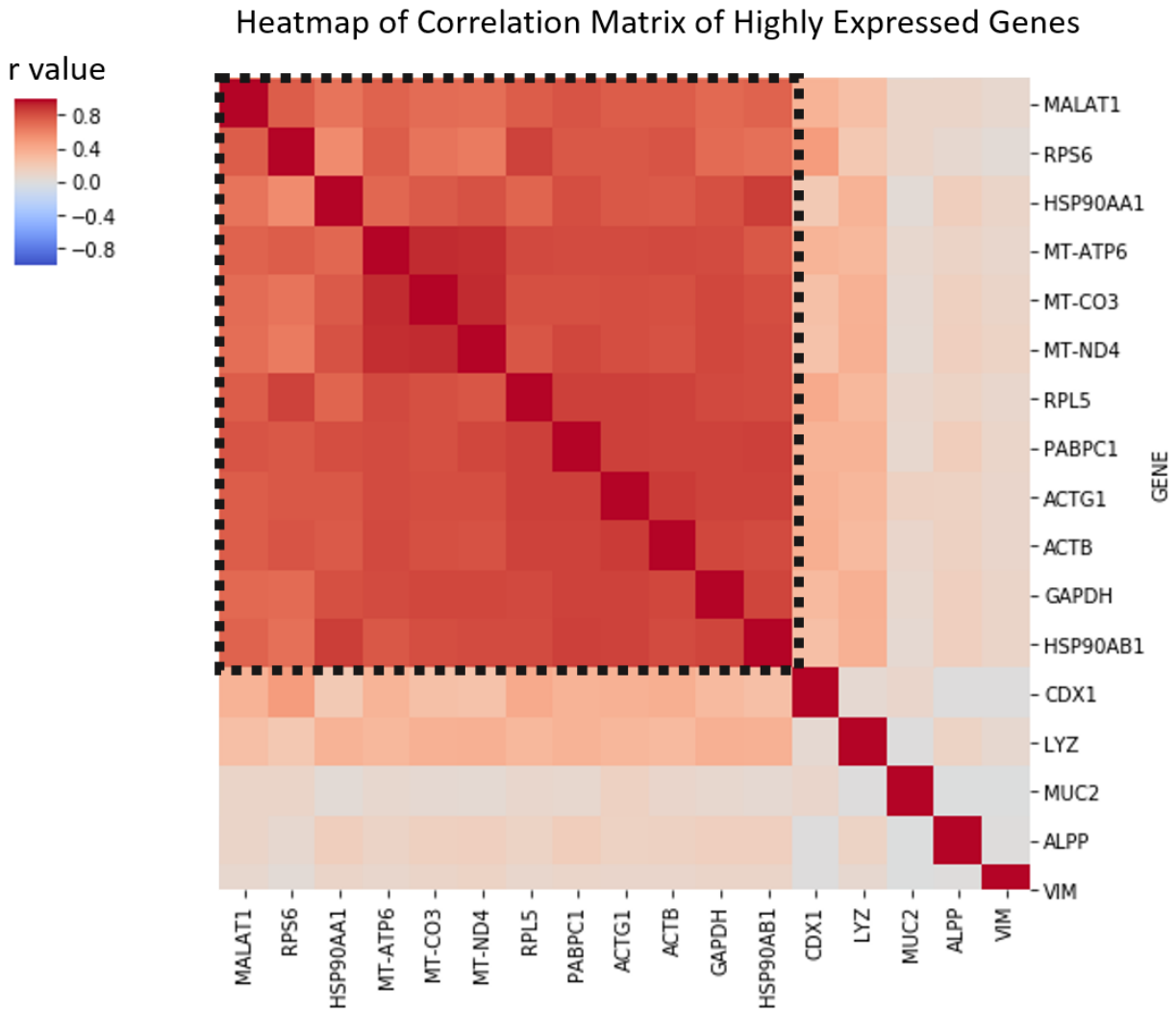


Figure 7.4 Correlation Matrix of Highly Expressed Genes: Heatmap of the correlation matrix done on a set of genes to show that highly expressed genes are highly correlated in expression pattern across individual cells, delineated by the dashed black box. Transcription factors such as CDX1 and genes such as MUC2, ALPP, LYZ, and VIM have little or no correlation to these highly expressed genes. This suggests that highly expressed genes including housekeeping genes, mitochondrial genes, and genes encoding for ribosomal proteins are highly confounding variables that can greatly influence the result of clustering if not properly dealt with.

Establishing GMMchi-based Single Cell Sequencing Pipelines

Before establishing a pipeline that attempts to deal with separating very similar cell types, we established a list of prior assumptions*:*

1) Barcodes (i.e. cells) consisting of a similar transcriptomic profile blur the difference between the two cell lines.
2) Poor differentiation between similar cell types can be due to low-quality barcodes representing RNA soup or cells with relatively high dropout that result in a poor representation of the true biological transcriptomic profile.
3) Housekeeping genes are genes that are stably expressed across different cell types with expression levels that fall under a single normal distribution are the best choice for assessing barcode quality.
4) Low-quality barcodes express distinctly low-level expression of certain housekeeping genes that form a bimodal distribution with high-quality barcodes that express normal levels of housekeeping genes. This may explain the reason behind the failure to capture the overall representation of the transcriptome when clustering such cells leading to a bias towards clustering based on highly expressed genes.
5) Importance of removing barcodes with low library complexity. This complexity quantifies how many unique UMIs an individual barcode expresses and removes low-level complexity that effectively removes RNA soup.

In the following, we discuss our approach towards dealing with scRNA-seq on the CRC cell line experiment and the result of postprocessing the scRNA-seq data with our pipeline. We established the library ***GMMchi_scs_pipeline*** based in Python that automatically runs through every step of our pipeline. The input data of ***GMMchi_scs_pipeline*** is a matrix with rows representing UMIs and columns representing barcodes. An example of a part of such a matrix inputted is shown in Table 7.1:

		Barcodes				
		CTCTCTAT	GCGTAAGA	TATCCTCT	AGATAGA	GTTAGGAG
UMI	ZEB1	0	0	0	1	1
	VIM	0	2	0	5	1
	CDX2	1	0	0	0	0
	OTOP2	3	4	0	0	1
	CDH1	12	9	0	3	5

Table 7.1 An Example of Input Data for the GMMchi_scs_pipeline: An example of the input data for *GMMchi_scs_pipeline* consists of rows representing UMIs and columns representing barcodes.

The pipeline consists of:

1) Removing doublet and empty barcodes.
2) Removing barcodes with low library complexity.
3) Remove low-quality barcodes based on the distribution of housekeeping genes (GAPDH, ACTB, and ACTG1).
4) Remove UMIs/genes with little or no expression.
5) Library normalization.
6) Remove barcodes with high mitochondrial gene expression.

The UMAP of the resulting post-processed scRNA-seq data will be shown in a figure for each step to visualize the improvement of the resulting cluster as we progress through the steps.

7.2.2 Overview of GMMchi_scs_pipeline

7.2.2.1 Remove Doublet and Empty Barcodes

The first step of ***GMMchi_scs_pipeline*** consists of removing doublets and empty barcodes. This step is identical to ***scprep***. In this step, we quantify the library size of each barcode by summing the total counts for each UMI. In Figure 7.5, we show the $\log_2$(library size) before filtering and after filtering as well as the UMAP of the post-processed scRNA-seq data after filtering based on the library size. The cutoff threshold for filtering doublets and empty barcodes is arbitrary and depends on each input sample data. Similar to ***scprep***, we chose to use > 25,000 UMIs/barcode as our filter for removing doublets and < 500 UMIs/barcode for removing empty barcodes. As seen in Figure 7.5a, there is a large number of barcodes that represent empty cells. After removing doublets and empty barcodes, the total available barcodes went from 50,000 to 3,443. Effectively filtering 93.1% of the possible 50,000 barcodes that we began with. This is a common phenomenon and so by no means, every available barcode is used for further analysis.

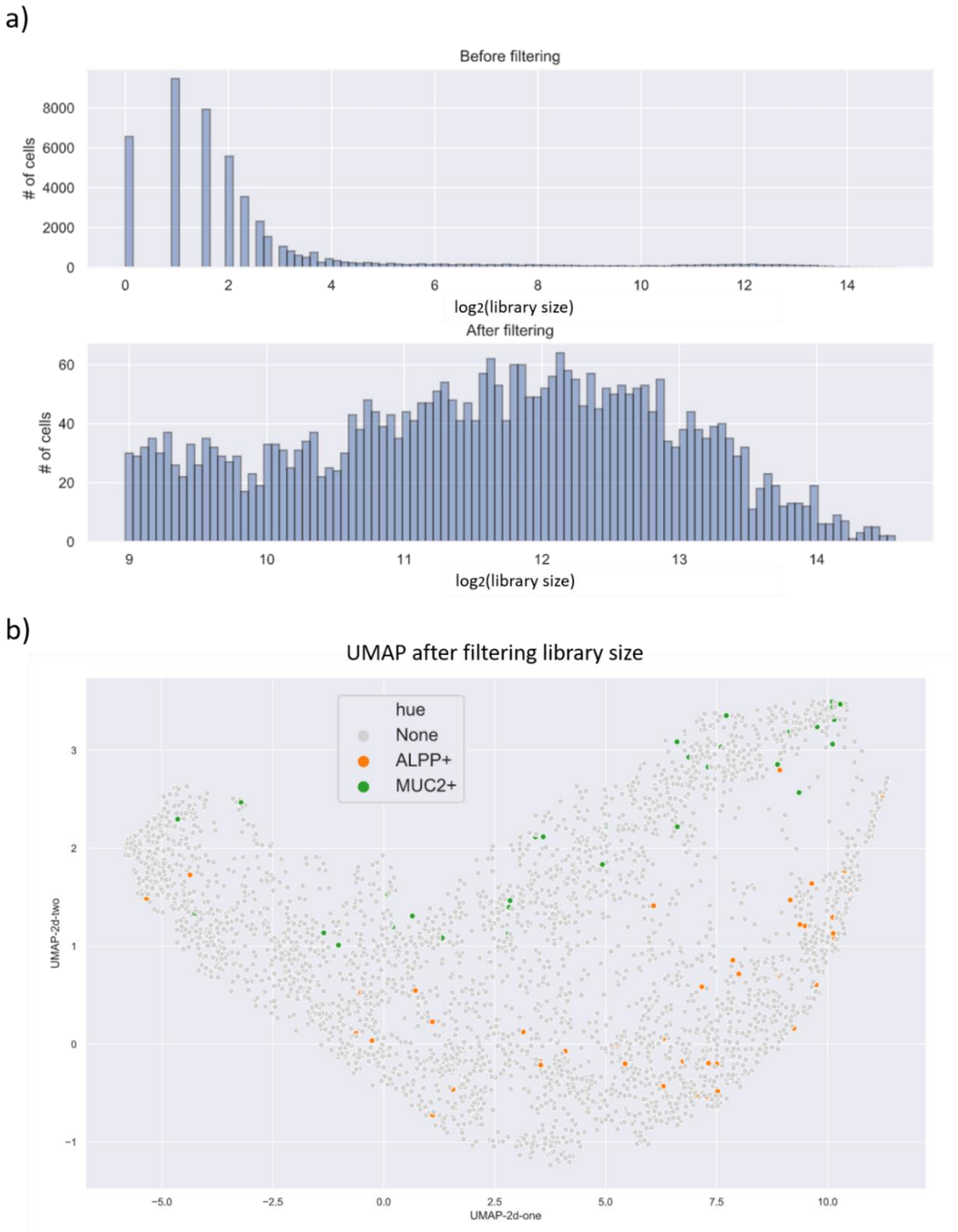


Figure 7.5 Library Size Filter Removes Doublets and Empty Droplets: (a) Graph showing the library size of each cell before and after filtering library size > 25,000 UMIs/barcode and < 500 UMIs/barcode. (b) UMAP of post-processed scRNA-seq data after filtering based on the library size. 3,443 barcodes left after this step.

7.2.2.2 Remove Barcodes with Low Library Complexity

The purpose of measuring library complexity is to measure how many UMIs are contributing to the total library size in a given barcode. For example, a cell with low library complexity can pass the library size filtering step with a total library size of 5,000 total counts; however, most of the counts can come from 30 housekeeping genes which the cell happens to express at a much higher level. Compare this to a cell with a high library complexity which also has a total library size of 5,000 total counts, yet the total counts are distributed across 500 different UMIs. We define the first case as low library complex barcodes and the second case as high library complex barcodes. In both cases, we are more interested in the latter case where a cell exhibits a higher library complexity. The definition of the library complexity for each cell is thus the sum of the UMIs that is measured with at least 1 count. Barcodes with low library complexity could potentially originate from the RNA soup or an incomplete cell due to cell death or debris. The concept of RNA soup was introduced by *Sam Behjati et al.* where cell-free RNA that originates from cellular apoptosis or free-floating RNA has the potential to contaminate and confound the correct biological interpretation of the transcriptomic profile of individual cells in scRNA-seq (Young and Behjati, 2018). The common characteristic of cell-free RNA is reflected in barcodes with low library complexity since they represent bits and pieces of RNA that can potentially be high in terms of expression level (i.e. housekeeping or mitochondrial genes) but low in terms of the variety of unique UMIs in a single barcode. Thus, we next remove barcodes with low library complexity to account for cell-free RNA. During this process shown in Figure 7.6, we measure the skewness of the distribution of the library complexity of all barcodes using the Fisher-Pearson coefficient of skewness and automatically determine the threshold cutoff by

defining the point where the skewness returns to 0. The skewed tail towards the left is assumed to represent barcodes with low library complexity. The threshold is shown as the orange dashed line in Figures 7.6a and 7.6b. After this step, we have reduced the barcodes from 3,443 to 2,731 or filtering out a further 20.7% of the barcodes left after the previous step. We see that in Figure 7.6c, there is relatively better separation between the two clusters compared to the UMAP in the previous step.

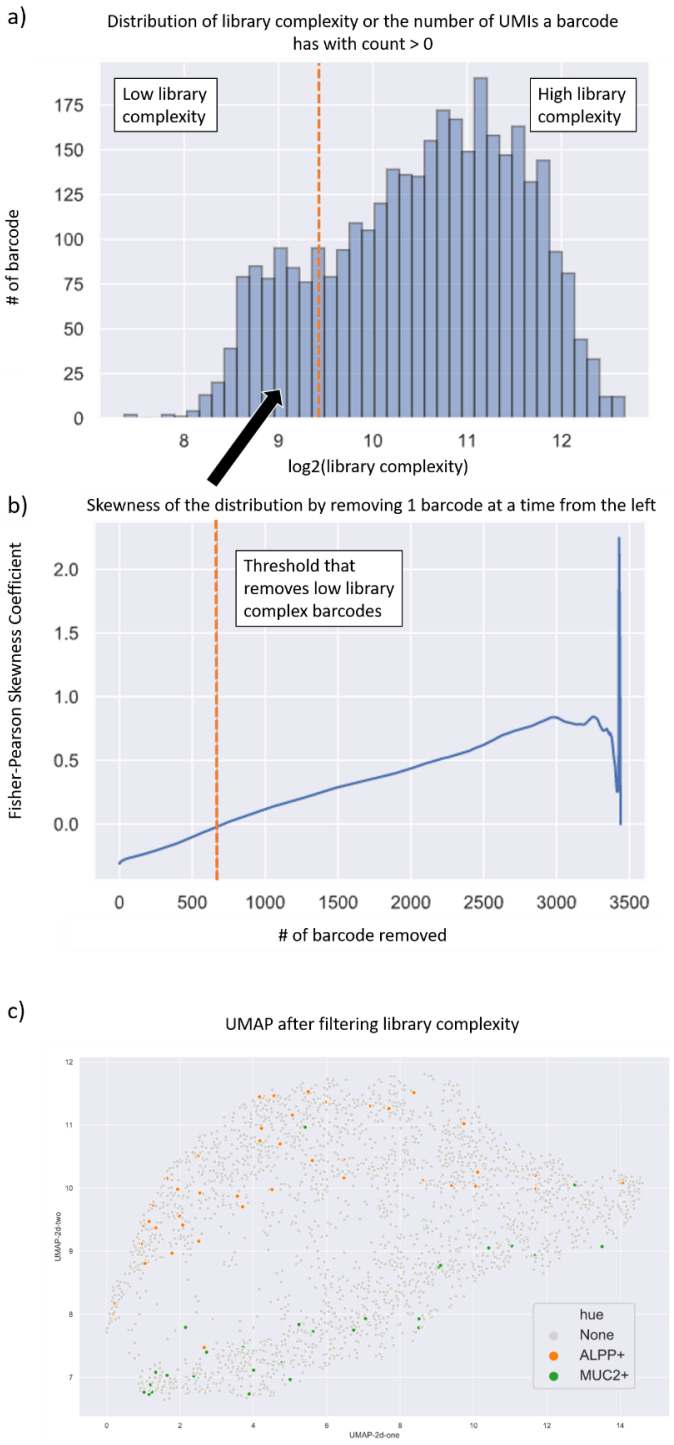
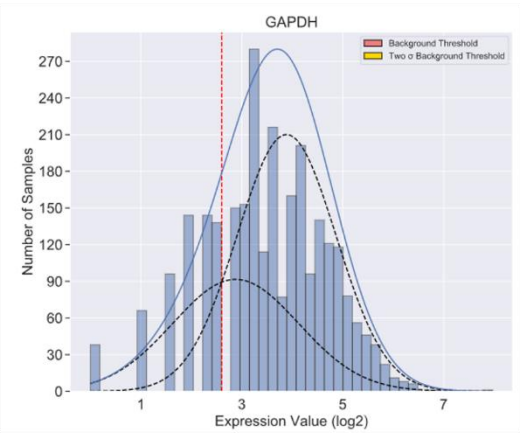


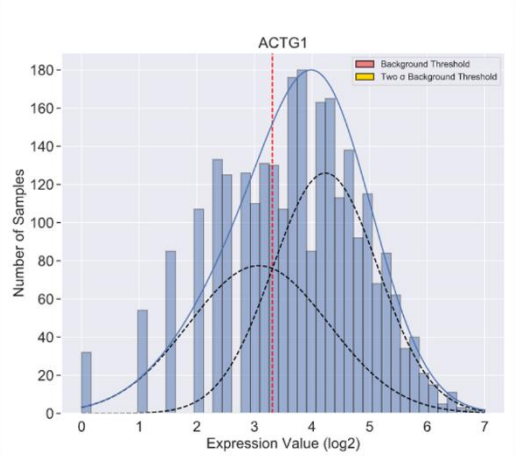
Figure 7.6 Library Complexity Filter Removes Low-quality Barcodes: (a) Graph shows the distribution of the library complexity which is the number of genes with count > 0 a cell has. (b) Graph showing the Fisher-Pearson skewness coefficient of the distribution in (a) when removing one cell at a time from the left until it reaches a skewness of 0, which is the threshold cutoff (orange line). (c) UMAP of post-processed scRNA-seq data after filtering based on the library complexity. 2,731 barcodes left after this step.

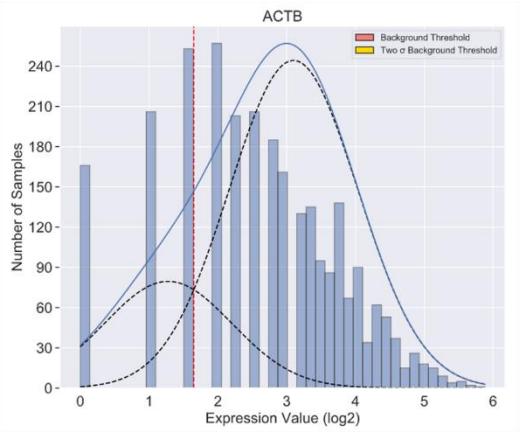
7.2.2.3 Remove Low-Quality Barcodes using Housekeeping Genes

In the next step, we remove low-quality barcodes by investigating the pattern of distribution exhibited by housekeeping genes. We use housekeeping genes including GAPDH, ACTG1, and ACTB, experimentally determined to be the most stably expressed housekeeping genes across all cell types (Noriega, Kohama and Urbanski, 2010; Caracausi *et al.*, 2017; Roy, McElhaney and Verschoor, 2020). We applied GMMchi to the housekeeping genes GAPDH, ACTB, and ACTG1 to identify whether the distribution exhibited a unimodal or bimodal distribution. The assumption is that all cells should normally express housekeeping genes in a range that falls under a single normal distribution due to the nature of housekeeping genes stably expressed across different cells. Thus, if a gene (UMI) exhibits a bimodal distribution or a unimodal distribution with a tail, we assume that the distribution with a lower range of expression represents low-quality barcodes consisting of abnormal, or even absent, level of housekeeping gene expression. The pipeline ***GMMchi_scs_pipeline*** iterates through the housekeeping genes (GAPDH, ACTG1, ACTB), fits with GMMchi, identifies the threshold cutoff of low-quality barcodes, and removes low-quality barcodes. In Figure 7.7a, the threshold delineates the threshold for cells that express distinctively low housekeeping genes. Here, we see that in the case of ACTG1 and GAPDH, there is a more obvious bimodal pattern with a high degree of overlap. However, in ACTB there is not a clear bimodal pattern. It is nevertheless clear that the spread and distribution in the lower end are noisy and we see a lot of cells that express little to no ACTB. In Figure 7.7b, the UMAP drastically improved and the separation between the clusters is now much more obvious. Post this step, 1487 out of 2731 barcodes were left after filtering, effectively removing a further 45.6% of the probes.

a)







b) UMAP after filtering using Housekeeping Genes

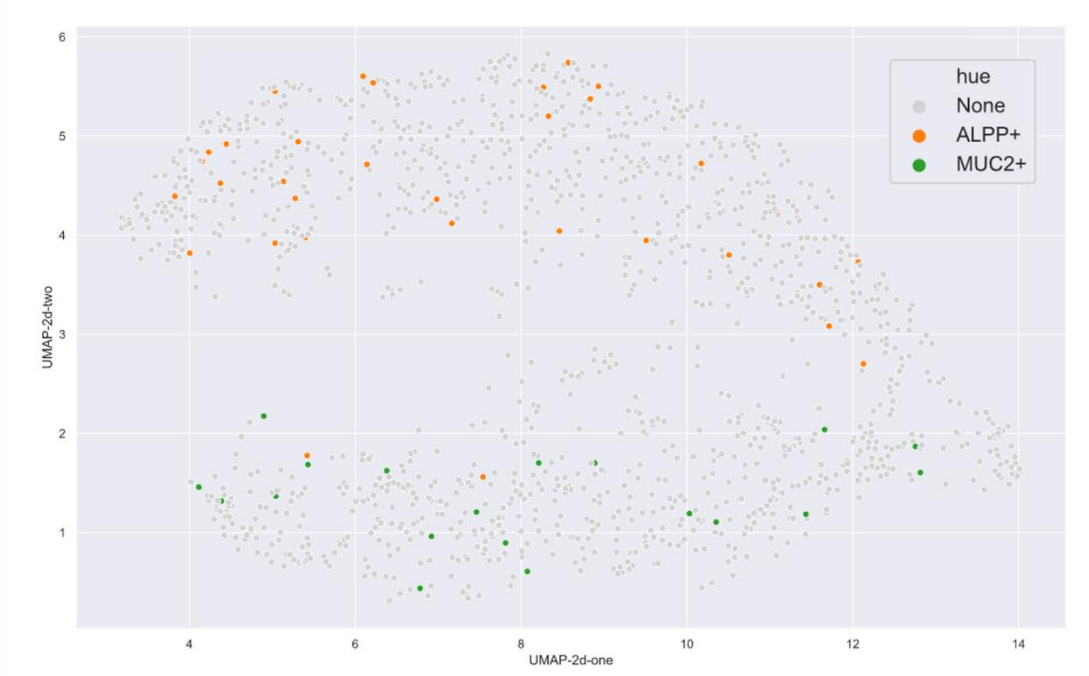


Figure 7.7 Housekeeping Gene Filter Results in a Better Seperation: (a) Graph shows GMMchi analysis on housekeeping genes including GAPDH, ACTG1, and ACTB. We show the distribution as well as the identified threshold using GMMchi for removing low-quality barcodes that express statistically low levels of GAPDH, ACTB, and ACTG1 located left to the red dashed line. The removal of the lower distribution is done separately for each of the three genes. (b) UMAP of post-processed scRNA-seq data after filtering based on the expression distribution of housekeeping genes (GAPDH, ACTB, and ACTG1). 1,487 barcodes left after this step.

7.2.2.4 Remove UMIs with little to no expression

The dropout rate is the percentage of genes that are not detected in a cell in which they are expressed. It has been widely suggested that a high dropout rate is mostly attributed to transcripts being lost or inefficiently reverse-transcribed during library preparation. Especially in single-cell sequencing where the target transcriptome is only extracted from a single cell, accurate reverse transcription of the transcriptome proves to be a technical challenge during library preparation. The dropout rate can thus be anywhere between 60-90% in scRNA-seq. As a result, the scRNA-seq output contains large numbers of zeros where many genes are not detected. To reduce computational load and to remove genes that have little to no expression, we remove any UMIs/genes with a total count of < 3 across all barcodes.

7.2.2.5 Normalization and Remove Barcodes with High Mitochondrial Gene Expression

In addition, we remove barcodes with high mitochondrial gene expression. The reason is that a high detection of mitochondrial RNAs usually suggests the degradation of the mitochondrial membrane as a result of cellular apoptosis. This can be attributed to stress introduced during sample handling. As shown in Figure 7.8, the right tail of the distribution represents barcodes with distinctly high mitochondrial gene expression. After removing the tail, we are left with 1417 out of 1487 barcodes, effectively removing 4.8% of the barcodes.

Library size normalization is an important step that normalizes the UMIs by the total size of the library of any given barcode. This is equivalent to comparing the ratio of the UMIs expressed within a cell rather than comparing the absolute count of each UMI. In other words, this step normalizes against variations that can originate from technical inconsistencies. We explored whether normalization improves the post-processed output in Figure 7.9. As shown, we see an improvement in the separation between the two clusters. Note that normalization is only used for visualization and identifying clusters but not in the actual analysis for any quantitative downstream analysis i.e. gene expression analysis. This is our final post-processed sRNA-seq output. Initially, we started with 50,000 barcodes and ended up with 1417 barcodes or 2.8% of the starting possible barcodes.

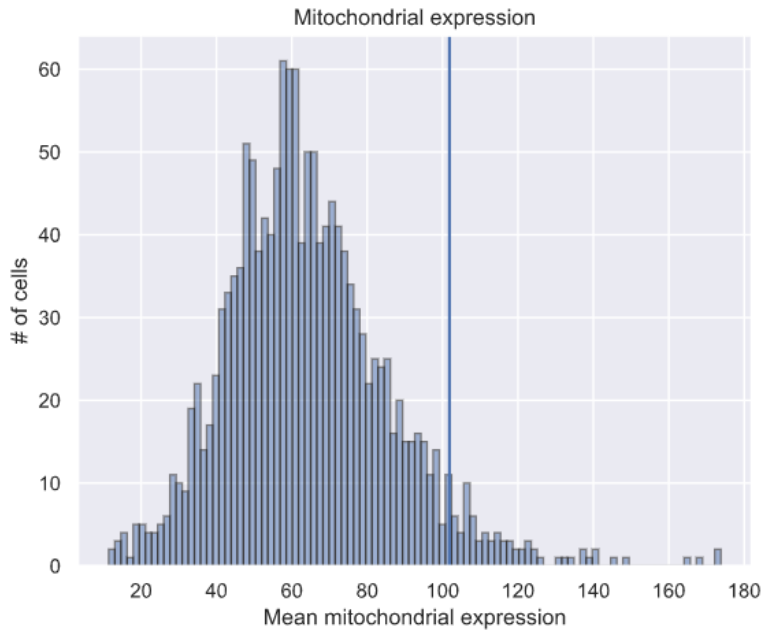


Figure 7.8 Removing Barcodes Expressing High Levels of Mitochondrial Genes Removes Apoptotic Cells: Graph shows the mean mitochondrial expression of each barcode/cell. We remove cells that occur beyond the top 95th percentile (blue line). 1,417 barcodes left after this step.

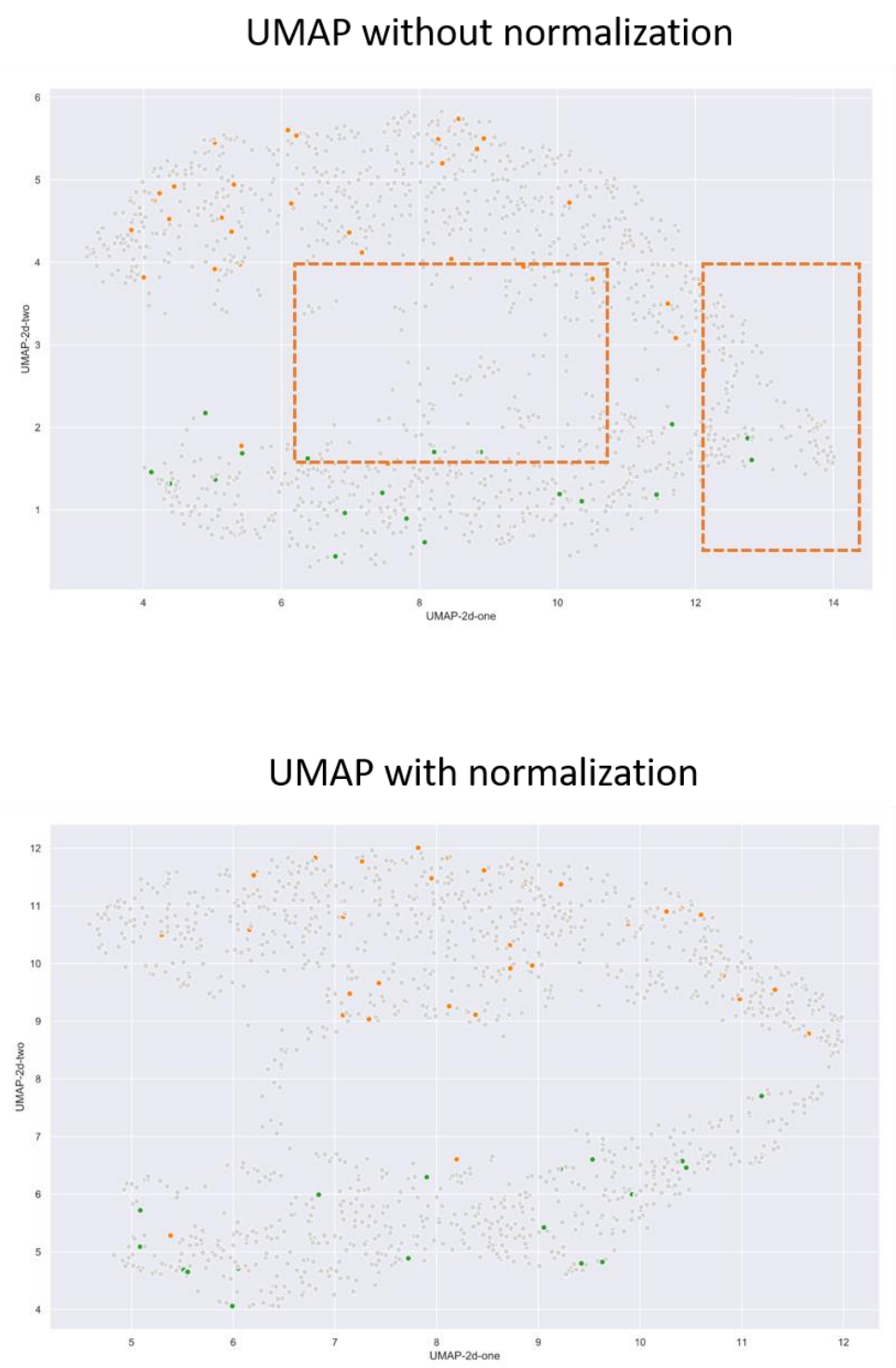


Figure 7.9 Normalization Results in a Better Separation: In the Figure, we compare the UMAP result of normalized vs non-normalized data. The orange box delineates cells with vague cluster membership. We show that normalized data has a much clearer separation between the two clusters.

7.2.2.6 Postprocessed Output of GMMchi-based Single Cell Sequencing Pipeline

In the ***GMMchi_scs_pipeline***, we built ***UMAP_scs_graph*** on top of the ***UMAP*** method from the ***sklearn*** library to include the capability to annotate UMIs before visualizing the post-processed sRNA-seq data with UMAP. UMAP is a dimension reduction technique that has been overtaking t-SNE recently in visualizing non-linear high dimensionality input samples such as scRNA-seq. The final result is shown in Figure 7.10. In the result from the ***GMMchi_scs_pipeline***, we resulted in 1417 barcodes while ***scprep*** resulted in 3270 barcodes (Figure 7.1). The bottom cluster in Figure 7.10 marked with MUC2+ cells as well as the top cluster marked with ALPP+ cells confirms that our pipeline has efficiently separated the two individual cell lines. Compared to the result from ***scprep***, there is nearly a 2-fold downsize in total resulting barcodes (from 3270 to 1417 barcodes) after postprocessing which resulted in a clear separation of the two cell lines. This shows that there is nearly more than 50% of cells may be of poor quality and affect downstream analysis when using the ***scprep*** pipeline.

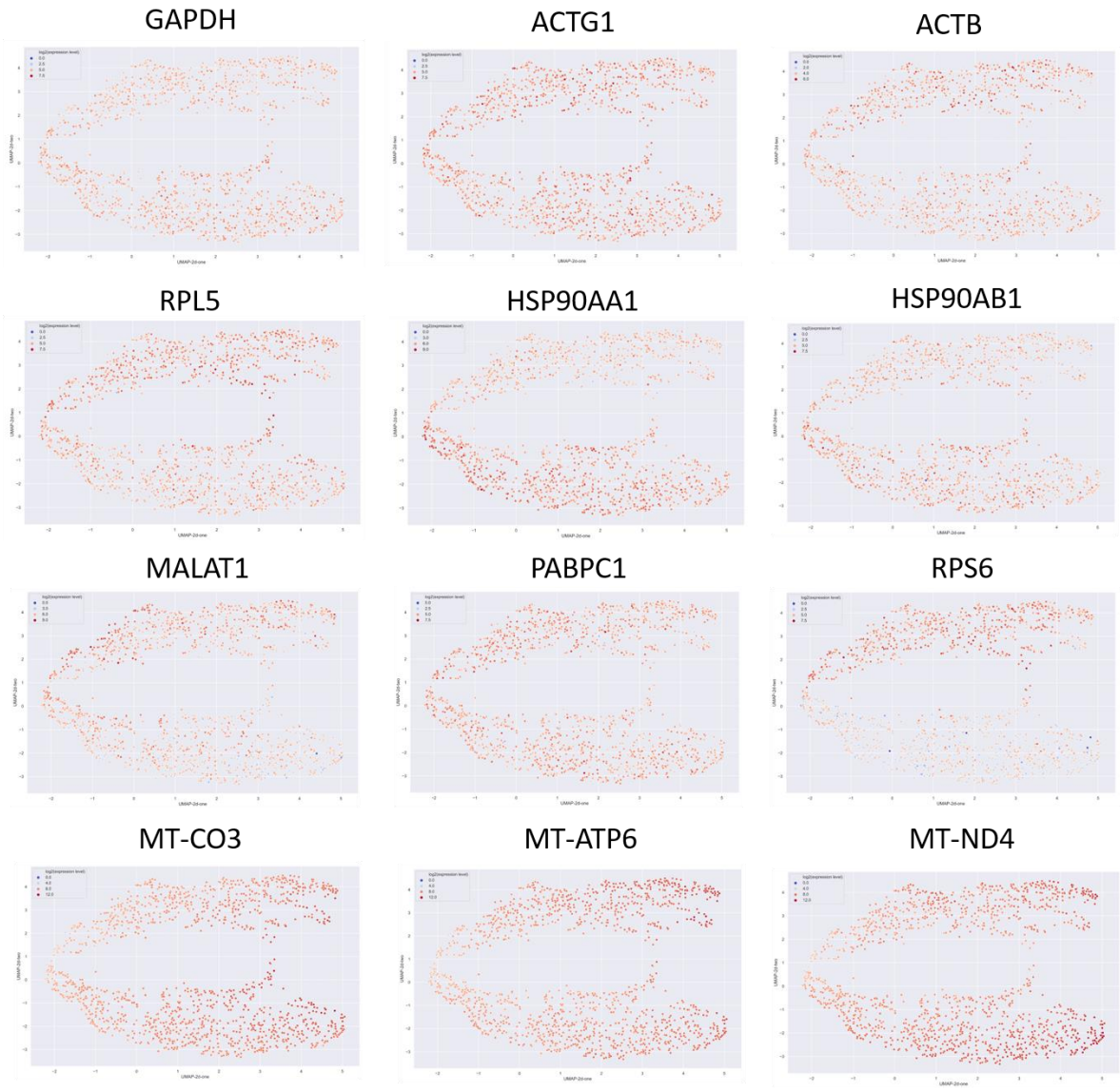


Figure 7.10 Final Output of GMMchi_scs_pipeline: The final output of post-processed scRNA-seq using the GMM-chi-based pipeline and mapping using UMAP. The ALPP+ and MUC2+ cells indicate the cluster membership for the Colo678 and LS180 clusters, respectively. The result from our pipeline outputs 1,417 barcodes.

In addition to achieving a better separation between the two cell lines, we show that in Figure 7.11a, the GMMchi-based pipeline successfully removed the gradient in highly expressed genes. We believe this is an important step in improving the identification and integrity of clusters that are based on markers that although lowly expressed, are, nevertheless, differentially expressed instead of being influenced by the large variance exhibited by highly expressed genes. There are two important observations when looking at the result in Figure 7.11a. First, our filtering step based on the distribution of housekeeping genes only relied on GAPDH, ACTB, and ACTG1. Despite relying on only 3 genes, we observe the removal of the gradient across all highly expressed genes. This proves the high inter-correlation between these genes and the danger of

not removing them as they tend to cluster together in the same cells. Moreover, this shows that using 3 housekeeping genes is sufficient for achieving our goal. Second, in the case of RPS6 and HSP90AA1, we noticed a difference in expression level between the two cell lines and further show the clear division in expression levels highlighted in our GMManalysis in Figure 7.11b. Nonetheless, we observe uniform expression within each cluster which shows that the gradient within each cluster is removed. This further suggests that the differential level of expression between the two cell lines despite filtering based on housekeeping genes is a real difference.

a)


GAPDH
ACTG1
ACTB
RPL5
HSP90AA1
HSP90AB1
MALAT1
PABPC1
RPS6
MT-CO3
MT-ATP6
MT-ND4

b)

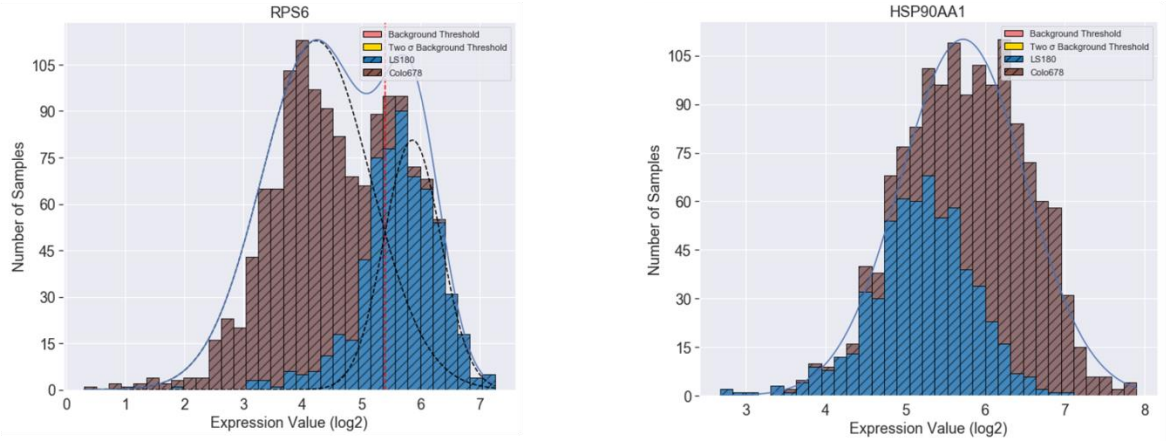

RPS6
HSP90AA1
Background Threshold
Two σ Background Threshold
LS180
Colo678
Number of Samples
Expression Value (log2)

Figure 7.11 GMMchi_scs_pipeline Effectively Removes Expression Gradients in Highly Expressed Genes: (a) UMAP showing the distribution of highly expressed genes (housekeeping genes, mitochondrial genes, and ribosomal protein). GMMchi-based scRNA postprocessing effectively removes the gradient in highly expressed genes. RPS6 and HSP90AA1 are exceptions where we see a distinctive difference in expression pattern between the two cell lines, suggesting that this may be a meaningful difference between the two cell lines rather than due to low-quality barcodes. (b) GMManalysis shows the clear division of the two cell lines in RPS6 and HSP90AA1 expression levels.

Lastly, we show that visualization or clustering was not influenced by other factors such as library size or library complexity, which also exhibits a relatively uniform distribution across the cells (Figure 7.12).

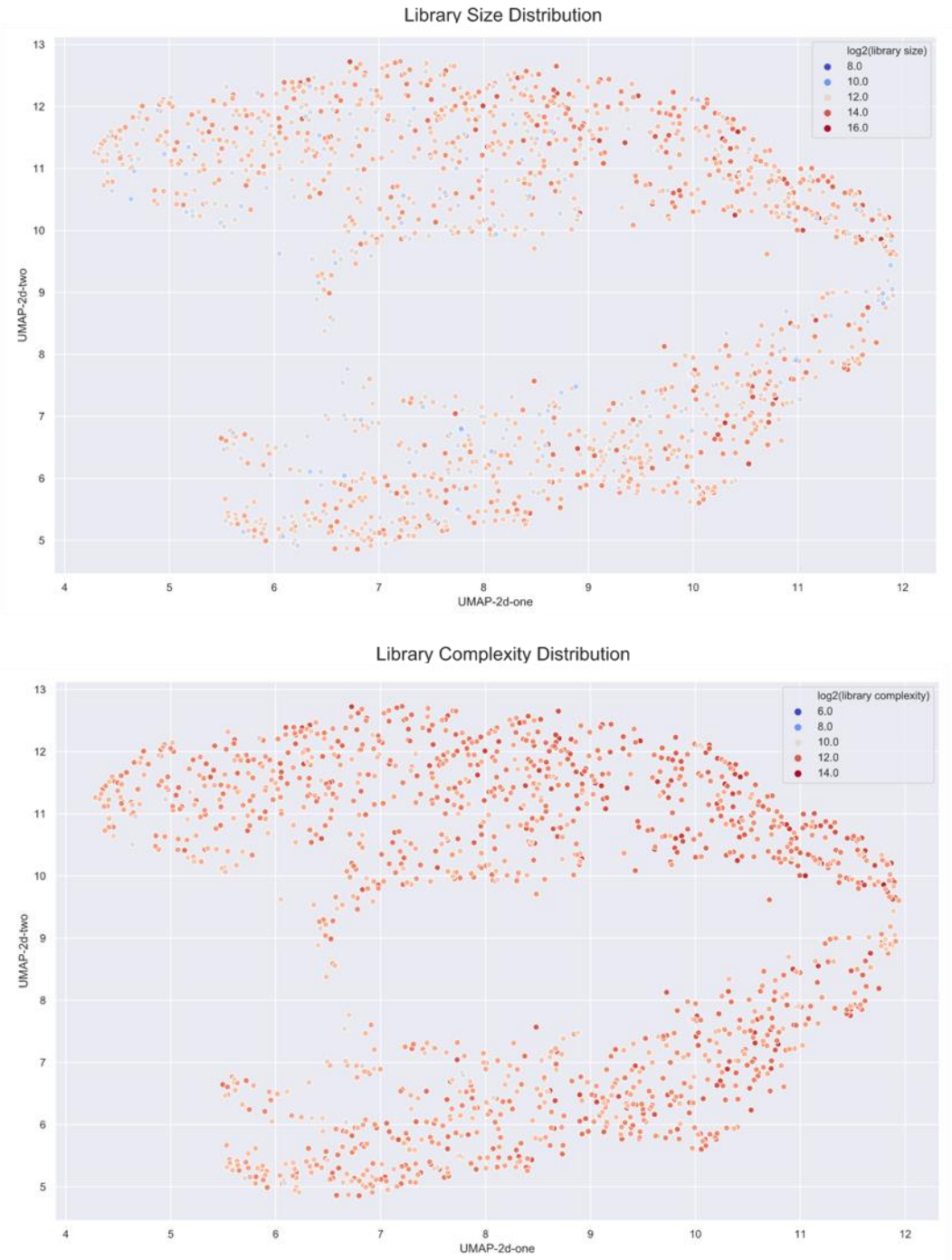


Figure 7.12 Library Complexity and Size are Evenly Distributed: Graph shows the UMAP of our post-processed scRNA-seq with library size and library complexity. The lack of a gradient shows that clusters were not influenced by library size or library complexity.

7.2.3 Dealing with Dropout Rates in scRNA-seq

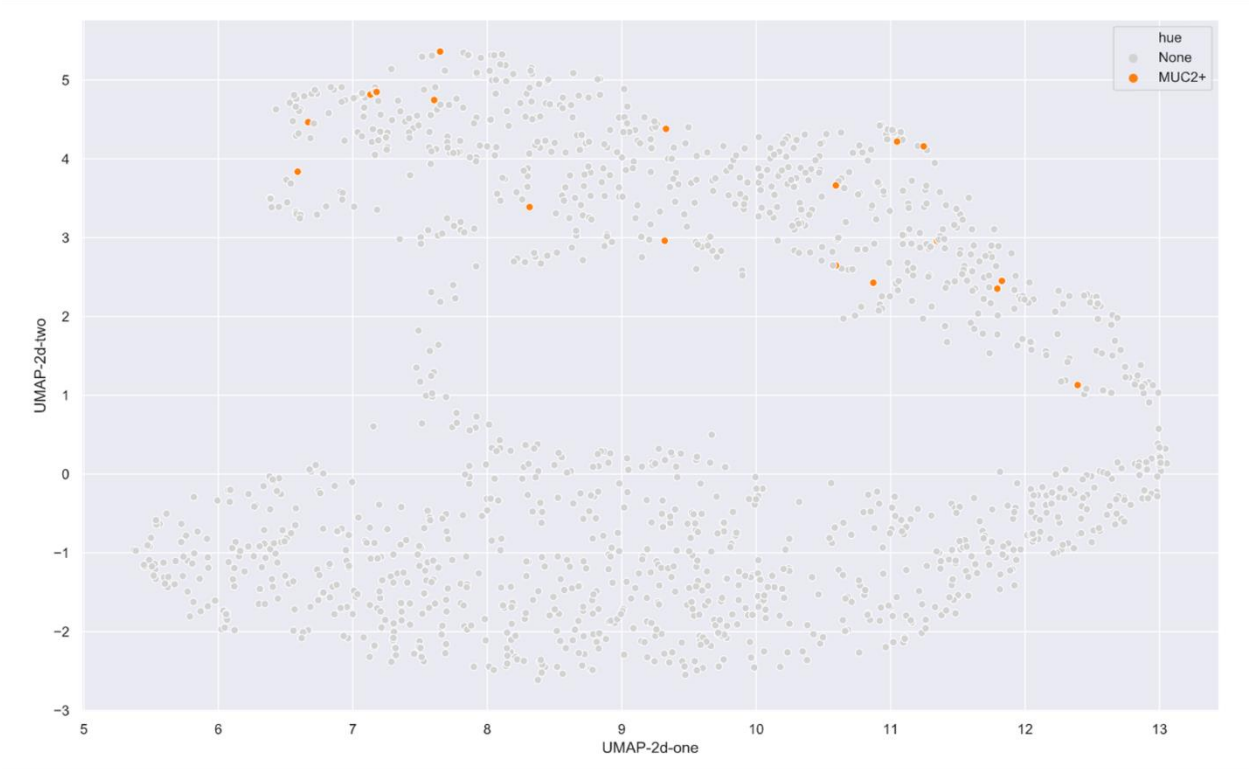
In the literature, many groups including 10X genomics had proposed algorithms that aim to compensate for the dropout rate of scRNA-seq such as Markov Affinity-based Graph Imputation of Cells (***MAGIC)*** and Recovery of Single-Cell Under-detected Expression (***RESCUE)*** (Dijk van *et al.*, 2017; Tracy, Yuan and Dries, 2019b)**.** We have experimented with many of these algorithms such as MAGIC shown in Figure 7.13 but found that there is a fundamental assumption that is not suitable in our case. The fundamental assumption of most dropout-compensation algorithms assumes clusters of cells are more or less 'identical'. The algorithm essentially groups cells and compensates the lost signal (dropout) by matching what should 'theoretically' be expressed in the same group of cells. A simplified version of explaining the logic behind this algorithm is for example if group A of cells expresses gene A, B, C, E. Following this logic, another cell that expresses gene A, C, E that is categorized in group A should also express B thus the expression of B is compensated.

This method may work more efficiently if there are distinct groups of cells such as those often found in a tissue or biopsy. The rationale is that the methods proposed are built based on the assumption of being able to find distinct clusters of cells. This is much more complex to parse out within a cell line, which most likely consists of cells that are of the same basic cell type (i.e. epithelial, endodermal, mesenchymal…etc.) compared to the diversity of cell types in tissue samples. In our case where we are mixing two cell lines of the same cell type and tissue of origin, the method most likely overfits and over-compensates. In Figure 7.13, the percentage of MUC2+ cells found in the cluster prior to applying ***MAGIC*** accurately reflects the percentage of MUC2+ cells by immunofluorescence, which is around 2.5%. Nonetheless, after ***MAGIC***, nearly all of the

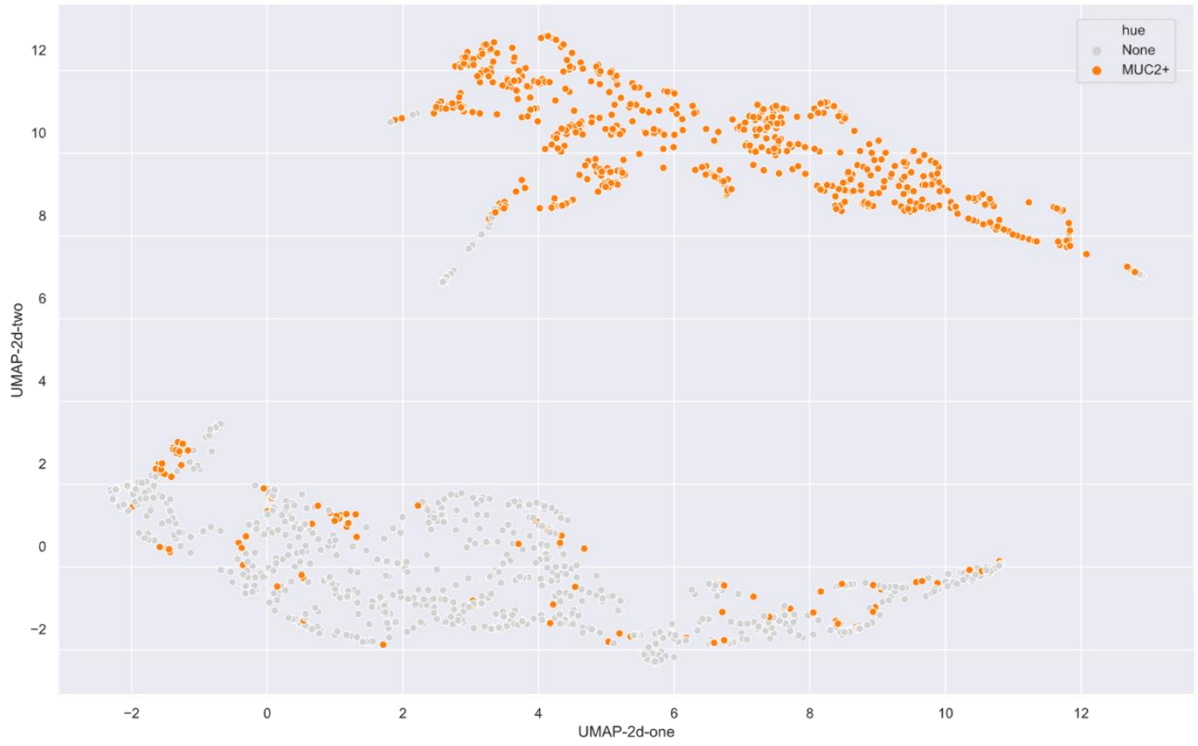
lower cluster is labeled as MUC2+ and even some in the top cluster, suggesting serious overfitting and overcompensation from the algorithm.

Figure 7.13 Comparing UMAP Prior and After MAGIC: Comparing the UMAP of post-processed scRNA-seq via our pipeline before and after applying *MAGIC*. We observe that in the UMAP post-*MAGIC*, the cluster separates better after compensation; nonetheless, nearly all cells in the top cluster are expressing MUC2. In IF experiments on MUC2 staining, the UMAP pre-*MAGIC* accurately reflects the percentage of cells expressing MUC2 (~2.5%), thus *MAGIC* is seriously overcompensating and so skewing downstream analysis.

UMAP of Postprocessed Input Prior ***MAGIC***



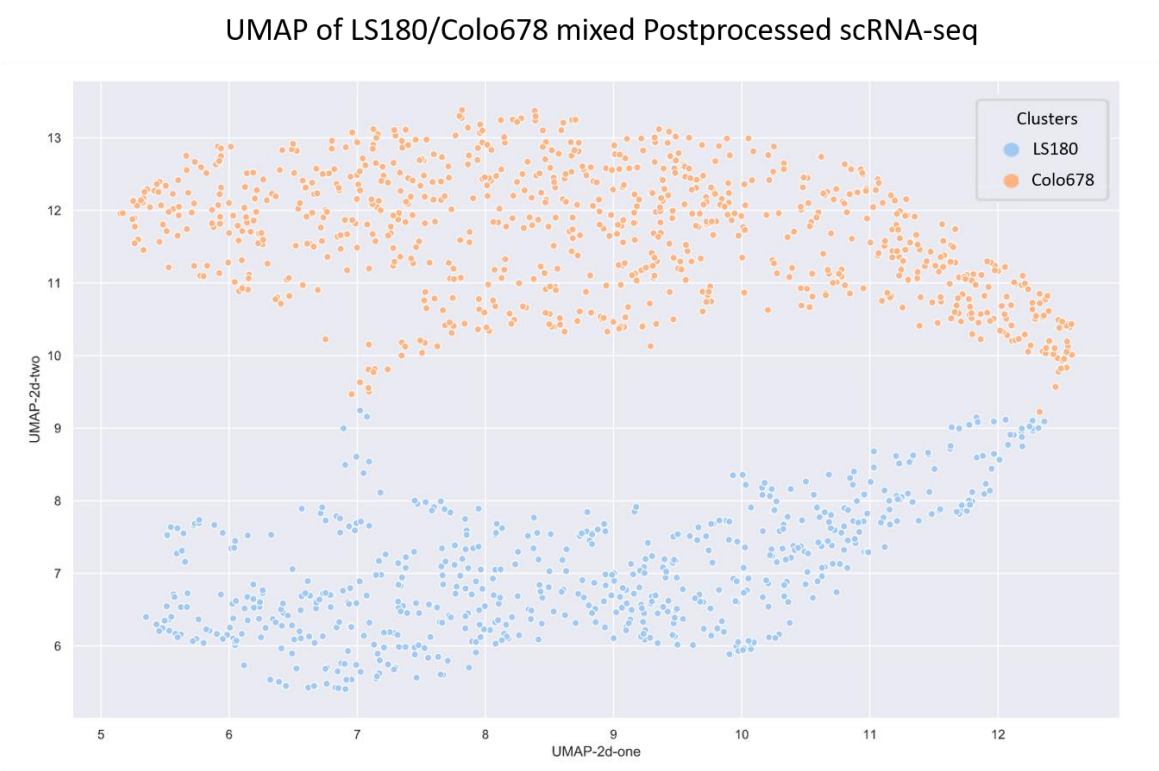
UMAP of Postprocessed Input After ***MAGIC***



7.2.4 Application of GMMchi-based Single Cell Sequencing Pipeline

We used k-means clustering (n=2) to extract the two clusters assumed to be LS180 and Colo678 as shown in Figure 7.14a. Post-separation of the two clusters, all downstream analysis was based on the raw counts and not the normalized counts. The two cell lines were different in their phenotypes (discussed in Chapter 4) in lumen formation as well as whether the cell line differentiated into goblet cells as shown in Figure 7.14b.

a)



b)

Table of Phenotypic Differences Between LS180 and Colo678

	Lumen Formation	Goblet Cell Forming
LS180	**Yes**	**Yes**
Colo678	**No**	**No**

Figure 7.14 K-means Clustering: (a) (K-means clustering (n=2) on the UMAP matrix of our scRNA-seq data delineates the two clusters labeled in blue (LS180) and orange (Colo678). (b) Table showing the phenotypic differences between LS180 and Colo678

Genes Controlling Lumen formation on a Single Cell Level

In Figure 7.15, we show that CDX1 and CDX2 are almost exclusively expressed in LS180 (bottom cluster), demonstrating a striking distinction between the two cell lines on a single cell level for markers involved in controlling lumen formation. However, there is a small subset of cells expressing CDX1 and CDX2 in the Colo678 (Figure 7.15a, 7.15c). Based on the heatmap in Figures 7.15a, 7.15c, we observe that the levels of CDX1 and CDX2 in Colo678 are comparable to the average level of expression seen in LS180. We do recognize that cells expressing the highest level of CDX1 and CDX2 are always found in LS180. However, the GMMchi analysis shown in Figure 7.15b, 7.15d demonstrates that the distribution of CDX1 and CDX2 expression for Colo678 (brown) is evenly spread across the overall distribution. This observation suggests that CDX1 and CDX2 expression in Colo678 was not substantially lower than those found in LS180 which prompts us to hypothesize whether CDX1 and CDX2-expressing cells in Colo678 are similar to those observed both in LS180 as well as on a cell line level.

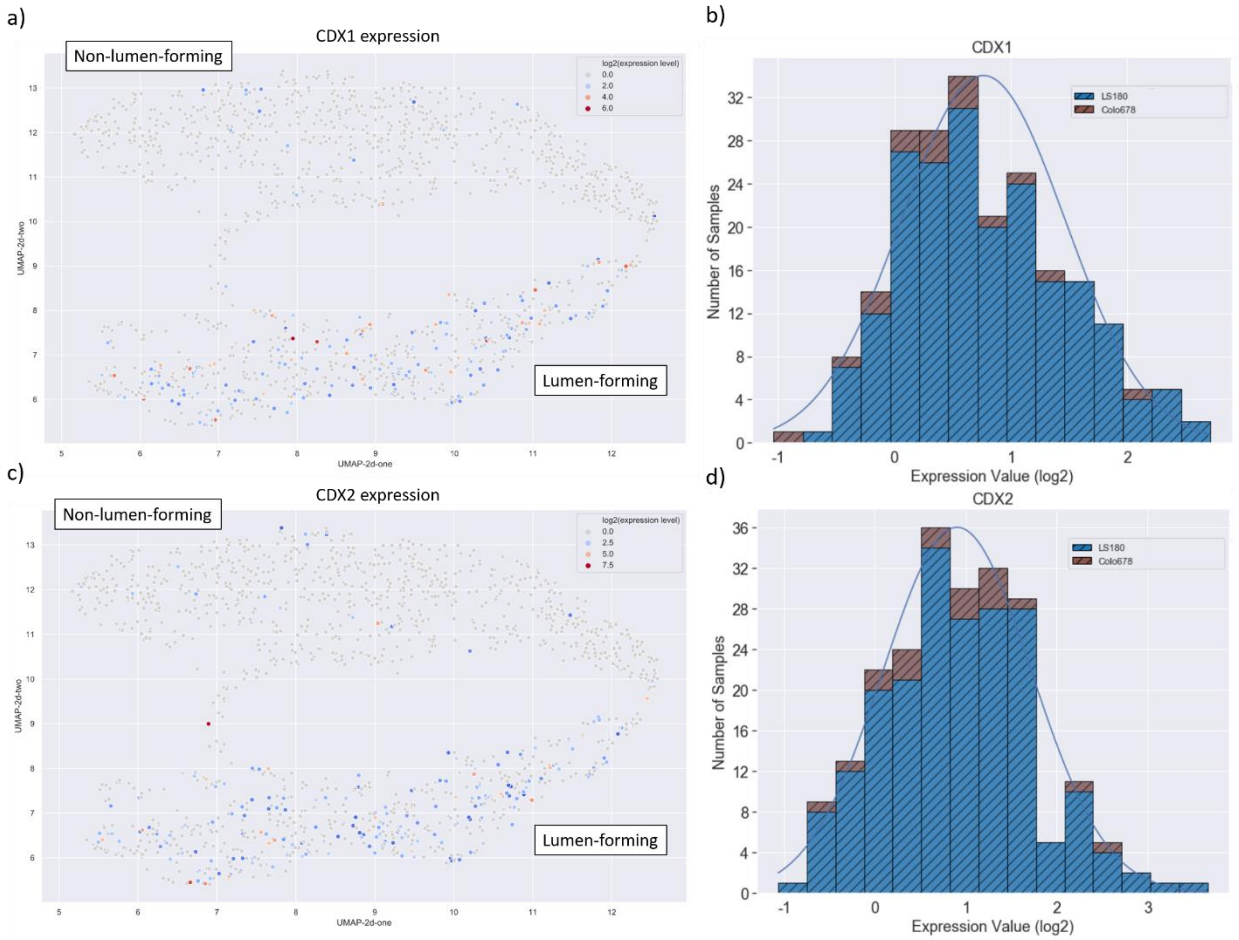


Figure 7.15 scRNA-seq Shows a Striking Difference in CDX1 and CDX2 Levels in Lumen vs Non-Lumen-Forming Cell Lines: CDX1 and CDX2 are important transcription factors for differentiation found highly associated in lumen forming cell lines. Both expressions are mapped onto the UMAP of our scRNA-seq data to demonstrate the heterogeneity of their expression on a single-cell level. (a) and (b) shows the expression levels of CDX1 and CDX2, respectively. The GMMchi analysis shows that Colo678 (brown) exhibits a distribution of expression across the overall distribution similar to LS180 (blue). This suggests that although CDX1 and CDX2-expressing cells in Colo678 are rare, they are not expressing levels substantially lower than those found in LS180.

This was an interesting observation because lumen formation experiments start with seeding a single cell in a layer of Matrigel and observing whether the single-cell differentiates into a lumen structure after some time. The observation of a CDX1 and CDX2 expressing subset in Colo678 begs the question of whether the failure of observing lumen formation in our previous experimental work is due to the high chance of seeding a non-CDX1 or CDX2-expressing Colo678

cell. Moreover, whether lumen formation is possible if we isolate and seed CDX1 and CDX2-expressing Colo678 cells or whether other mechanisms make it unable to form lumen structures despite expressing CDX1 or CDX2. This example further shows the strength of scRNA-seq and important questions we can ask by understanding single-cell expression.

Genes Controlling Goblet Cell Formation on a Single Cell Level

In Figure 7.13, we showed that MUC2 is only expressed in a small subpopulation that accounts for 21 out of a total of 578 cells in the LS180 cluster. This equates to a subset of 3.6% of cells expressing MUC2. In previous experimental work investigating the MUC2+ population in LS180, we observed a range of 2.5% - 3.5% cells expressing MUC2 via IF staining. This is strikingly close to the proportion of cells observed in our post-processed scRNA-seq data. Supposing that the 3.6% of MUC2+ cells overlap completely with the MUC2+ cells observed in IF staining, we will be able to draw this inference: our postprocessing pipeline results in an accurate estimate of viable cells that reflect the proportion of MUC2 expressing cells. MUC2 is the marker for mature goblet cells that secrete mucus. This would mean the probability of detection of mature goblet cells is close to, if not, 100%. Thus, we will base on this assumption that we have detected 100% of mature goblet cells seen in IF staining. We, again, need to assume that the antibody perfectly captures all MUC2 expression. This may vary if there exist other epitopes our antibody failed to detect. Nonetheless, under the assumption that our IF staining detected all MUC2+ cells and that our scRNA-seq showed the same proportion thus accurately detected all MUC2+ cells, we can now calculate the compound probability of detecting any other genes and whether the actual scRNA-seq supports our observation. For example, FCGBP is expressed in 9 out of 578 cells, accounting for 1.6% of the LS180 cluster. If we calculate the compound probability of a cell

expressing both MUC2 and FCGBP, we have to consider the high dropout rates in scRNA-seq thus need to account for the observation of MUC2+ and FCGBP+ expressing cells using data from IF staining and assume it to be the ground truth. We can calculate the probability of detection *r* and the proportion of cells expressing x gene p_x:

$$r_{MUC2} = \frac{p_{MUC2\ in\ scRNA-seq\ Data}}{p_{MUC2\ in\ IF}}$$

In our case, we continue our assumption mentioned above that we have accurately detected 100% of all MUC2+ cells in our scRNA-seq data:

$$r_{MUC2} = \frac{0.036}{0.036} * 100\% = 100\%$$

Since we do not have FCGBP data, we may, for example, assume that they have the same percentage of expression as MUC2. The r_{FCGBP} will be:

$$r_{FCGBP} = \frac{p_{FCGBP\ in\ scRNA-seq\ Data}}{p_{FCGBP\ in\ IF}}$$

In our case:

$$r_{FCGBP} = \frac{0.016}{0.036} * 100\% = 44.4\%$$

Two scenarios emerge and would lead to a different result where if we expect FCGBP and MUC2 to be expressed in the same cells, we will calculate the compound probability as:

$$r_{FCGBP} \times r_{MUC2} \times p_{MUC2\ and\ FCGBP} = p_{MUC2,\ \ FCGBP}$$

For example, in our case we would get:

$$1 \times 0.44 \times 0.036 = 0.016 \; or \; 1.6\%$$

$$1.6\% \times 578 = 9 \; cells \; expressing \; both \; MUC2 \; and \; FCGBP$$

In this case, the expected number of cells expressing both genes is close to the actual observation in our scRNA-seq data where we found 7 cells expressing both genes.

If we did not expect the two genes to be expressed in the same cells, we will then calculate the compound probability of the expected cells expressing both FCGBP and MUC2 to be:

$$r_{MUC2} \times p_{MUC2} \times r_{FCGBP} \times p_{FCGBP} = p_{MUC2,\ FCGBP}$$

In our case:

$$1 \times 0.036 \times 0.44 \times 0.036 = 0.00057 \; or \; 0.057\%$$

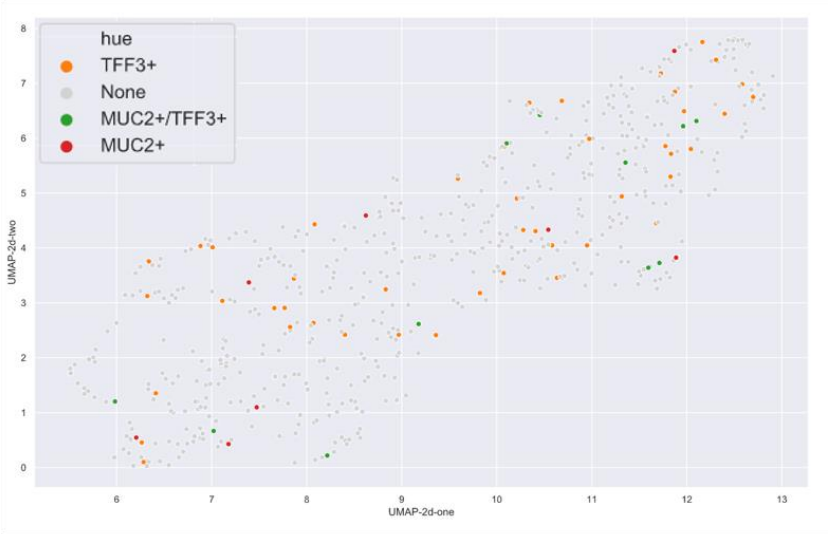
$$0.057\% \times 578 = 0.33 \; cells \; expressing \; both \; MUC2 \; and \; FCGBP$$

This equates to 0.33 cells expected to express both MUC2 and TFF3. This means that the observed 7 cells expressing both genes are nearly 21 times the expected number, making this observation highly significant in this scenario and so providing strong evidence from the single-cell data that MUC2 and FCGBP are expressed in the same cells, namely goblet cells.

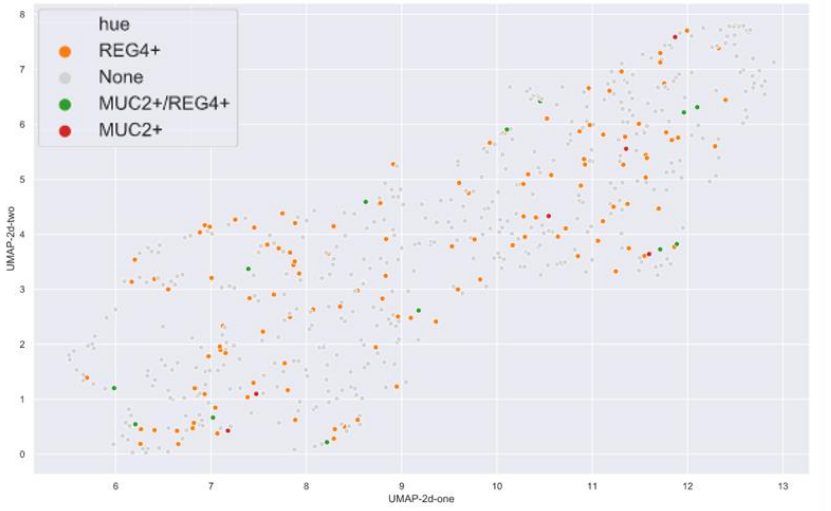
This suggests that another way of analyzing scRNA-seq data is to use methods shown in Chapter 4 based on using such as 2x2 analysis for clustering cell lines, treating individual cells as unique samples as compared to cell lines. We explored known MUC2 expression and ran ***GMM.run_hits*** to explore other genes associated with MUC2 expressing single cells in LS180. In Table 7.2, we find that MUC2 expression in individual LS180 cells is correlated with genes

previously found on a cell line level including GUCY2C, TFF2, TFF3, FCGBP, and SPINK4. TFF3, consistent with our previous analysis, exhibits an inclusion pattern with MUC2 skewed towards -/+. In Figure 7.16, we further show REG4 and TFF3 expressing cells do not always express MUC2 while MUC2 expressing cells are much more likely to express TFF3. FCGBP, on the other hand, contradicted our observation on the cell line level discussed in Chapter 4 and exhibits the opposite inclusion pattern skewed towards +/-. In fact, there is a striking list of genes that share this pattern in which they are rarely expressed without the expression of MUC2. This is an important piece of evidence that reinforces the idea that TFF3 and REG4 may be expressed before MUC2 maturation in goblet cells on a single-cell level. In this example, it shows that previous observations on the cell line level, which represents the global phenotype or the average expression level of all single cells, are validated within individual cells.

MUC2/TFF3 expressing cells



MUC2/REG4 expressing cells



MUC2/FCGBP expressing cells

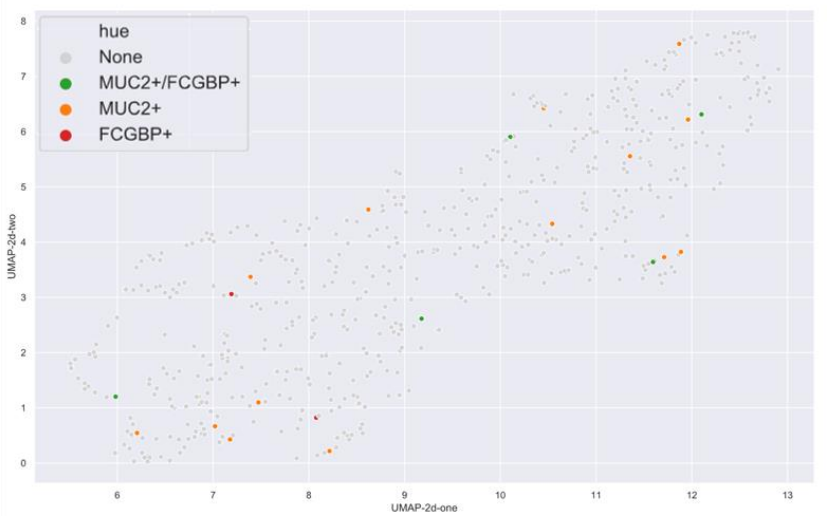


Figure 7.16 REG4 and TFF3 Shows Inclusion of MUC2 on a Single Cell Level: TFF3 and REG4 showing an inclusion of MUC2 where we observe a MUC2+ subpopulation of cells within the TFF3+ and REG4+ subpopulation. We observe a much rarer case of cells solely expressing MUC2, fitting the definition of MUC2+ cells being an inclusion of TFF3+ cells. On the other hand, we observe the opposite relationship with FCGBP where most FCGBP+ cells are expressing MUC2 while we observe significantly fewer cells solely expressing FCGBP, suggesting an inclusion of FCGBP within the MUC2+ subpopulation.

Table 7.2 Genes Associated with MUC2+ Single Cells: 2x2 table showing significantly associated genes with MUC2+ single cells.

Association with MUC2-expressing cells without Inclusions

MUC2	+/+	+/-	-/+	-/-	p-value		r value		Inclusion Criterion
GUCY2C	7	14	23	534	3.46E-05		0.25		No
NCALD	4	17	8	549	5.31E-04		0.23		No
TFF2	5	16	15	542	4.19E-04		0.22		No

Association with MUC2-expressing cells with Inclusions (skewed to -/+)

MUC2	+/+	+/-	-/+	-/-	p-value		r value		Inclusion Criterion
TFF3	12	9	48	509	7.61E-08		0.30		Significant
SLC52A2	13	8	82	475	1.97E-06		0.24		Significant

Association with MUC2-expressing cells with Inclusions (skewed to +/-)

MUC2	+/+	+/-	-/+	-/-	p-value		r value		Inclusion Criterion
FCGBP	7	14	2	555	9.72E-10		0.50		Significant
HMGCS2	9	12	32	525	2.95E-06		0.27		Significant
DLL4	3	18	2	555	3.96E-04		0.28		Significant
SPINK4	3	18	2	555	3.96E-04		0.28		Significant

Chapter 8: General Discussion

8.1 GMMchi Efficiently Identifies Mixtures of Normal and Non-normal Components in Gene Expression Analysis

Tumorigenesis is an evolutionary process that is driven by advantageous changes on the genetic or epigenetic level. Changes evident on the message level may consequently lead to a change on the functional level. In this chapter, we explored the GMMchi pipeline as a means of deciphering bimodal patterns within gene expression data across panels of cell lines or patient samples. The expectation is that samples with a mutation or epigenetic change will tend to cluster in a normal distribution while samples without the mutation or epigenetic change will cluster in a second normal distribution leading to a bimodal distribution pattern.

To address the shortcomings of standard Gaussian Mixture Modeling in expression pattern analysis, we built a novel GMM pipeline, coined GMMchi, that leverages an iterative process to deal with mixtures of normal and non-normal components. Since standard GMM was built on identifying mixtures of normal components, GMM fails to identify patterns in expression data consisting of a non-normal tail problem' as seen in Figure 3.4a. In Chapter 5, we demonstrated how we applied GMMchi in identifying CDH1-/VIM+ EMT cell lines (Figure 5.3). Another example is MUC2, the gene encoding a secreted glycoprotein found in goblet cells, that exhibited a unimodal distribution with a tail. The non-normal tail suggests a group of outliers that exhibit a unique phenotype separate from the main normal component which was later shown to have a significant degree of overlap with goblet cell formation in CRC cell lines confirmed via IF staining (Table 4.2b).

In the GMMchi pipeline, we included four scenarios when dealing with the tail problem including unimodal with a left tail, unimodal with a right tail, bimodal with a left tail, and bimodal

with a right tail. One important scenario that we did not include in the pipeline is the consideration of a unimodal with tails on both ends (right and left). The reason we did not include this scenario in the pipeline is due to the lack of such a pattern in our expression data; however, it will be ideal to incorporate this in future versions to consider all possible scenarios.

During validation tests on simulated data, we can see that GMMchi is sensitive in identifying non-normal tails in a mixture of normal and non-normal components shown in Figure 3.7a, 3.7b. The validation tests showed that GMMchi is more sensitive at identifying unimodal distributions with a tail by consistently outperforming identifying bimodal distribution with a tail. Even more significant is the degrading accuracy of identifying a bimodal distribution with a tail as the spread of the tail increases beyond 7 standard deviations. We explained that the degrading accuracy was a direct consequence of the method GMMchi used to determine the number of bins. The Mann and Wald criterion (Bowman, D'Agostino and Stephens, 1988) proposes $b = 1.88 \times n^{2/5}$ where *b* represents the number of bins. In other words, the formula relied solely on the sample size *n* to determine the number of bins *b* irrespective of the spread of the data. As the data points move further away from one another (increasing spread), the histogram loses granularity with a static number of bins, resulting in bimodal distributions with means that are close to one another being squished into bins that can be mistakenly represented as a unimodal distribution (Figure S1). This can be addressed by incorporating a bin number criterion that takes into account the spread of the total distribution. Nonetheless, in gene expression data, it is rare to find tails that span beyond 7 standard deviations from the main component. Thus, GMMchi performs stably and robustly in the context of gene expression analysis.

Overall, GMMchi performs consistently well (Figure 3.7c) and reaches an accuracy of 85% once the sample data are sufficiently large at around a minimum of 90 samples. Our microarray data consists of 78 CRC cell lines, which according to the data given above should achieve an overall accuracy of nearly 80%. The Cancer Genome Atlas (TCGA) data on colorectal cancer consists of 637 patient samples thus will be able to achieve higher accuracy. The initial motivation for establishing the GMMchi pipeline is to identify patterns that represent a well-defined bimodal distribution that can correlate back to a well-defined genetic or epigenetic change. In Supplementary Figure S2, we show that the sensitivity of GMMchi increases with the reducing overlap between the mixture of components, suggesting the high sensitivity in identifying well-defined bimodal distributions.

GMMchi has shown its ability to accurately define the cutoff between mixtures of normal and non-normal tails in gene expression data. Ultimately, we believe GMMchi provides a new method for researchers to perform large-scale pattern analysis to search for novel distinctive cues for downstream applications and marker discovery.

8.2 GMMchi Provides a Powerful Approach to Many of Aspects in Gene Expression Analysis

The wide range of GMMchi applications is demonstrated through phenotype associations, mutation and methylation associations, cluster analysis, and multi-cell type purification in TCGA patient samples. We have shown the simplicity and advantages of using 2x2 tables as the basis of the analysis of GMMchi categorized expression value. We also introduced the notion of inclusion and exclusion criteria and their significance when interpreting 2x2 association studies in a biological context. In the phenotype association analysis, genes that exhibit an inclusion skewed towards either -/+ or +/- imply an entirely different role in the larger context related to the associated phenotype. For example, NRCAM, a differentiation marker (Conacci-Sorrell *et al.*, 2002), is found to be an inclusion under lumen formation skewed towards +/-, which suggested that it is likely not required for lumen formation but is expressed when cells differentiate to form lumen structures. On the other hand, genes like ARG2 and AGR3 are found to be inclusions skewed towards -/+ under goblet-cell formation. The AGR family is thought to play an important role in protein folding by catalyzing disulfide bond formation. This pattern of association suggests that AGR2 and AGR3 are likely required for goblet cell formation. In the literature, intestinal mucin 2 (MUC2) secreted by goblet cells consists of units of larger polymers formed by disulfide bond linking between mucin subunits (Ambort *et al.*, 2011; Bäckström *et al.*, 2013). This suggests the critical role AGR proteins play in MUC2 maturation required for goblet cell maturation, reinforcing the association pattern observed. In the association study with RER status, we compared our results with *Wilding et al.* published work on genes associated with MMR deficiency (Wilding *et al.*, 2010). We found that 22 out of the 30 differentially expressed RER-related genes overlapped with our results (Table 4.5). In addition, we showed that 7 out of the 8

non-overlapping genes did not return as significantly associated genes in our study due to their lack of bimodality. All 7 genes exhibited a unimodal distribution (Figure 4.7). This particular example further shows the advantages and disadvantages of 2x2 table analysis. In this example, the disadvantage of 2x2 table analysis is the limitation to studying the association between bimodal genes despite an observable trend of association in unimodal genes. This can be solved by analyzing data using differential gene expression (DGE) or any other analysis based on continuous data common in many studies. In the paper, the top 30 differentially expressed RER-related genes were determined using DGE analysis. Nonetheless, this example also demonstrates the advantage of 2x2 table analysis in its capability of selecting the most significantly associated genes consisting of distinctive groups. This will mean that associations in a 2x2 analysis leveraging GMMchi can point to a very significant associations where the genes exhibit a categorical difference in expression level instead of a general trend of association.

In the next section, we looked for associations between well-defined bimodal genes and mutations, methylation, and enhancers. Similar to previous studies (Kawaguchi *et al.*, 2019), we found SMAD4 and p53 mutations negatively associated with p53 expression. Interestingly, we noticed a striking exclusion criterion with no cell lines in the -/- category, suggesting not every p53 mutation inactivated p53 expression. In the literature, we found other upstream regulators such as lysine methylation and acetylation were shown to control p53 expression (Ivanov *et al.*, 2007; West and Gozani, 2011). CDX1 was previously shown to be controlled by promoter methylation in our lab (Wong *et al.*, 2004). We found the same association between CDX1 methylation and CDX1 expression in this study. In addition, we found a novel methylation association across all datasets for KDM5D. This was a striking discovery as the association of

KDM5D was repeatedly found between both our microarray data and CCLE as well as the two available methylation datasets in CCLE. The significance of the novel mutations and methylation patterns found in this study would require further experimental work to validate and explore to understand their significance in tumor biology. In our GMMchi-based cluster analysis, our ability to categorize data enabled us to incorporate biological context into the analysis, which includes background threshold, non-expressing cell lines, low expressing cell lines, and high expressing cell lines. In traditional trend-based association methods using continuous data, it will be difficult to engineer an easy way to include these attributes.

Furthermore, we also demonstrated the application of GMMchi in purifying non-tumor expression from bulk tumor expression in the TCGA patient samples. In the literature, algorithms such as CIBERSORT developed by *Newman et al* (Newman *et al.*, 2015) perform deconvolution on bulk multi-cell type RNA-seq data that aims to provide an estimation of the abundances of cell subsets in tissue expression profiles (Chen *et al.*, 2018). The CIBERSORT algorithm takes in bulk RNA-seq data as well as RNA-seq data that represent individual cell types to deconvolute the bulk RNA-seq data and create an estimation of the cell-type composition. The main usability issue with CIBERSORT is the need for an accurate 'ground truth' for each cell type. This requires cell-type-specific RNA-seq or single-cell sequencing data to maximize the value of CIBERSORT, yet such data are either difficult or expensive to retrieve. Not to mention the lack of a definite definition for individual cell types where researchers are still currently attempting to discover, address, and establish a common definition for different cell types (Amit *et al.*, 2018). The main advantages of our model include the ease of use as well as a method to examine commonly paired experiments (i.e., normal and tumor) where data can be relatively easily retrieved and explored. Moreover,

this allows users to analyze purified tumor expression from the vast databases already available online. As shown, we believe the generalizability and robustness of GMMchi will be able to benefit researchers with a new tool to perform pattern analysis and incorporate it into existing pipelines that may shed new insights onto a vast genre of datasets.

8.3 Methylation is a Novel Mechanism Regulating ZEB1 Expression in CRC Cell Lines

EMT occurs during stages of development, wound healing, tissue fibrosis, as well as cancer progression (Kalluri and Weinberg, 2009; Brabletz *et al.*, 2018). It is a dynamic process where epithelial cells transform into cells exhibiting a mesenchymal phenotype. However, epithelial tumor cells in a cancer setting undergo incomplete EMT that has been suggested to enhance their invasiveness and cellular plasticity (Saitoh, 2018). These phenotypic changes are believed to play an important role in tumor metastasis. In our panel of 78 CRC cell lines, we identified 6 cell lines that exhibited low CDH1 and high VIM expression that represented an EMT-like phenotype, not due to mutations in E-cadherin. We found that ZEB1 was the only known EMT TF differentially upregulated in EMT cell lines. ZEB1 is a gene of the ZEB family known to directly bind to the TATA region located in the promoter region of many EMT-related genes (Dykes *et al.*, 2017). We established ZEB1 knockout clones from EMT cell lines to assess the effect of ZEB1 ablation in relation to the EMT phenotype. Out of the 6 EMT cell lines, we performed ZEB1-KO, only 2 EMT cell lines OXCO1 and SW620 changed expression profile resembling MET with strong upregulation of CDH1 and a loss of VIM expression. In addition, we observed a dramatic decrease in motility without any effects on the doubling time. Nonetheless, in the 4 ZEB1-KO cell lines that did not exhibit a change in CDH1 and VIM expression despite the loss of ZEB1 expression, we suspected MITF, an E-box binding transcription factor, may play a role in mediating EMT-related genes via promoter binding since MITF targets the same binding domain as ZEB1 (Terragni *et al.*, 2011; Pogenberg *et al.*, 2012; Um *et al.*, 2012). In addition, we found that CDH1 methylation may potentially play a role in the lack of CDH1 upregulation in Colo320DM shown in Chapter 4 (Supplementary Table S1). Although the CCLE data lacks the methylation

profile on Colo320HSR, this observation implies the high possibility of CDH1 methylation in the duplicate of Colo320DM, Colo320HSR.

In our ZEB1-KO characterization, the colocalization of β-catenin and CDH1 emerged in the OXCO1 and SW620 ZEB1-KO clones. This is a significant finding since it reinforces previous observations on β-catenin membrane expression in CDH1 expressing cell lines and β-catenin nuclear expression in CDH1 non-expressing cell lines. It is known that β-catenin contributes to the activation of Wnt target genes through the formation of a transcription complex with members of the TCF/LEF family in the nucleus (Cadigan and Waterman, 2012). Previous work by *Felix et al.* demonstrated that siCDH1 knockdown experiments led to the nuclear translocation of β-catenin and an enhancement in β-catenin/TCF-dependent reporter activity (Kuphal and Behrens, 2006). Furthermore, multiple studies have shown an enhanced Wnt/β-catenin signaling activity associated with EMT (Howard *et al.*, 2011; Basu, Cheriyamundath and Ben-Ze'ev, 2018). This evidence suggests that E-cadherin expression regulates the availability of cytoplasmic β-catenin by CDH1/β-catenin membrane colocalization, effectively preventing β-catenin from translating into the nucleus. In OXCO1 and SW620 ZEB1-KO clones, the observation of CDH1/β-catenin colocalization reinforces these findings and supports a possible downregulation of Wnt signaling activity that may contribute to the reversal of the EMT phenotype.

In the literature, a wide range of EMT TF upstream regulators have been proposed including TGFβ, Wnt, Notch, Hedgehog, and the hypoxic pathway (Zhang, Tian and Xing, 2016; Zhu *et al.*, 2018). In this thesis, we showed a novel mechanism of ZEB1 promoter methylation. Previous microarray analysis on the effect of 5-aza treatment on 8 non-EMT cell lines and 3 EMT cell lines revealed uniform ZEB1 upregulation coincided with VIM downregulation and CDH1

upregulation. We further repeated this experiment across a panel of 20 non-EMT cell lines and found a strong trend in ZEB1 upregulation. The level of upregulation varied among the cell lines. Nonetheless, this suggests a likely scenario of ZEB1 methylation in CRC cell lines. We also ran association studies using TCGA and CCLE methylation data. In this analysis, we identified likely candidates of CpG island methylation located on chromosome 10, 21320962 to 31320976 as well as 31606675 to 31606814. Unfortunately, due to the COVID-19 pandemic, we were unable to experimentally verify these observations. These CpG islands will serve as important cues to future targeted methylation sequencing experiments. Until now, there have been no studies indicating the specific sites of ZEB1 methylation or the role of ZEB1 methylation in tumors undergoing EMT. We only found studies showing the implication of ZEB1 methylation through indirect observations (Meidhof *et al.*, 2015; Lindner *et al.*, 2020). The validation of methylation-mediated ZEB1 regulation will further increase our understanding of the upstream regulators of EMT, especially in CRC, as well as potential targets for reversing tumor EMT in future research.

8.4 GMMchi-based scRNA-seq Postprocessing Pipeline Efficiently Removes Low-Quality Barcodes in a Mix of Similar Cell Type

The rapid progress in the development of single-cell sequencing technologies in recent years has provided an important tool in uncovering biological diversity at the single cell level. These single-cell technologies will allow researchers to reveal rare cell populations, complex cell-cell networks, as well as single-cell lineage tracking, entirely transforming the avenue of our understanding in cell biology (Hwang, Lee and Bang, 2018; Tang *et al.*, 2019). Inherently, the significance of single cell technologies makes postprocessing single cell sequencing output a critical step to maximize the value of single cell data. Common challenges in postprocessing single-cell data include identifying doublets, empty barcodes, RNA soup as well as accommodating the high drop rate of scRNA-seq (Young and Behjati, 2018; Qiu, 2020; Slyper *et al.*, 2020). In this thesis, we established the GMMchi-based scRNA-seq postprocessing pipeline to handle scRNA-seq data consisting of a mix of similar cell types. We have shown that standard pipelines such as ***scprep*** result in a poorly defined cluster membership for a significant portion of the data when separating cell types with less drastic differences in their molecular profile (Figure 7.1). The result of our pipeline efficiently removes low-quality barcodes to achieve a better separation of clusters as well as eliminates clustering and visualization biases towards highly expressed genes.

In Figure 7.3, we demonstrated the gradient of expression in highly expressed genes; moreover, we showed the high intercorrelation between these highly expressed genes (Figure 7.4) which presents a confounding set of genes that can together lead to the misrepresentation of the overall molecular profile of each cell during clustering and visualization. In the literature,

multiple studies have shown the large standard deviations of housekeeping gene expressions in scRNA-seq analysis (Thellin *et al.*, 1999; Eisenberg and Levanon, 2013; Lin *et al.*, 2019) which presents an issue when selecting genes that are thought to be stably expressed across all cells. *Lin et al.* identified a set of stably expressed genes (SEG) and proposed a framework for calculating gene stability index scSEGIndex for scRNA-seq data normalization (Lin *et al.*, 2019). Nonetheless, other studies using bulk and single-cell RNA-seq showed conflicting results with known housekeeping genes including GAPDH and ACTB stably expressed across different cell types (Noriega, Kohama and Urbanski, 2010; Caracausi *et al.*, 2017; Roy, McElhaney and Verschoor, 2020). In our pipeline, we took a different angle and hypothesized that the large standard deviations observed in housekeeping genes were mainly due to poor quality barcodes that exhibited a bimodal distribution. Similar to our findings of a bimodal distribution in GAPDH and ACTB (Figure 7.7a), *Lin et al.* observed a clear bimodality in GAPDH and ACTB for both the human and mouse scRNA-seq data (Lin *et al.*, 2019). We further showed that removing the barcodes/cells with distribution expressing distinctively lower levels of housekeeping gene via GMMchi analysis results in a significantly better separation of clusters (Figure 7.7b). Using our pipeline, we found interesting observations in CDX1 and CDX2 expressing cells in the non-lumen forming Colo678 cluster. We suspected that isolating CDX1 and CDX2 expressing cells from Colo678 may give rise to cells capable of differentiating into lumen structure observed in LS180 (Figure 7.15). Furthermore, we also showed methods for calculating the compound probability of cells co-expressing one or more genes using IF data. This was critical in determining the significance of any given observations found in scRNA-seq in order to compensate for the high potential drop rate of any given gene. Going through the literature for studies related to scRNA-

seq, we noticed that most studies were performed on tissue or biopsies consisting of clearly distinctive cell types (i.e. red blood cells, immune cell types, epithelial cells … etc.). The large differences in their molecular profile simplify the separation between clusters of different cell types. Nonetheless, we believe as research progresses into understanding the more granular differences between similar cell types, methods such as our pipeline demonstrated in this thesis will be an important step towards aiding researchers in maximizing the quality and value of their scRNA-seq data.

8.5 Limitations

However, this study has potential limitations that should be noted. First, ***GMMchi*** is based on the observation that gene expression data behave closely to a mixture of normal components and, in some cases, with non-normal outliers represented as a tail. In unique cases where the distribution patterns do not behave closely to our observation, ***GMMchi*** may return a poor fit. Therefore, it is suggested that any input data should first be observed by the user to make sure the data resembles closely to the model's assumption to prevent any misfit or misinterpretation of the analysis. Another limitation is the underperformance of ***GMMchi*** within small sample sizes. This should be carefully noted as inaccurate categorization may lead to misinterpretation of the data in smaller sample sizes. Second, the observation of promoter methylation of ZEB1 in colorectal cancer cell lines was made based on 5-azacytidine experiments as well as performing correlation studies using large public datasets (methylation and expression datasets). Since 5-azacytidine experiments remove methylation in all methylated sites, the increase in ZEB1 expression is at most an indirect observation that will require subsequent targeted methylation sequencing using candidates suggested in this study to provide strong direct evidence for ZEB1 promoter methylation. Third, although we have revalidated the observation for the myofibroblast recruitment phenotype across multiple myofibroblast cell lines, it is important that the PDGF-KO experiment is carried out to provide direct evidence of PDGF regulation of the observed recruitment phenotype

References

A.P.Dempster, N.M.Laird and D.B.Rubin (1997) 'maximum likelihood from incomplete data via the Em algorithm', *Journal of Royal ststistical society*, 39(1), pp. 1–38. Available at: http://www.jstor.org/stable/2984875.

Adegboyega, P. A. *et al.* (2002) 'Immunohistochemical study of myofibroblasts in normal colonic mucosa, hyperplastic polyps, and adenomatous colorectal polyps', *Archives of Pathology and Laboratory Medicine*, 126(7), pp. 829–836. doi: 10.5858/2002-126-0829-isomin.

Aiello, N. M. *et al.* (2018) 'EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration', *Developmental Cell*, 45(6), pp. 681-695.e4. doi: 10.1016/j.devcel.2018.05.027.

Aiello, N. M. and Kang, Y. (2019) 'Context-dependent EMT programs in cancer metastasis', *Journal of Experimental Medicine*, 216(5), pp. 1016–1026. doi: 10.1084/jem.20181827.

Aleksandrova, K. *et al.* (2015) 'A prospective study of the immune system activation biomarker neopterin and colorectal cancer risk', *Journal of the National Cancer Institute*, 107(4). doi: 10.1093/jnci/djv010.

Ambort, D. *et al.* (2011) 'Function of the CysD domain of the gel-forming MUC2 mucin', *Biochemical Journal*, 436(1), pp. 61–70. doi: 10.1042/BJ20102066.

Amit, I. *et al.* (2018) 'The human cell atlas', *arXiv*.

An, J., Zheng, Y. and Dann, C. T. (2017) 'Mesenchymal to Epithelial Transition Mediated by CDH1 Promotes Spontaneous Reprogramming of Male Germline Stem Cells to Pluripotency', *Stem Cell Reports*, 8(2), pp. 446–459. doi: 10.1016/j.stemcr.2016.12.006.

Ansieau, S., Collin, G. and Hill, L. (2014) 'EMT or EMT-promoting transcription factors, where to focus the light?', *Frontiers in Oncology*, 4(NOV), pp. 1–4. doi: 10.3389/fonc.2014.00353.

Aran, D., Sirota, M. and Butte, A. J. (2015) 'Systematic pan-cancer analysis of tumour purity', *Nature Communications*, 6, pp. 1–11. doi: 10.1038/ncomms9971.

Arechederra, M. *et al.* (2018) 'Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer', *Nature Communications*, 9(1). doi: 10.1038/s41467-018-05550-5.

Ashley, N., Yeung, T. M. and Bodmer, W. F. (2013) 'Stem cell differentiation and lumen formation in colorectal cancer cell lines and primary tumors', *Cancer Research*, 73(18), pp. 5798–5809. doi: 10.1158/0008-5472.CAN-13-0454.

Auclair, B. A. *et al.* (2007) 'Bone Morphogenetic Protein Signaling Is Essential for Terminal Differentiation of the Intestinal Secretory Cell Lineage', *Gastroenterology*, 133(3), pp. 887–896. doi: 10.1053/j.gastro.2007.06.066.

Bach, L. A. (2015) 'Recent insights into the actions of IGFBP-6', *Journal of Cell Communication and Signaling*, 9(2), pp. 189–200. doi: 10.1007/s12079-015-0288-4.

Bacigalupo, M. L. *et al.* (2015) 'Galectin-1 Triggers Epithelial-Mesenchymal Transition in Human Hepatocellular Carcinoma Cells', *Journal of Cellular Physiology*, 230(6), pp. 1298–1309. doi: 10.1002/jcp.24865.

Bäckström, M. *et al.* (2013) 'Increased understanding of the biochemistry and biosynthesis of MUC2 and other gel-forming mucins through the recombinant expression of their protein domains', *Molecular Biotechnology*, 54(2), pp. 250–256. doi: 10.1007/s12033-012-9562-3.

Bailey, M. H. *et al.* (2018) 'Comprehensive Characterization of Cancer Driver Genes and Mutations', *Cell*, 174(4), pp. 1034–1035. doi: 10.1016/j.cell.2018.07.034.

Ban, H. S. *et al.* (2019) 'Epigenetic alterations of Heat Shock Proteins (HSPs) in cancer', *International Journal of Molecular Sciences*, 20(19). doi: 10.3390/ijms20194758.

Banyard, J. and Bielenberg, D. R. (2015) 'The role of EMT and MET in cancer dissemination', *Connective Tissue Research*, 56(5), pp. 403–413. doi: 10.3109/03008207.2015.1060970.

Basu, S., Cheriyamundath, S. and Ben-Ze'ev, A. (2018) 'Cell–cell adhesion: Linking wnt/β-catenin signaling with partial emt and stemness traits in tumorigenesis [version 1; peer review: 4 approved]', *F1000Research*, 7. doi: 10.12688/F1000RESEARCH.15782.1.

Batlle, E. *et al.* (2002) 'β-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/EphrinB', *Cell*, 111(2), pp. 251–263. doi: 10.1016/S0092-8674(02)01015-2.

Baum, J. and Duffy, H. S. (2011) 'Fibroblasts and myofibroblasts: What are we talking about?', *Journal of Cardiovascular Pharmacology*, 57(4), pp. 376–379. doi: 10.1097/FJC.0b013e3182116e39.

Bergstrom, K. S. B. and Xia, L. (2013) 'Mucin-type O-glycans and their roles in intestinal homeostasis', *Glycobiology*, 23(9), pp. 1026–1037. doi: 10.1093/glycob/cwt045.

Bhat, H. and Kumar, N. (2010) 'On the derivation of the Bayesian Information Criterion. Technical report. http://nscs00.ucmerced.edu/~nkumar4/BhatKumarBIC.pdf', *School of Natural Sciences, University of California*, (November), pp. 1–4. Available at: http://nscs00.ucmerced.edu/~nkumar4/BhatKumarBIC.pdf.

Bochaton-Piallat, M. L., Gabbiani, G. and Hinz, B. (2016) 'The myofibroblast in wound healing and fibrosis: Answered and unanswered questions', *F1000Research*, 5(0), pp. 1–8. doi: 10.12688/f1000research.8190.1.

Bodmer, W. F. *et al.* (1988) 'Localization of the gene for familial adenomatous polyposis on chromosome 5', *Nature*, 328(6131), pp. 614–616. doi: 10.1038/328614a0.

Bowman, A. W., D'Agostino, R. B. and Stephens, M. A. (1988) 'Goodness-of-Fit Techniques.', *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 151(1), p. 225. doi: 10.2307/2982198.

Brabletz, T. *et al.* (2018) 'EMT in cancer', *Nature Reviews Cancer*, 18(2), pp. 128–134. doi: 10.1038/nrc.2017.118.

Brown, R. A. *et al.* (2002) 'Myofibroblasts and mechano-regulation of connective tissue remodelling', *Nature Reviews Molecular Cell Biology*, 3(5), pp. 349–363.

Cadigan, K. M. and Waterman, M. L. (2012) 'TCF/LEFs and Wnt signaling in the nucleus', *Cold Spring Harbor Perspectives in Biology*, 4(11). doi: 10.1101/cshperspect.a007906.

Calon, A. *et al.* (2012) 'Dependency of colorectal cancer on a TGF-beta-driven programme in stromal cells for metastasis initiation $watermark-text $watermark-text $watermark-text', *Cancer Cell*, 22(5), pp. 571–584. doi: 10.1016/j.ccr.2012.08.013.Dependency.

Caracausi, M. *et al.* (2017) 'Systematic identification of human housekeeping genes possibly useful as references in gene expression studies', *Molecular Medicine Reports*, 16(3), pp. 2397–2410. doi: 10.3892/mmr.2017.6944.

Chatterjee, A., Rodger, E. J. and Eccles, M. R. (2018) 'Epigenetic drivers of tumourigenesis and cancer metastasis', *Seminars in Cancer Biology*, 51(August 2017), pp. 149–159. doi: 10.1016/j.semcancer.2017.08.004.

Chen, B. *et al.* (2018) 'Profiling tumor infiltrating immune cells with CIBERSORT', *Methods in Molecular Biology*, 1711, pp. 243–259. doi: 10.1007/978-1-4939-7493-1_12.

Chen, K. S. *et al.* (2020) 'Transcription factors NFIA and NFIB induce cellular differentiation in high-grade astrocytoma', *Journal of Neuro-Oncology*, 146(1), pp. 41–53. doi: 10.1007/s11060-019-03352-3.

Cheng, F. *et al.* (2016) 'Vimentin coordinates fibroblast proliferation and keratinocyte differentiation in wound healing via TGF-β-Slug signaling', *Proceedings of the National Academy of Sciences of the United States of America*, 113(30), pp. E4320–E4327. doi: 10.1073/pnas.1519197113.

Choi, B. J. *et al.* (2017) 'Hypoxia induces epithelial-mesenchymal transition in colorectal cancer cells through ubiquitin-specific protease 47-mediated stabilization of Snail: A potential role of Sox9', *Scientific Reports*, 7(1), pp. 1–15. doi: 10.1038/s41598-017-15139-5.

Chu, C. H. *et al.* (2018) 'Prognostic values of EPDR1 hypermethylation and its inhibitory function on tumor invasion in colorectal cancer', *Cancers*, 10(10). doi: 10.3390/cancers10100393.

Chung, D. C. (2000) 'The genetic basis of colorectal cancer: Insights into critical pathways of tumorigenesis', *Gastroenterology*, 119(3), pp. 854–865. doi: 10.1053/gast.2000.16507.

Chung, Y. W. *et al.* (2015) 'Targeted disruption of PDE3B, but not PDE3A, protects murine heart from ischemia/reperfusion injury', *Proceedings of the National Academy of Sciences of the United States of America*, 112(17), pp. E2253–E2262. doi: 10.1073/pnas.1416230112.

Conacci-Sorrell, M. E. *et al.* (2002) 'Nr-CAM is a target gene of the β-catenin/LEF-1

pathway in melanoma and colon cancer and its expression enhances motility and confers tumorigenesis', *Genes and Development*, 16(16), pp. 2058–2072. doi: 10.1101/gad.227502.

D.W., P. *et al.* (1999) 'Myofibroblasts. II. Intestinal subepithelial myofibroblasts', *American Journal of Physiology - Cell Physiology*, 277(2 46-2), pp. C183–C201. Available at: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L29415032.

David, M. B. *et al.* (2020) 'Distinct populations of crypt-associated fibroblasts act as signaling hubs to control colon homeostasis', *PLoS Biology*, 18(12), pp. 1–20. doi: 10.1371/journal.pbio.3001032.

Dijk van, D. *et al.* (2017) 'MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data', *bioRxiv*. doi: 10.1101/111591.

Do, E. K. *et al.* (2017) 'Trib2 regulates the pluripotency of embryonic stem cells and enhances reprogramming efficiency', *Experimental and Molecular Medicine*, 49(11), pp. e401-11. doi: 10.1038/emm.2017.191.

Dykes, S. S. *et al.* (2017) 'Zinc finger E-box binding homeobox-1 (Zeb1) drives anterograde lysosome trafficking and tumor cell invasion via upregulation of Na+/H+ Exchanger-1 (NHE1)', *Molecular Carcinogenesis*, 56(2), pp. 722–734. doi: 10.1002/mc.22528.

Eisenberg, E. and Levanon, E. Y. (2013) 'Human housekeeping genes, revisited', *Trends in Genetics*, 29(10), pp. 569–574. doi: 10.1016/j.tig.2013.05.010.

Ertel, A. (2010) 'Article Commentary: Bimodal Gene expression and Biomarker Discovery', *Cancer Informatics*, 9, pp. 11–14. doi: 10.4137/CIN.S3456.

Fernández-Briera, A. *et al.* (2010) 'Effect of human colorectal carcinogenesis on the neural cell adhesion molecule expression and polysialylation', *Oncology*, 78(3–4), pp. 196–204. doi: 10.1159/000313699.

Fici, P. *et al.* (2017) 'Splicing factor ratio as an index of epithelial-mesenchymal transition and tumor aggressiveness in breast cancer', *Oncotarget*, 8(2), pp. 2423–2436. doi: 10.18632/oncotarget.13682.

Filip, S. *et al.* (2020) 'Distant metastasis in colorectal cancer patients—do we have new predicting clinicopathological and molecular biomarkers? A comprehensive review', *International Journal of Molecular Sciences*, 21(15), pp. 1–24. doi: 10.3390/ijms21155255.

Fisher RA (1918) 'The Correlation between Relatives on the Supposition of Mendelian Inheritance.', *Transactions of the Royal Society of Edinburgh*, 52, pp. 399–433.

Fre, S. (2015) 'Intestinal stem cells', *Stem Cell Biology and Regenerative Medicine*, 12(5), pp. 455–475. doi: 10.5937/mckg47-3311.

Gao, J. *et al.* (2015) 'Epithelial-to-Mesenchymal transition induced by TGF-β1 Is mediated by AP1-Dependent EpCAM expression in MCF-7 cells', *Journal of Cellular Physiology*, 230(4), pp. 775–782. doi: 10.1002/jcp.24802.

Garg, M. (2013) 'Epithelial-mesenchymal transition - activating transcription factors - multifunctional regulators in cancer', *World Journal of Stem Cells*, 5(4), p. 188. doi: 10.4252/wjsc.v5.i4.188.

Geiser, J., De Lisle, R. C. and Andrews, G. K. (2013) 'The zinc transporter Zip5 (Slc39a5) regulates intestinal zinc excretion and protects the pancreas against zinc toxicity', *PLoS ONE*, 8(11). doi: 10.1371/journal.pone.0082149.

Ghandi, M. *et al.* (2019) 'Next-generation characterization of the Cancer Cell Line Encyclopedia', *Nature*, 569(7757), pp. 503–508. doi: 10.1038/s41586-019-1186-3.

Gordon, C. A. *et al.* (2017) 'NUSAP1 promotes invasion and metastasis of prostate cancer', *Oncotarget*, 8(18), pp. 29935–29950. doi: 10.18632/oncotarget.15604.

Gould, S. J. and Eldredge, N. (1993) 'Punctuated equilibrium comes of age', *Nature*, 366(6452), pp. 223–227. doi: 10.1038/366223a0.

Gualtero, D. F. and Suarez Castillo, A. (2016) 'Biomarkers in saliva for the detection of oral squamous cell carcinoma and their potential use for early diagnosis: A systematic review', *Acta Odontologica Scandinavica*, 74(3), pp. 170–177. doi: 10.3109/00016357.2015.1110249.

Haque, A. *et al.* (2017) 'A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications', *Genome Medicine*, 9(1), pp. 1–12. doi: 10.1186/s13073-017-0467-4.

Haramis, A. P. G. *et al.* (2004) 'De Novo Crypt Formation and Juvenile Polyposis on BMP Inhibition in Mouse Intestine', *Science*, 303(5664), pp. 1684–1686. doi: 10.1126/science.1093587.

Hardwick, J. C. *et al.* (2008) 'Bone morphogenetic protein signalling in colorectal cancer', *Nature Reviews Cancer*, 8(10), pp. 806–812. doi: 10.1038/nrc2467.

Hawinkels, L. J. A. C. *et al.* (2014) 'Interaction with colon cancer cells hyperactivates TGF-β signaling in cancer-associated fibroblasts', *Oncogene*, 33(1), pp. 97–107. doi: 10.1038/onc.2012.536.

He, X. C. *et al.* (2004) 'BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt-β-catenin signaling', *Nature Genetics*, 36(10), pp. 1117–1121. doi: 10.1038/ng1430.

Heerboth, S. *et al.* (2015) 'EMT and tumor metastasis', *Clinical and Translational Medicine*, 4(1). doi: 10.1186/s40169-015-0048-3.

Henriksson, M. L. *et al.* (2011) 'Colorectal cancer cells activate adjacent fibroblasts resulting in FGF1/FGFR3 signaling and increased invasion', *American Journal of Pathology*, 178(3), pp. 1387–1394. doi: 10.1016/j.ajpath.2010.12.008.

Hoffman, J. A. *et al.* (2020) 'Single-cell RNA sequencing reveals a heterogeneous response to Glucocorticoids in breast cancer cells', *Communications Biology*, 3(1). doi: 10.1038/s42003-020-0837-0.

Howard, S. *et al.* (2011) 'A positive role of cadherin in wnt/β-catenin signalling during epithelial-mesenchymal transition', *PLoS ONE*, 6(8). doi: 10.1371/journal.pone.0023899.

Hsia, L. T. *et al.* (2016) 'Myofibroblasts are distinguished from activated skin fibroblasts by the expression of AOC3 and other associated markers', *Proceedings of the National Academy of Sciences of the United States of America*, 113(15), pp. E2162–E2171. doi: 10.1073/pnas.1603534113.

Hu, L. *et al.* (2015) 'Serine threonine tyrosine kinase 1 is a potential prognostic marker in colorectal cancer', *BMC Cancer*, 15(1), pp. 1–8. doi: 10.1186/s12885-015-1285-y.

Huber, A. H. and Weis, W. I. (2001) 'The structure of the β-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by β-catenin', *Cell*, 105(3), pp. 391–402. doi: 10.1016/S0092-8674(01)00330-0.

Humphries, A. and Wright, N. A. (2008) 'Colonic crypt organization and tumorigenesis', *Nature Reviews Cancer*, 8(6), pp. 415–424. doi: 10.1038/nrc2392.

Hwang, B., Lee, J. H. and Bang, D. (2018) 'Single-cell RNA sequencing technologies and bioinformatics pipelines', *Experimental and Molecular Medicine*, 50(8). doi: 10.1038/s12276-018-0071-8.

Iizuka, M. and Konno, S. (2011) 'Wound healing of intestinal epithelial cells', *World Journal of Gastroenterology*, 17(17), pp. 2161–2171. doi: 10.3748/wjg.v17.i17.2161.

Ivanov, G. S. *et al.* (2007) 'Methylation-Acetylation Interplay Activates p53 in Response to DNA Damage', *Molecular and Cellular Biology*, 27(19), pp. 6756–6769. doi: 10.1128/mcb.00460-07.

Jie, X. X., Zhang, X. Y. and Xu, C. J. (2017) 'Epithelial-to-mesenchymal transition, circulating tumor cells and cancer metastasis: Mechanisms and clinical applications', *Oncotarget*, 8(46), pp. 81558–81571. doi: 10.18632/oncotarget.18277.

Jones, B. W. *et al.* (2012) 'Cardiomyocytes from AKAP7 knockout mice respond normally to adrenergic stimulation', *Proceedings of the National Academy of Sciences of the United States of America*, 109(42), pp. 17099–17104. doi: 10.1073/pnas.1215219109.

Kalluri, R. and Weinberg, R. A. (2009) 'The basics of epithelial-mesenchymal transition', *Journal of Clinical Investigation*, 119(6), pp. 1420–1428. doi: 10.1172/JCI39104.

Kawaguchi, Y. *et al.* (2019) 'Mutation status of Ras, TP53, and Smad4 is superior to mutation status of Ras alone for predicting prognosis after resection of colorectal liver metastases', *Clinical Cancer Research*, 25(19), pp. 5843–5851. doi: 10.1158/1078-0432.CCR-19-0863.

Killeen, A. A. and Killeen, A. A. (2004) 'Molecular Origins of Cancer', *Principles of Molecular Pathology*, pp. 185–209. doi: 10.1007/978-1-59259-431-3_6.

Kim, J. K. *et al.* (2015) 'Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression', *Nature Communications*, 6. doi:

10.1038/ncomms9687.

Kim, Y. S. and Ho, S. B. (2010) 'Intestinal goblet cells and mucins in health and disease: Recent insights and progress', *Current Gastroenterology Reports*, 12(5), pp. 319–330. doi: 10.1007/s11894-010-0131-2.

Kinzler, K. W. and Vogelstein, B. (1996) 'Lessons from hereditary colorectal cancer', *Cell*, 87(2), pp. 159–170. doi: 10.1016/S0092-8674(00)81333-1.

Klingberg, F., Hinz, B. and White, E. S. (2013) 'The myofibroblast matrix: Implications for tissue repair andfibrosis', *Journal of Pathology*, 229(2), pp. 298–309. doi: 10.1002/path.4104.

Kosinski, C. *et al.* (2007) 'Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors', *Proceedings of the National Academy of Sciences of the United States of America*, 104(39), pp. 15418–15423. doi: 10.1073/pnas.0707210104.

Krishnaswamy, S. *et al.* (2014) 'Conditional density-based analysis of T cell signaling in single-cell data', *Science*, 346(6213). doi: 10.1126/science/1250689.

Kröger, C. *et al.* (2019) 'Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells', *Proceedings of the National Academy of Sciences of the United States of America*, 116(15), pp. 7353–7362. doi: 10.1073/pnas.1812876116.

Kuhnert, F. *et al.* (2004) 'Essential requirement for Wnt signaling in proliferation of adult small intestine and colon revealed by adenoviral expression of Dickkopf-1', *Proceedings of the National Academy of Sciences of the United States of America*, 101(1), pp. 266–271. doi: 10.1073/pnas.2536800100.

Kuphal, F. and Behrens, J. (2006) 'E-cadherin modulates Wnt-dependent transcription in colorectal cancer cells but does not alter Wnt-independent gene expression in fibroblasts', *Experimental Cell Research*, 312(4), pp. 457–467. doi: 10.1016/j.yexcr.2005.11.007.

Lapierre, L. A. *et al.* (2011) 'Transformation of rat intestinal epithelial cells by overexpression of Rab25 is microtubule dependent', *Cytoskeleton*, 68(2), pp. 97–111. doi: 10.1002/cm.20497.

Larsen, J. E. *et al.* (2016) 'ZEB1 drives epithelial-to-mesenchymal transition in lung cancer', *Journal of Clinical Investigation*, 126(9), pp. 3219–3235. doi: 10.1172/JCI76725.

Leigh, S. (2000) 'The importance of breast cancer research from a patient's view: the voices and visions of advocates', *Breast Cancer Research*, 2(S1). doi: 10.1186/bcr127.

Li, J. *et al.* (2019) 'Abstract LB-125: KDM5D as a novel gender-specific gene target for metastatic colorectal cancer', p. LB-125-LB-125. doi: 10.1158/1538-7445.sabcs18-lb-125.

Li, Z. *et al.* (2013) 'ETV1 induces epithelial to mesenchymal transition in human gastric cancer cells through the upregulation of Snail expression', *Oncology Reports*, 30(6), pp. 2859–2863. doi: 10.3892/or.2013.2776.

Lin, Y. *et al.* (2019) 'Evaluating stably expressed genes in single cells', *GigaScience*, 8(9). doi: 10.1093/gigascience/giz106.

Lindner, P. *et al.* (2020) 'EMT transcription factor ZEB1 alters the epigenetic landscape of colorectal cancer cells', *Cell Death and Disease*, 11(2). doi: 10.1038/s41419-020-2340-4.

Liu, H., Bebu, I. and Li, X. (2010) 'Microarray probes and probe sets', *Frontiers in Bioscience - Elite*, 2 E(1), pp. 325–338. doi: 10.2741/e93.

Loboda, A. *et al.* (2011) 'EMT is the dominant program in human colon cancer', *BMC Medical Genomics*, 4(1), p. 9. doi: 10.1186/1755-8794-4-9.

Lu, W. and Kang, Y. (2019) 'Epithelial-Mesenchymal Plasticity in Cancer Progression and Metastasis', *Developmental Cell*, 49(3), pp. 361–374. doi: 10.1016/j.devcel.2019.04.010.

Lygren, B. *et al.* (2007) 'AKAP complex regulates Ca2+ re-uptake into heart sarcoplasmic reticulum', *EMBO Reports*, 8(11), pp. 1061–1067. doi: 10.1038/sj.embor.7401081.

Ma, S. *et al.* (2018) 'The lncRNA LINC00675 regulates cell proliferation, migration, and invasion by affecting Wnt/β-catenin signaling in cervical cancer', *Biomedicine and Pharmacotherapy*, 108(October), pp. 1686–1693. doi: 10.1016/j.biopha.2018.10.011.

Van Der Maaten, L. and Hinton, G. (2008) 'Visualizing data using t-SNE', *Journal of Machine Learning Research*, 9, pp. 2579–2625.

Malik, S., Huang, C. -F and Schmidt, J. (1995) 'The Role of the CANNTG Promoter Element (E box) and the Myocyte-enhancer-binding-factor-2 (MEF-2) Site in the Transcriptional Regulation of the Chick Myogenin Gene', *European Journal of Biochemistry*, 230(1), pp. 88–96. doi: 10.1111/j.1432-1033.1995.0088i.x.

Manning, B. D. and Toker, A. (2017) 'AKT/PKB Signaling: Navigating the Network', *Cell*, 169(3), pp. 381–405. doi: 10.1016/j.cell.2017.04.001.

Mar, J. C. (2019) 'The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond', *Biophysical Reviews*, 11(1), pp. 89–94. doi: 10.1007/s12551-018-0494-4.

Maric, I. *et al.* (2012) 'Bone Morphogenetic Proteins and Signaling Pathway in Inflammatory Bowel Disease', *Inflammatory Bowel Disease - Advances in Pathogenesis and Management*. doi: 10.5772/29427.

Marshman, E., Booth, C. and Potten, C. S. (2002) 'The intestinal epithelial stem cell', *BioEssays*, 24(1), pp. 91–98. doi: 10.1002/bies.10028.

Marx, V. (2013) 'The big challenges of big data', *Nature*, 498(7453), pp. 255–260. doi: 10.1038/498255a.

Mayr, E. and Provine, W. B. (1998) 'The evolutionary synthesis: Perspectives on the unification of biology'.

McCarthy, A. J. *et al.* (2019) 'Heterogenous loss of mismatch repair (MMR) protein

expression: a challenge for immunohistochemical interpretation and microsatellite instability (MSI) evaluation', *Journal of Pathology: Clinical Research*, 5(2), pp. 115–129. doi: 10.1002/cjp2.120.

McInnes, L., Healy, J. and Melville, J. (2018) 'UMAP: Uniform manifold approximation and projection for dimension reduction', *arXiv*.

McKaig, B. C. *et al.* (1999) 'Normal human colonic subepithelial myofibroblasts enhance epithelial migration (restitution) via TGF-β3', *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 276(5 39-5), pp. 1087–1093. doi: 10.1152/ajpgi.1999.276.5.g1087.

Mclachlan, G. J. and Rathnayake, S. (2014) 'On the number of components in a Gaussian mixture model', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5), pp. 341–355. doi: 10.1002/widm.1135.

Medema, J. P. and Vermeulen, L. (2011) 'Microenvironmental regulation of stem cells in intestinal homeostasis and cancer', *Nature*, 474(7351), pp. 318–326. doi: 10.1038/nature10212.

Meidhof, S. *et al.* (2015) ' ZEB 1-associated drug resistance in cancer cells is reversed by the class I HDAC inhibitor mocetinostat ', *EMBO Molecular Medicine*, 7(6), pp. 831–847. doi: 10.15252/emmm.201404396.

Merlo, L. M. F. *et al.* (2006) 'Cancer as an evolutionary and ecological process', *Nature Reviews Cancer*, 6(12), pp. 924–935. doi: 10.1038/nrc2013.

Miyazono, K. (2009) 'Transforming growth factor-β signaling in epithelial-mesenchymal transition and progression of cancer', *Proceedings of the Japan Academy Series B: Physical and Biological Sciences*, 85(8), pp. 314–323. doi: 10.2183/pjab.85.314.

Moore, R., Carlson, S. and Madara, J. L. (1989) 'Villus contraction aids repair of intestinal epithelium after injury', *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 257(2). doi: 10.1152/ajpgi.1989.257.2.g274.

Moyer, R. A. *et al.* (2007) 'Rho activation regulates CXCL12 chemokine stimulated actin rearrangement and restitution in model intestinal epithelia', *Laboratory Investigation*, 87(8), pp. 807–817. doi: 10.1038/labinvest.3700595.

Newman, A. M. *et al.* (2015) 'Robust enumeration of cell subsets from tissue expression profiles', *Nature Methods*, 12(5), pp. 453–457. doi: 10.1038/nmeth.3337.

Noah, T. K. *et al.* (2010) 'SAM pointed domain ETS factor (SPDEF) regulates terminal differentiation and maturation of intestinal goblet cells', *Experimental Cell Research*, 316(3), pp. 452–465. doi: 10.1016/j.yexcr.2009.09.020.

Noriega, N. C., Kohama, S. G. and Urbanski, H. F. (2010) 'κMicroarray analysis of relative gene expression stability for selection of internal reference genes in the rhesus macaque brain', *BMC Molecular Biology*, 11. doi: 10.1186/1471-2199-11-47.

Nowell, P. C. (1976) 'The clonal evolution of tumor cell populations', *Science*, 194(4260), pp. 23–28. doi: 10.1126/science.959840.

Osaki, L. H. and Gama, P. (2013) 'MAPKs and signal transduction in the control of gastrointestinal epithelial cell proliferation and differentiation', *International Journal of Molecular Sciences*, 14(5), pp. 10143–10161. doi: 10.3390/ijms140510143.

Pakyari, M. *et al.* (2013) 'Critical Role of Transforming Growth Factor Beta in Different Phases of Wound Healing', *Advances in Wound Care*, 2(5), pp. 215–224. doi: 10.1089/wound.2012.0406.

Palacio-Rúa, K. A. *et al.* (2014) 'Genetic analysis in APC, KRAS, and TP53 in patients with stomach and colon cancer', *Revista de Gastroenterologia de Mexico*, 79(2), pp. 79–89. doi: 10.1016/j.rgmx.2014.05.001.

Park, S. W. *et al.* (2009) 'The protein disulfide isomerase AGR2 is essential for production of intestinal mucus', *Proceedings of the National Academy of Sciences of the United States of America*, 106(17), pp. 6950–6955. doi: 10.1073/pnas.0808722106.

Patrik, D. (2005) 'How does gene expression clustering work ?', *Nature biotechnology*, 23(12), pp. 1499–1501.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: http://scikit-learn.sourceforge.net.

Pena, C. *et al.* (2013) 'STC1 expression by cancer-associated fibroblasts drives metastasis of colorectal cancer', *Cancer Research*, 73(4), pp. 1287–1297. doi: 10.1158/0008-5472.CAN-12-1875.

Peter Kalugin (2017) *Myofibroblast Migration and Signaling in in vitro Colorectal Cancer Models*. University of Oxford.

Pinto, D. *et al.* (2003) 'Canonical Wnt signals are essential for homeostasis of the intestinal epithelium', *Genes and Development*, 17(14), pp. 1709–1713. doi: 10.1101/gad.267103.

Pogenberg, V. *et al.* (2012) 'Restricted leucine zipper dimerization and specificity of DNA recognition of the melanocyte master regulator MITF', *Genes and Development*, 26(23), pp. 2647–2658. doi: 10.1101/gad.198192.112.

Pomerantz, M. M. and Freedman, M. L. (2011) 'The genetics of cancer risk', *Cancer Journal*, 17(6), pp. 416–422. doi: 10.1097/PPO.0b013e31823e5387.

Poulogiannis, G., Frayling, I. M. and Arends, M. J. (2010) 'DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome', *Histopathology*, 56(2), pp. 167–179. doi: 10.1111/j.1365-2559.2009.03392.x.

Powell, S. M. *et al.* (1992) 'APC mutations occur early during colorectal tumorigenesis', *Nature*, 359(6392), pp. 235–237. doi: 10.1038/359235a0.

Qiu, P. (2020) 'Embracing the dropouts in single-cell RNA-seq analysis', *Nature Communications*, 11(1), pp. 1–9. doi: 10.1038/s41467-020-14976-9.

Qutub, A. A. and Popel, A. S. (2008) 'Reactive Oxygen Species Regulate Hypoxia-Inducible Factor 1α Differentially in Cancer and Ischemia', *Molecular and Cellular Biology*, 28(16), pp. 5106–5119. doi: 10.1128/mcb.00060-08.

Ran, D. *et al.* (2019) 'scDoc : Correcting Drop-out Events in Single-cell RNA-seq Data', pp. 1–28.

Rauluseviciute, I., Drabløs, F. and Rye, M. B. (2020) 'DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation', *BMC Medical Genomics*, 13(1). doi: 10.1186/s12920-020-0657-6.

Reischmann, N. *et al.* (2020) 'BRAFV600E drives dedifferentiation in small intestinal and colonic organoids and cooperates with mutant p53 and Apc loss in transformation', *Oncogene*, 39(38), pp. 6053–6070. doi: 10.1038/s41388-020-01414-9.

Reya, T. and Clevers, H. (2005) 'Wnt signalling in stem cells and cancer', *Nature*, 434(7035), pp. 843–850. doi: 10.1038/nature03319.

Rieder, F. *et al.* (2007) 'Wound healing and fibrosis in intestinal disease', *Gut*, 56(1), pp. 130–139. doi: 10.1136/gut.2006.090456.

Roca, H. *et al.* (2013) 'Transcription Factors OVOL1 and OVOL2 Induce the Mesenchymal to Epithelial Transition in Human Cancer', *PLoS ONE*, 8(10), pp. 1–20. doi: 10.1371/journal.pone.0076773.

Rogler, G. *et al.* (2001) 'Differential activation of cytokine secretion in primary human colonic fibroblast/myofibroblast cultures', *Scandinavian Journal of Gastroenterology*, 36(4), pp. 389–398. doi: 10.1080/003655201300051216.

Roy, J. G., McElhaney, J. E. and Verschoor, C. P. (2020) 'Reliable reference genes for the quantification of mRNA in human T-cells and PBMCs stimulated with live influenza virus', *BMC Immunology*, 21(1), pp. 1–7. doi: 10.1186/s12865-020-0334-8.

Rupp, C. *et al.* (2015) 'IGFBP7, a novel tumor stroma marker, with growth-promoting effects in colon cancer through a paracrine tumor-stroma interaction', *Oncogene*, 34(7), pp. 815–825. doi: 10.1038/onc.2014.18.

Sahai, E. *et al.* (2020) 'A framework for advancing our understanding of cancer-associated fibroblasts', *Nature Reviews Cancer*, 20(3), pp. 174–186. doi: 10.1038/s41568-019-0238-1.

Saitoh, M. (2018) 'Involvement of partial EMT in cancer progression', *The Journal of Biochemistry*, 164(4), pp. 257–264. doi: 10.1093/jb/mvy047.

Santamaria, P. G. *et al.* (2017) 'EMT: Present and future in clinical oncology', *Molecular Oncology*, 11(7), pp. 718–738. doi: 10.1002/1878-0261.12091.

Sasaki, N. *et al.* (2016) 'Reg4+ deep crypt secretory cells function as epithelial niche for Lgr5+ stem cells in colon', *Proceedings of the National Academy of Sciences of the United States of America*, 113(37), pp. E5399–E5407. doi: 10.1073/pnas.1607327113.

Satoh, K. *et al.* (2014) 'Tumor budding in colorectal carcinoma assessed by cytokeratin immunostaining and budding areas: Possible involvement of c-Met', *Cancer Science*, 105(11), pp. 1487–1495. doi: 10.1111/cas.12530.

Saxena, K. *et al.* (2020) 'OVOL1/2: Drivers of Epithelial Differentiation in Development, Disease, and Reprogramming', *Cells Tissues Organs*. doi: 10.1159/000511383.

Schimmack, S. *et al.* (2014) 'A mechanistic role for the chromatin modulator, NAP1L1, in pancreatic neuroendocrine neoplasm proliferation and metastases', *Epigenetics and Chromatin*, 7(1), pp. 1–12. doi: 10.1186/1756-8935-7-15.

Schliekelman, M. J. *et al.* (2015) 'Molecular portraits of epithelial, mesenchymal, and hybrid states in lung adenocarcinoma and their relevance to survival', *Cancer Research*, 75(9), pp. 1789–1800. doi: 10.1158/0008-5472.CAN-14-2535.

Serrano-Gomez, S. J., Maziveyi, M. and Alahari, S. K. (2016) 'Regulation of epithelial-mesenchymal transition through epigenetic and post-translational modifications', *Molecular Cancer*, 15(1), pp. 1–14. doi: 10.1186/s12943-016-0502-x.

Shen, J. *et al.* (2016) 'Role of DUSP1/MKP1 in tumorigenesis, tumor progression and therapy', *Cancer Medicine*, 5(8), pp. 2061–2068. doi: 10.1002/cam4.772.

Silen, W. and Ito, S. (1985) 'Mechanisms for rapid re-epithelialization of the gastric mucosal surface', *Annual Review of Physiology*, VOL. 47, pp. 217–229. doi: 10.1146/annurev.physiol.47.1.217.

Sinha, A. *et al.* (2013) 'Abstract B60: Analysis of the endocytic receptor MRC2 as a target for mesenchymal pancreatic cancer cells', pp. B60–B60. doi: 10.1158/1538-7445.tim2013-b60.

Sinicrope, F. A. (2010) 'DNA mismatch repair and adjuvant chemotherapy in sporadic colon cancer', *Nature Reviews Clinical Oncology*, 7(3), pp. 174–177. doi: 10.1038/nrclinonc.2009.235.

Slonim, D. K. and Yanai, I. (2009) 'Getting started in gene expression microarray analysis', *PLoS Computational Biology*, 5(10), pp. 3–6. doi: 10.1371/journal.pcbi.1000543.

Slyper, M. *et al.* (2020) 'A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors', *Nature Medicine*, 26(5), pp. 792–802. doi: 10.1038/s41591-020-0844-1.

Smith, T. G., Robbins, P. A. and Ratcliffe, P. J. (2008) 'The human side of hypoxia-inducible factor', *British Journal of Haematology*, 141(3), pp. 325–334. doi: 10.1111/j.1365-2141.2008.07029.x.

Snippert, H. J. *et al.* (2010) 'Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells', *Cell*, 143(1), pp. 134–144. doi: 10.1016/j.cell.2010.09.016.

Sood, R. *et al.* (1992) 'Regulation of CFTR expression and function during differentiation of intestinal epithelial cells', *EMBO Journal*, 11(7), pp. 2487–2494. doi: 10.1002/j.1460-2075.1992.tb05313.x.

Steffens, F. E. (1983) 'What is principal components analysis?', *Seminar on principal components analysis in the atmospheric and Earth sciences, Pretoria, 1983, (Council for Scientific & Industrial Research, Pretoria, National Programme for Weather, Climate & Atmosphere Research; CSIR-S-334)*, 26(3), pp. 3–16.

Stemmler, M. P. *et al.* (2019) 'Non-redundant functions of EMT-TFs', *Nature Cell Biology*, 21(January), pp. 102–112. Available at: https://doi.org/10.1038/s41556-018-0196-y.

Stoeckius, M. *et al.* (2018) 'Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics', *Genome Biology*, 19(1). doi: 10.1186/s13059-018-1603-1.

Sun, D. Y. *et al.* (2019) 'Cancer-associated fibroblast regulate proliferation and migration of prostate cancer cells through TGF-β signaling pathway', *Life Sciences*, 235(August), p. 116791. doi: 10.1016/j.lfs.2019.116791.

Takahashi, D. *et al.* (2011) 'The epithelia-specific membrane trafficking factor AP-1B controls gut immune homeostasis in mice', *Gastroenterology*, 141(2), pp. 621–632. doi: 10.1053/j.gastro.2011.04.056.

Takaku, K. *et al.* (1998) 'Intestinal tumorigenesis in compound mutant mice of both Dpc4 (Smad4) and Apc genes', *Cell*, 92(5), pp. 645–656. doi: 10.1016/S0092-8674(00)81132-0.

Takatsuna, M. *et al.* (2016) 'Myofibroblast distribution is associated with invasive growth types of colorectal cancer', *Oncology Reports*, 36(6), pp. 3154–3160. doi: 10.3892/or.2016.5202.

Tang, X. *et al.* (2019) 'The single-cell sequencing: New developments and medical applications', *Cell and Bioscience*, 9(1), pp. 1–9. doi: 10.1186/s13578-019-0314-y.

Terragni, J. *et al.* (2011) 'The E-box binding factors max/Mnt, MITF, and USF1 act coordinately with FoxO to regulate expression of proapoptotic and cell cycle control genes by phosphatidylinositol 3-kinase/Akt/glycogen synthase kinase 3 signaling', *Journal of Biological Chemistry*, 286(42), pp. 36215–36227. doi: 10.1074/jbc.M111.246116.

Thellin, O. *et al.* (1999) 'Housekeeping genes as internal standards: Use and limits', *Journal of Biotechnology*, 75(2–3), pp. 291–295. doi: 10.1016/S0168-1656(99)00163-7.

Tian, X. *et al.* (2011) 'E-Cadherin/β-catenin complex and the epithelial barrier', *Journal of Biomedicine and Biotechnology*, 2011. doi: 10.1155/2011/567305.

Tilburgs, T. *et al.* (2017) 'NLRP2 is a suppressor of NF-κB signaling and HLA-C expression in human trophoblasts', *Biology of Reproduction*, 96(4), pp. 831–842. doi: 10.1093/biolre/iox009.

Ting-Yu, C. *et al.* (2020) 'Identification of the potential biomarkers in patients with glioma: A weighted gene co-expression network analysis', *Carcinogenesis*, 41(6), pp. 743–750. doi: 10.1093/carcin/bgz194.

Title, A. C. *et al.* (2018) 'Genetic dissection of the miR-200–Zeb1 axis reveals its

importance in tumor differentiation and invasion', *Nature Communications*, 9(1). doi: 10.1038/s41467-018-07130-z.

Tomaselli, P. J. *et al.* (2017) 'Mutations in noncoding regions of GJB1 are a major cause of X-linked CMT', *Neurology*, 88(15), pp. 1445–1453. doi: 10.1212/WNL.0000000000003819.

Tracy, S., Yuan, G. C. and Dries, R. (2019a) 'RESCUE: Imputing dropout events in single-cell RNA-sequencing data', *BMC Bioinformatics*, 20(1). doi: 10.1186/s12859-019-2977-0.

Tracy, S., Yuan, G. C. and Dries, R. (2019b) 'RESCUE: Imputing dropout events in single-cell RNA-sequencing data', *BMC Bioinformatics*, 20(1), pp. 1–11. doi: 10.1186/s12859-019-2977-0.

Tsujino, T. *et al.* (2007) 'Stromal myofibroblasts predict disease recurrence for colorectal cancer', *Clinical Cancer Research*, 13(7), pp. 2082–2090. doi: 10.1158/1078-0432.CCR-06-2191.

Um, J. M. *et al.* (2012) 'A small molecule inhibitor of Mitf-E-box DNA binding and its depigmenting effect in melan-a cells', *Journal of the European Academy of Dermatology and Venereology*, 26(10), pp. 1291–1297. doi: 10.1111/j.1468-3083.2011.04286.x.

Uribe, D. J. *et al.* (2017) 'The receptor tyrosine kinase AXL promotes migration and invasion in colorectal cancer', *PLoS ONE*, 12(7), pp. 1–16. doi: 10.1371/journal.pone.0179979.

Vu, T. and Datta, P. K. (2017) 'Regulation of EMT in colorectal cancer: A culprit in metastasis', *Cancers*, 9(12), pp. 1–22. doi: 10.3390/cancers9120171.

Wallace, J. L. and Granger, D. N. (1996) 'The cellular and molecular basis of gastric mucosal defense', *The FASEB Journal*, 10(7), pp. 731–740. doi: 10.1096/fasebj.10.7.8635690.

Walther, A. *et al.* (2009) 'Genetic prognostic and predictive markers in colorectal cancer', *Nature Reviews Cancer*, 9(7), pp. 489–499. doi: 10.1038/nrc2645.

Walz, H. A. *et al.* (2007) 'β-cell PDE3B regulates Ca2+-stimulated exocytosis of insulin', *Cellular Signalling*, 19(7), pp. 1505–1513. doi: 10.1016/j.cellsig.2007.01.030.

Wang, J. *et al.* (2009) 'The bimodality index: A criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data', *Cancer Informatics*, 7, pp. 199–216. doi: 10.4137/cin.s2846.

Wang, Y. J. *et al.* (2019) 'Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues', *bioRxiv*. doi: 10.1101/541433.

Wang, Z. *et al.* (2010) 'The Role of Notch Signaling Pathway in Epithelial-Mesenchymal Transition (EMT) During Development and Tumor Aggressiveness', *Current Drug Targets*, 11(6), pp. 745–751. doi: 10.2174/138945010791170860.

Weidenfeld, K. and Barkan, D. (2018) 'EMT and stemness in tumor dormancy and outgrowth: Are they intertwined processes?', *Frontiers in Oncology*, 8(SEP), pp. 1–6. doi: 10.3389/fonc.2018.00381.

West, L. E. and Gozani, O. (2011) 'Regulation of p53 function by lysine methylation', *Epigenomics*, 3(3), pp. 361–369. doi: 10.2217/epi.11.21.

Wilding, J. L. *et al.* (2010) 'Replication error deficient and proficient colorectal cancer gene expression differences caused by 3'UTR polyT sequence deletions', *Proceedings of the National Academy of Sciences of the United States of America*, 107(49), pp. 21058–21063. doi: 10.1073/pnas.1015604107.

Wolf, F. A., Angerer, P. and Theis, F. J. (2018) 'SCANPY: Large-scale single-cell gene expression data analysis', *Genome Biology*, 19(1). doi: 10.1186/s13059-017-1382-0.

Wong, N. A. C. S. *et al.* (2004) 'Loss of CDX1 expression in colorectal carcinoma: Promoter methylation, mutation, and loss of heterozygosity analyses of 37 cell lines', *Proceedings of the National Academy of Sciences of the United States of America*, 101(2), pp. 574–579. doi: 10.1073/pnas.0307190101.

Wu, C. *et al.* (2017) 'Hedgehog signaling pathway in colorectal cancer: Function, mechanism, and therapy', *OncoTargets and Therapy*, 10, pp. 3249–3259. doi: 10.2147/OTT.S139639.

Wu, S. *et al.* (2020) 'Cellular, transcriptomic and isoform heterogeneity of breast cancer cell line revealed by full-length single-cell RNA sequencing', *Computational and Structural Biotechnology Journal*, 18, pp. 676–685. doi: 10.1016/j.csbj.2020.03.005.

Xu, J., Lamouille, S. and Derynck, R. (2009) 'TGF-B-induced epithelial to mesenchymal transition', *Cell Research*, 19(2), pp. 156–172. doi: 10.1038/cr.2009.5.

Xu, Y. and Pasche, B. (2007) 'TGF-β signaling alterations and susceptibility to colorectal cancer', *Human Molecular Genetics*, 16(R1), pp. 1–13. doi: 10.1093/hmg/ddl486.

Yang, L. and Zhang, Y. (2017) 'Tumor-associated macrophages: from basic research to clinical application', *Journal of hematology & oncology*, 10(1), p. 58. doi: 10.1186/s13045-017-0430-2.

Yates, F. (1934) 'Contingency Tables Involving Small Numbers and the χ 2 Test', *Supplement to the Journal of the Royal Statistical Society*, 1(2), p. 217. doi: 10.2307/2983604.

Yeung, T. M. *et al.* (2013) 'Myofibroblast activation in colorectal cancer lymph node metastases', *British Journal of Cancer*, 108(10), pp. 2106–2115. doi: 10.1038/bjc.2013.209.

Young, M. D. and Behjati, S. (2018) 'SoupX removes ambient RNA contamination from droplet based single-cell RNA sequencing data', *bioRxiv*. doi: 10.1101/303727.

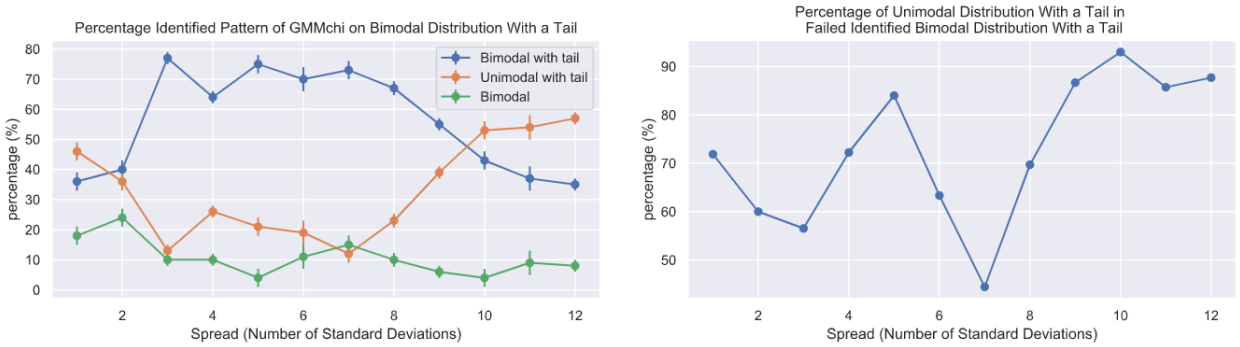
Zhang, J., Tian, X.-J. and Xing, J. (2016) 'Signal Transduction Pathways of EMT Induced by TGF-β, SHH, and WNT and Their Crosstalks', *Journal of Clinical Medicine*, 5(4), p. 41. doi: 10.3390/jcm5040041.

Zhang, W. *et al.* (2015) 'HIF-1α promotes epithelial-mesenchymal transition and metastasis through direct regulation of ZEB1 in colorectal cancer', *PLoS ONE*, 10(6), pp. 1–16. doi: 10.1371/journal.pone.0129603.

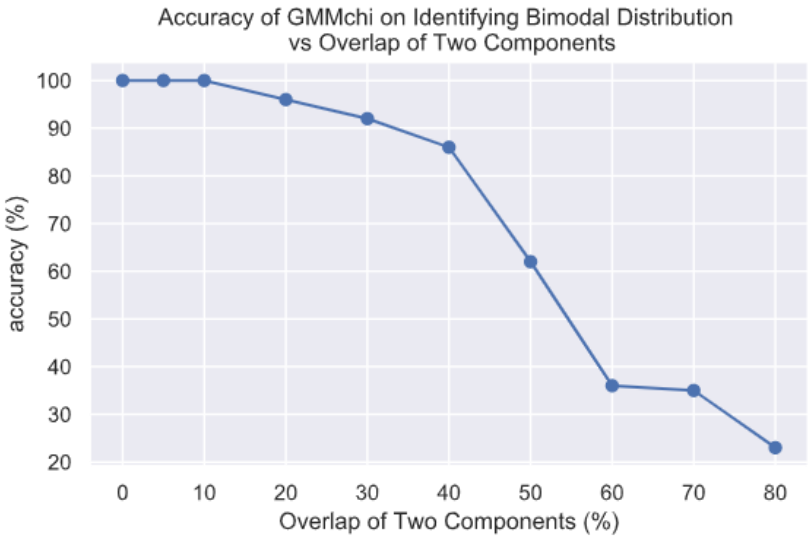
Zhou, S. *et al.* (2020) 'The role of epithelial-mesenchymal transition in regulating radioresistance', *Critical Reviews in Oncology/Hematology*, 150(150), p. 102961. doi: 10.1016/j.critrevonc.2020.102961.

Zhu, J. *et al.* (2018) 'HIF-1α promotes ZEB1 expression and EMT in a human bladder cancer lung metastasis animal model', *Oncology Letters*, 15(3), pp. 3482–3489. doi: 10.3892/ol.2018.7764.

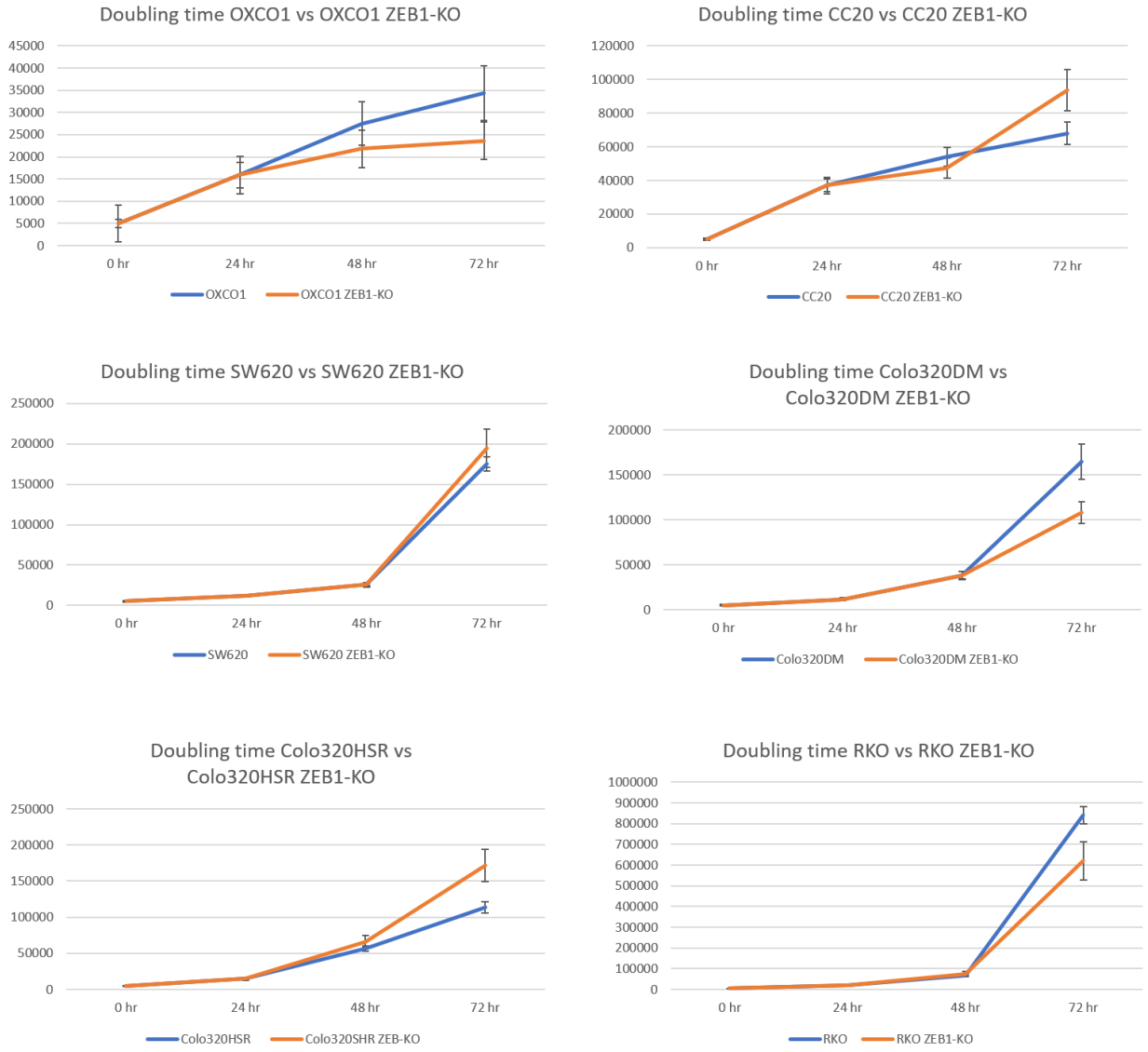
Appendix



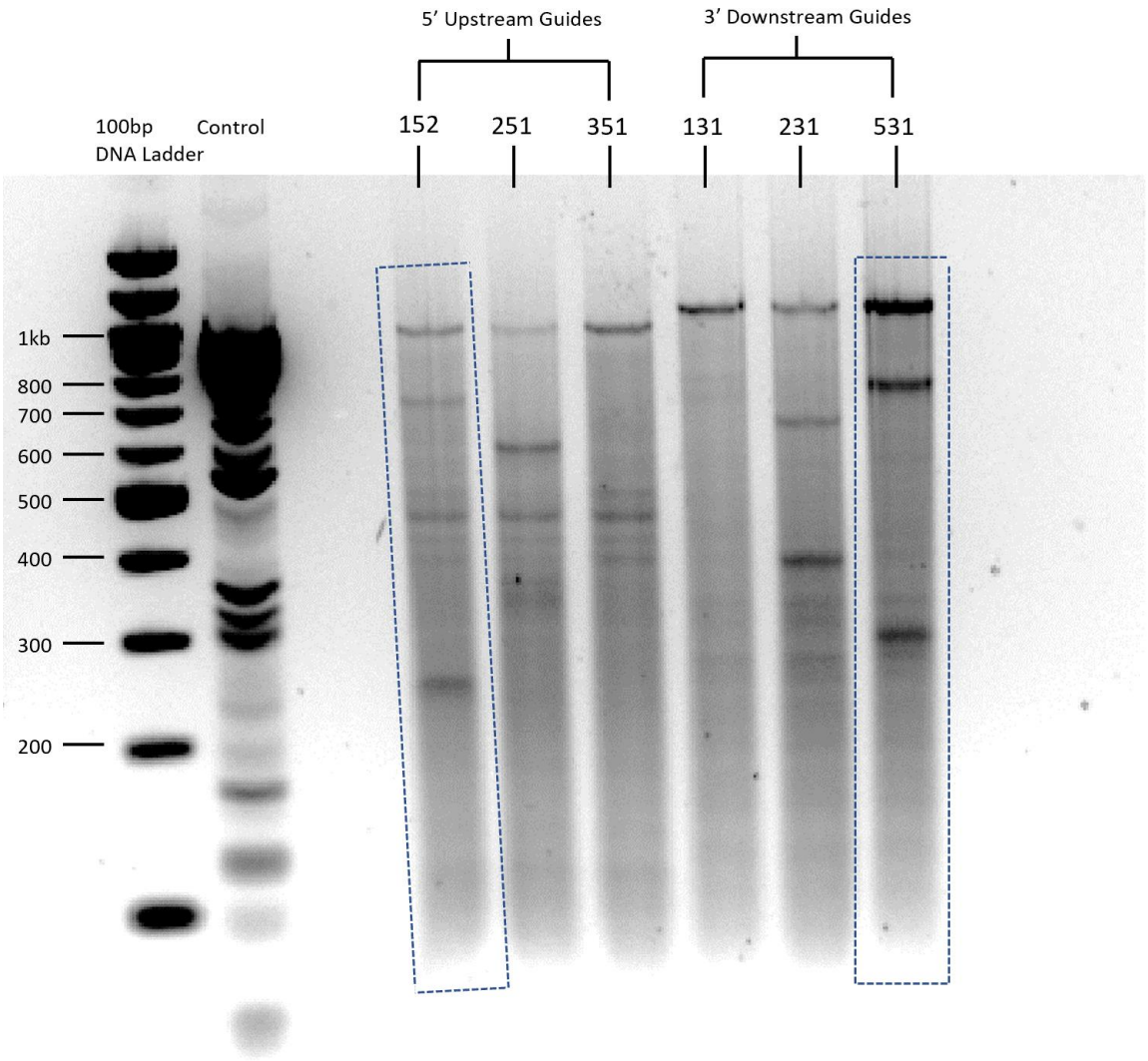
Supplementary Figure S1 Graph Showing GMMchi's Accuracy of Each Category: The left distribution shows the percentage of each returned pattern from the GMMchi on 5000 simulated bimodal distribution with a tail. The right plot shows the percentage of unimodal distribution with a tail identified in the failed identification of bimodal distribution with a tail, demonstrating that the widespread phenomenon squishes bimodal distribution into unimodal distribution due to the static number of bins representing the same number of sample data.



Supplementary Figure S2 Accuracy of GMMchi Based on the Overlap Between Two Normal Components: The plot reveals an increase in accuracy of GMMchi in identifying bimodal distribution as the overlap between the two components overlap decreases.



Supplementary Figure S3 Doubling Time Assay of ZEB1-KO vs Wildtype: Shows the scratch assay of ZEB1-KO vs control on EMT cell lines RKO, SW620, OXCO1, and CC20. Brightfield images are taken every 24 hours for 2 days. OXCO1 and SW620 which have the most drastic reversal of EMT phenotypes with no ZEB1 expression upregulated CDH1 and downregulated VIM shows a significant slow down in wound closure. RKO and CC20, which have a re-expression in ZEB1 and little change in CDH1 and VIM expression confirmed via immunofluorescence, showed little difference in the speed of wound closure when compared to their wildtype control.



Supplementary Figure S4 Surveyor Assay to Determine which Guides to Use for CRISPR Knockout: Surveyor assay is performed to ascertain the best pair of guides used for excising the critical exon. The result shows that 152 and 531 (boxed) gives the expected bands suggesting it is cutting correctly and has a similar efficiency in cutting thus chosen as the pair for subsequent CRISPR-KO experiments.

Supplementary Table S1 List of All Genes Significantly Associated with Methylation: Table showing the list of genes found to be significantly associated with methylation of CpG clusters near the promoter region of the genes.

Gene	p-value	r	CpG_cluster_id	CpG_sites_hg19 (chromosome: location)
ADAM23--2	2.72E-193	-1.00	ADAM23_2	2:207308081 to 2:207308465
AGT	4.22E-05	-0.69	AGT_1	1:230848675 to 1:230848702
ANK2--1	3.12E-07	-0.80	ANK2_4	4:113627479 to 4:113627503
AQP6--2	4.22E-05	-0.69	AQP6_2	12:50367487
ARHGAP6--1	4.22E-05	-0.69	ARHGAP6_5	X:11445222 to X:11445738
ARHGEF1--1	7.65E-06	-0.74	ARHGEF1_4	19:42386936 to 19:42386962
ATP6V0A4	4.22E-05	-0.69	ATP6V0A4_6	7:138458827 to 7:138458847
	4.22E-05	-0.69	BOC_2	3:112930438 to 3:112931299
BOC--2	4.22E-05	-0.69	BOC_3	3:112931429 to 3:112931434
C11orf96	2.72E-193	-1.00	C11orf96_2	11:43963707 to 11:43964930
C14orf132_weighted	4.22E-05	-0.69	C14orf132_2	14:96505281 to 14:96506628
CACNA2D4_weighted	4.22E-05	-0.69	CACNA2D4_3	12:2027354 to 12:2028356
CAMK4_weighted	4.94E-205	-1.00	CAMK4_2	5:110559798 to 5:110560377
CAPN9--1	2.72E-193	-1.00	CAPN9_2	1:230883259 to 1:230883707
CCR10	4.94E-205	-1.00	CCR10_3	17:40832215 to 17:40835778
CDH1_weighted	7.02E-05	-0.68	CDH1_3	16:68772281 to 16:68772331
CDYL--2	4.22E-05	0.69	CDYL_8	6:4775722 to 6:4776188
CHGA	4.22E-05	-0.69	CHGA_3	14:93389472 to 14:93389717
CLP1--3	4.94E-205	1.00	CLP1_1	11:57423148
CNPY4--2	4.22E-05	-0.69	CNPY4_3	7:99717840; 7:99717841; 7:99718005
COLEC11	4.22E-05	-0.69	COLEC11_6	2:3653231 to 2:3654619
CPEB2--3	4.22E-05	-0.69	CPEB2	4:15003069 to 4:15006223
CSGALNACT1--2	4.22E-05	-0.69	CSGALNACT1_1	8:19614225 to 8:19615098
DLX1--1	4.22E-05	-0.69	DLX1_2	2:172949263 to 2:172949988
DTX3--1	4.94E-205	-1.00	DTX3_3	12:57998582 to 12:57999054
EPB41L3_weighted	4.22E-05	-0.69	EPB41L3_2	18:5542879 to 18:5543700
FAM107B_weighted	4.22E-05	0.69	FAM107B_1	10:14814938 to 10:14817317
FOXL1--2	4.22E-05	-0.69	FOXL1_2	16:86609390 to 16:86610658
FRMD4A--6	4.22E-05	0.69	FRMD4A_3	10:14374079 to 10:14374096
GABRB2--3	4.22E-05	-0.69	GABRB2_1	5:160974560 to 5:160976113
GABRB3--4	5.94E-06	-0.74	GABRB3_5	15:26874330 to 15:26874350
GAS2	4.22E-05	-0.69	GAS2_2	11:22688481 to 11:22688509
GLI3--1	2.72E-193	-1.00	GLI3_2	7:42275523 to 7:42277722
GNAQ_weighted	4.22E-05	0.69	GNAQ_2	9:80645442 to 9:80647503
GNG7--2	4.22E-05	-0.69	GNG7_2	19:2702563 to 19:2702905
GOLGA7B	5.94E-06	-0.74	GOLGA7B_1	10:99608564
HECW2--1	4.94E-205	-1.00	HECW2_1	2:197457425 to 2:197458333
HEYL--2	4.22E-05	-0.69	HEYL_3	1:40105508 to 1:40105680
ITGBL1_weighted	7.02E-05	-0.68	ITGBL1_2	13:102105986 to 13:102106351
KCNH3	3.32E-197	-1.00	KCNH3_3	12:49933165 to 12:49933332
KCNJ3--3	4.94E-205	-1.00	KCNJ3_1	2:155554095 to 2:155554961

KDM5D	4.05E-201	1.00	KDM5D	Y:21906412 to Y:21907084
KLK4--1	4.22E-05	-0.69	KLK4_1	19:51412375 to 19:51412392
	4.22E-05	-0.69	KRT15_1	17:39674572 to 17:39674590; 17:39676664
KRT15	4.22E-05	-0.69	KRT15_2	17:39676172 to 17:39676419; 17:39676420
KRT16	4.22E-05	-0.69	KRT16_1	17:39767441 to 17:39767456
KRT17_weighted	4.22E-05	-0.69	KRT17_1	17:39780837
	4.22E-05	-0.69	KRT9_1	17:39728293
KRT9	4.22E-05	-0.69	KRT9_3	17:39728426 to 17:39728436
KSR1--4	4.22E-05	-0.69	KSR1_2	17:25798498 to 17:25798741
LAPTM5_weighted	4.22E-05	-0.69	LAPTM5_2	1:31229781 to 1:31230119
LYPD3	4.22E-05	-0.69	LYPD3_2	19:43969374 to 19:43969650
MMP14_weighted	2.72E-193	-1.00	MMP14_3	14:23307031 to 14:23307405
MMP14--3	2.72E-193	-1.00	MMP14_3	14:23307031 to 14:23307405
MMP28_weighted	4.22E-05	-0.69	MMP28_2	17:34121828 to 17:34122170
	4.22E-05	-0.69	MMP28_3	17:34122104 to 17:34122819
MMP28--1	3.32E-197	-1.00	MMP28_2	17:34121828 to 17:34122170
	3.32E-197	-1.00	MMP28_3	17:34122104 to 17:34122819
MOCS1--3	6.03E-196	-1.00	MOCS1_4	6:39901874 to 6:39901897
MYL3	4.22E-05	-0.69	MYL3_1	3:46903236
NCAM1_weighted	2.72E-193	-1.00	NCAM1_2	11:112831915 to 11:112832815
NR0B2	4.22E-05	-0.69	NR0B2_2	1:27240093 to 1:27240094
NUDT16L1	4.22E-05	0.69	NUDT16L1_1	16:4741032 to 16:4745392
	4.22E-05	0.69	NUDT16L1_2	16:4741476 to 16:4743052
	4.22E-05	0.69	NUDT16L1_3	16:4742988 to 16:4744433
OLFML3	7.02E-05	-0.68	OLFML3	1:114523774
OSBPL10--1	4.22E-05	-0.69	OSBPL10_3	3:32024278 to 3:32025392
PAX5--2	4.22E-05	-0.69	PAX5_3	9:37035056 to 9:37035850
PDLIM3--1	2.72E-193	-1.00	PDLIM3_2	4:186456032 to 4:186456801
PGGT1B--2	4.22E-05	0.69	PGGT1B_2	5:114600869 to 5:114601107
PHYHIPL	4.94E-205	-1.00	PHYHIPL_3	10:60936913 to 10:60936968
PI4K2A--3	4.22E-05	0.69	PI4K2A_2	10:99401176 to 10:99401181
PIWIL1	4.22E-05	-0.69	PIWIL1_3	12:130822210 to 12:130823312
PKD1L2--2	4.22E-05	-0.69	PKD1L2_1	16:81213279 to 16:81213293
PRKACB--3	4.22E-05	-0.69	PRKACB	1:84543431 to 1:84544261
PRSS12_weighted	4.22E-05	-0.69	PRSS12_1	4:119273341 to 4:119274742
	4.22E-05	-0.69	PTN_4	7:136952920 to 7:136953058
PTN_weighted	6.03E-196	-1.00	PTN_6	7:136953074 to 7:136953081
QSER1--2	2.72E-193	1.00	QSER1	11:32913998 to 11:32915702
RAP1GAP--1	4.22E-05	-0.69	RAP1GAP_4	1:21994865 to 1:21997199
RASL11B--1	2.72E-193	-1.00	RASL11B_1	4:53728088 to 4:53728768
RENBP	4.22E-05	-0.69	RENBP_2	X:153209990 to X:153210140
	4.22E-05	-0.69	RHBDL2_1	1:39407394 to 1:39407625
RHBDL2--1	4.22E-05	-0.69	RHBDL2_2	1:39407567 to 1:39407652
ROGDI	4.22E-05	-0.69	ROGDI_3	16:4851634 to 16:4853105
RPRM	2.72E-193	-1.00	RPRM_1	2:154333397 to 2:154333776
RPS20_weighted	4.22E-05	0.69	RPS20_3	8:56987900 to 8:56987912
SCG3	7.02E-05	-0.68	SCG3_2	15:51973921 to 15:51973945

SH3YL1--2	1.31E-06	-0.77	SH3YL1_1	2:263509 to 2:263595
SLC16A7--1	4.22E-05	-0.69	SLC16A7_2	12:59990437 to 12:59990591
SLC7A9	4.22E-05	-0.69	SLC7A9_2	19:33359747 to 19:33360545
SMAD9--1	4.94E-205	-1.00	SMAD9_1	13:37493569 to 13:37494564
SPARC--1	6.03E-196	-1.00	SPARC_3	5:151066598 to 5:151066599
SPTBN1--3	1.03E-05	-0.73	SPTBN1_3	2:54786597 to 2:54787464
STK4--5	4.22E-05	0.69	STK4_1	20:43594931 to 20:43595650
SUMF1	4.22E-05	0.69	SUMF1_1	3:4507763 to 3:4508092
SV2A	4.22E-05	-0.69	SV2A_3	1:149888710 to 1:149889388
TNFRSF10C_weighted	2.72E-193	-1.00	TNFRSF10C_2	8:22960568 to 8:22961190
	2.72E-193	-1.00	TNFRSF10C_3	8:22960617 to 8:22960859
TRIL_weighted	3.32E-197	-1.00	TRIL_1	7:28996031 to 7:28997885
TRIM2_weighted	2.08E-05	-0.71	TRIM2_2	4:154073424
ZADH2--7	4.22E-05	-0.69	ZADH2_4	18:72916694 to 18:72917153
ZBTB7C--1	4.22E-05	-0.69	ZBTB7C_4	18:45662577 to 18:45665951
	4.22E-05	-0.69	ZBTB7C_6	18:45664469 to 18:45664546
ZEB1--4	1.89E-07	-0.81	ZEB1_1	10:31606675 to 10:31606814

Supplementary Table S2 Full List of Genes Assigned to Each Gene Cluster

	Genes Associated That Are Not Inclusions	Genes Associated That Are Inclusions
Gene Cluster1	USH1C_weighted, AKR1C3, MYO15B_weighted, VIL1_weighted, TMPRSS2_weighted, MUC3A--1, RIPK3, ADD3_weighted, TFEB--2, CAPN5--1, LOC100134229, C20orf112--1, PIWIL4, PRR5L--2, SPATA13--1, LOC100505806 /// SNORD123	SGK2--1, DDC_weighted, ARSE, CYP2B6 /// CYP2B7P1, IHH--2, MYO15B--1, TM4SF5, A1CF, PPP1R14D, PLEKHA6_weighted, GPX2--2, CLDN3--2, KCNQ1--1, CYP3A5_weighted, AIFM3, nan--9565, TRIM31_weighted, KIAA1161--2, ABLIM2--3, C17orf73, LOC157860
Gene Cluster2	TMEM139--2, KCNK5, GOLT1A	RAB11FIP4_weighted, CYP2J2, MUC20_weighted, C2orf55--2, LOC158402, C1orf226, LPAR5--1
Gene Cluster3	PIK3C2B, MUC13_weighted, LAD1--1, ASCL2--2, CYP3A7--2	GPR35, EPS8L3, DOK4_weighted, ABP1, GABRE, TRIM10--3, C17orf76, GCOM1, ACOT11--1, CYP3A4--2, TRIM15--2
Gene Cluster4	LGALS4, nan--5651, ALDH2, AKR1C2_weighted, GPX2--1, SEMA4G--2, CDH17, CLDN3--1, FLJ46875	CLRN3, HEPH--2, FOXA3, BAIAP2L2--2, CALML4_weighted, AMACR--1, CDX1, ATP10B--1, GPA33, FABP1--1, MYO1A, GUCY2C, MYO7B--3, WNK4--1, SPPL2B--4, nan--9448, C10orf99_weighted, LOC151009, CDHR1--2, nan--9921, AOAH, nan--9239, GGT6, LGALS3--1, PDCD4--5
Gene Cluster5	MLLT3--2, FGFR4_weighted	EPHX2, CDX2, HSD11B2, nan--11227, PPFIBP2--1, SELENBP1--1, PAQR8--2, PKDCC
Gene Cluster8	BTG2--1, OVOL1--1, C10orf95, nan--6103	STARD5, FZD5--1
Gene Cluster9	nan--3489, nan--6657	
Gene Cluster12	GJB1, CHDH_weighted, LOC400573--2, TNFRSF1B, LOC100506781, TMEM229B--2	ATP2C2_weighted, ZFP36L2--2, CBLC--1, SLC12A2_weighted, ETS2_weighted, AGMAT--2, nan--5065, CDC42EP5, IL22RA1, ELL3 /// SERINC4--2, MYO5B--2
Gene Cluster13	ATP8B1--2, BCL2L14--3, PPP1R1B, C2orf89, LOC100507192, C9orf152, CEACAM5--1, TDGF1 /// TDGF3, TSEN2--2, FUT2_weighted, KIAA1199--2, CEACAM6_weighted, AGR2_weighted, FRAT1, ACSL5_weighted, VILL, PDZK1IP1_weighted, LOC113230, RALGAPA2--4, nan--8691, FAAH2	PRR15L, ERBB3_weighted, HNMT_weighted, TSPAN8
Gene Cluster14	IDH2--1, ATP8A1--3	NPNT--1, ABCC6--2, PAK1--2, FGFR3--1, PCSK9, AIM1, PAK1--4, FAM135A--2, TTC39A--1
Gene Cluster15	FAM108C1, UGT1A1 /// UGT1A10 /// UGT1A3 /// UGT1A4 /// UGT1A5 /// UGT1A6 /// UGT1A7 /// UGT1A8 /// UGT1A9--3, PIP5K1B--1, POF1B_weighted, JHDM1D--2, FA2H_weighted, ARHGAP32--4, FUT3_weighted, SPTBN1--3	SLC44A4_weighted, GATA6--1, KIAA1199--1, LGALS3--2, SULT1C2_weighted, CCRL2

Gene Cluster16	NOSTRIN, AUTS2--1, PLEKHH1_weighted, SAMD13, CKMT1A /// CKMT1B, BRP44--2, VDR_weighted, TLN2_weighted, LOC100289410, AKR1B10, AGR3, MCF2L_weighted, LOC100288092, TP53I11--2, KBTBD11, MGST1--4, SLC35D1--2, GIPC2_weighted, SLC46A3, CISD3--2	PAPSS2_weighted, MYB--1, HNF4G--2, ST6GALNAC1, IRF8, HOOK1--1
Gene Cluster17	ABCB1 /// ABCB4, CA12_weighted, CLDN2, CHPT1--3, NSMCE4A--4	ABCB1--1, PCK1, ANKS4B, NOX1_weighted, APOBEC1, FAM105A_weighted, ANKRD43, nan--6130, CALML4--3, CCL14-CCL15 /// CCL15, CFTR_weighted, IL33, PTPRO--2, OLFM4
Gene Cluster18	nan--10901, nan--10289, LRRCC1, HNMT--6, PPARGC1B--1, MOBKL2B_weighted, ZNF789_weighted, STX3--3	PIK3R1--3, ZNF789--2
Gene Cluster19	PIGZ, VAV3_weighted, PRUNE2--2, SLC3A1--1, LRRC31, DAPK2_weighted	FABP1--2, SGK2--2, PEX11A--2, LRRC19, CTTNBP2
Gene Cluster20	HMGCS2--1, TMEM176A, BIN2, TMEM176B--1, PBLD--2, PLCH1--1, C11orf93--1	MAP3K13--2, nan--8749, HMGCS2--2, NIPAL1--2
Gene Cluster21	MECOM_weighted, nan--7291, FLJ33630--1	ARL14, ACSM3_weighted, TEP1--2, PPARA_weighted, ISX, nan--7576, PXMP4--3, nan--5475
Gene Cluster22	DEPDC6--1, nan--5058, LRIG1--1, nan--11136, C18orf18	ZSWIM5, FAM84A_weighted, LOC653602, PPP1R13B
Gene Cluster23	LYZ_weighted, nan--983, ANXA13, SPINK4, FCGBP, SEC16B--4, PELI1_weighted, ENTPD2--2, nan--8884, DEGS2, BDKRB2	TM4SF4, DNAJC22_weighted, CREB3L1--1, MUC2, UGT1A8 /// UGT1A9, MUC17--1
Gene Cluster24	RNF128, HGD_weighted, CHN2--3, COBLL1_weighted, TINAG--2, C13orf31--3	AZGP1_weighted, CEACAM1--2, MEP1A, LOC400099, FLJ22763, CYP39A1--3
Gene Cluster25	TOX3_weighted, AMACR /// C1QTNF3_weighted, CEACAM1_weighted, ARHGAP26--4, SLC40A1--1	FOXA2_weighted, nan--3808, TFF3, SLC44A3, RSPH1, TST, LOC100130933, LOC100128893, SLC5A1_weighted, nan--5400, SH3PXD2A--2
Gene Cluster27	LOC253264, CCDC113, nan--10201	NOS3, FHIT
Gene Cluster28	UGT1A6, AHCYL2--1, AGPHD1, C2orf82_weighted, CASZ1--5	UGT1A1 /// UGT1A10 /// UGT1A3 /// UGT1A4 /// UGT1A5 /// UGT1A6 /// UGT1A7 /// UGT1A8 /// UGT1A9_weighted, UGT1A1 /// UGT1A10 /// UGT1A4 /// UGT1A6 /// UGT1A8 /// UGT1A9_weighted, AHCYL2--2, C15orf38
Gene Cluster32	PRAP1, NR1I2_weighted, C13orf18_weighted, ACOX2, DPEP1, nan--9357, CADPS--2, HOXB9--2, SLCO2B1_weighted, nan--5783, CELP, TLCD2--2, CADPS--4	SLC39A5--1, SLC22A18, CCL14 /// CCL14-CCL15, CYP2B6--1, LY6G6D, IL2RG, GTF2IRD1, C19orf77, IFITM1_weighted, ABCC6--3, SPINK1, CEL /// LOC100508206_weighted, KIF9--1
Gene Cluster33	TSNARE1--2, TRAPPC9--1, ARHGAP39	
Gene Cluster34	LMO7--2, ARSD--1	A2LD1--1, PSD4--2, CABLES1--2

Gene Cluster35	nan--1578, MGAT3--2, nan--4845, CIDEC--2, CYP39A1--2	RNF186, nan--8323, PMFBP1
Gene Cluster36	S100A14, SYBU, CYP2S1	GSTO2, HR--1, TRIM15--1, SERINC5, DGAT2--1
Gene Cluster37	SH2D2A, CENPV--1, LOC100505644, TESC--1, NMNAT3--1, KIAA1274, CEBPA, SNHG5, METTL7B, F12--1, LOC100134361, JPH1--2, VSNL1--1, PRR5_weighted, KDM5C--2	LOC645249, BTNL9--3, CGREF1, LOC440894
Gene Cluster38	TCF7--1, SLC7A8--4, CASK--5, SYT7--4	
Gene Cluster39	DEPDC7, ESPN--2, GRIN2D--2, C20orf54--2	LOC100506860--3, NR1H3
Gene Cluster40	ARID5B--3, ZNF703--1, NEBL--1, SYT7--3, HPDL, C14orf128, DUSP16--3, ND2, TMEM92--2	RNF43, TMEM117
Gene Cluster43	RBM47_weighted, SSBP3_weighted, CD24_weighted, DDAH1--2, KIAA1543_weighted, ANKRD5--1, PHLPP1--2, FAM110A, PCSK6--2, nan--11656, FGD3--2	CCDC88C--2, PARM1--2, HOOK1--2, LIMK2--1, nan--8701, PNKD--3
Gene Cluster44	BCL11A_weighted_2, TCEA3, SNTB1_weighted	PPARGC1B--4, RNF157--2, SAMD5_weighted, GLS2--4, KLK1, RPP25--1, GSTA4--1
Gene Cluster45	HOXA13_weighted, nan--8719, RIMKLA--1, CYP27B1	nan--11125, ICOSLG--1
Gene Cluster46	PRSS16, MARVELD3--2, nan--5999, ERCC6 /// PGBD3--2, MPZL3--3, DSC2--2	STX3--2, LPAR2--1
Gene Cluster47	CLDN7, TMEM125, GALNT3--1, PKP3_weighted, ST14_weighted, nan--8774, RAB17, LAD1--2, F2RL1--2, KRT8--1, CNKSR1, KRT19--1, SLPI, OVOL2--2	TJP2--1, KRT18, JUP, ITGB4--1
Gene Cluster48	AP1M2_weighted, OVOL1--2, RAB25, IRF6_weighted, DSC2--3, GRTP1--2, EPCAM, MAP7_weighted, BSPRY_weighted, ESRP1_weighted, MYCL1--1, TRAF3IP2--2, CCDC6--2, TMEM159	F11R--3, GRHL2, TMEM30B--1, KRTCAP3, MAL2, C1orf172
Gene Cluster49	PPM1H--2, PKP2--1, LYN_weighted, SLC27A3, RNASET2--3, MBNL3--2	TSPAN13, PLEKHG4, CDS1_weighted, HOOK2, EPB41L4B--4, F11R--1, SMARCA4_weighted_2, BAZ1A--1
Gene Cluster51	nan--2401, XK	FAAH, ARHGAP8 /// PRR5-ARHGAP8_weighted, CFD, nan--4857, ESRP2--2, GPR160, nan--5235, HR--3
Gene Cluster52	IRF2BP2--1, nan--10740, ICA1_weighted, COBL--2, DAAM1--1, C11orf9--1, ID2_weighted, GCNT1--2, PTPRK--1, HTATIP2_weighted, nan--5446, NGEF--1, CRIP1	
Gene Cluster53	MEST, SOX13--3	CLDN4--2, DUSP23, CCDC64--2, LOC151534, LIPG, SUSD1, FBXL16--1, PRRG2, GPR56_weighted

Gene Cluster54	MYO1D--2, VIPR1, ARHGAP27--2, CXXC5_weighted, PTPN6, KLHDC9, SLC43A2, ZNF513, ALDH3A1, HSPA1A /// HSPA1B--2	TBC1D16--5, LDLR--2
Gene Cluster55	PYCARD, HRCT1	SULT2B1, nan--10248, B3GALT4, nan--3159, RFFL--6, LOC643008
Gene Cluster56	C20orf134, nan--5685	OSBPL5--1, PRSS1 /// PRSS2 /// PRSS3, HKDC1_weighted, KIF21B, PRSS3_weighted, FOXP4--2
Gene Cluster60	SPINT1_weighted, CGN_weighted, EPDR1, nan--5072, RAB11FIP1--1, TNK2--4, YPEL3--1	RNLS--2, ABCD1--1
Gene Cluster61	LGALS9, UBA7--2, BATF2, CFB--1, HLA-C_weighted, HCG26, MUC1_weighted, EHF--5, IL32	HCP5, NLRC5, IDUA--2, IFI27
Gene Cluster63	MACC1_weighted, LOC100506130, nan--5334, STAMBPL1--2, LOC153546	LIPH--1, PTK6--2, BAG1--3, MANSC1, MBOAT1, C9orf95--2
Gene Cluster65	KRT20, LY75, TRIM2_weighted, GDA--2, ANKRD22_weighted, MYO5B--3, FABP6--3, nan--4623, TTC22--4, SH3RF2--3, SERPINB1_weighted, GALNT12_weighted, IL20RA_weighted	C15orf48, LOC100505633--2, C9orf150, RASEF_weighted, SLC6A20_weighted
Gene Cluster68	KRCC1_weighted, CAMK2N1_weighted, GANC--3	NRIP1--2, SEMA3C--2, MYO6_weighted, nan--5011, MAP3K5_weighted, TBC1D8B--2, NRIP1--1, PELI1--1, LONP2--2
Gene Cluster69	LOC283278, nan--9149, FGD4--1, SLC35D2--3, EEPD1--2	AKR1C1--3, C1orf106, nan--4013, nan--714, LLGL2--2, LOC100506098, C20orf54--1, SERINC2, ESRP2--1, RAB20--1
Gene Cluster70	MCTP2--2, S100P, OAS1_weighted, PLXNA2--2, ARSD_weighted, ADCK5	STARD10_weighted, PPARG, nan--6025, KLF5_weighted, FOS, TJP3--1, ERBB3--2
Gene Cluster71	LNX1--1, MGAT4A--3, FBP1, FXYD3--2, C1orf210, nan--10745	ELF3_weighted, DTX4, PRR15, TMEM45B_weighted, ELMO3, TNFSF13--1, LOC440335, nan--7559, SH3YL1--2, FBLIM1_weighted_2, DENND2D, TMC4, MYO5C
Gene Cluster72	LMO7--4, MYH14_weighted_2, ABHD2_weighted, CAPN8, nan--7439, ITGB6--3, TAOK3--1, TMC5_weighted, RPH3AL--1, COX19--2, ILDR1--2, LOC730102--2	ELF3--2, FAM83E, LIPH--2, VPS13C--4, DUSP8--1, LRP10--3, TRAK1--1, ZNF518A--3, nan--6402
Gene Cluster74	LOC728153--1, PPP6R2--1, nan--4663, ZC3H7B--3, LOC340085	
Gene Cluster75	nan--10751, nan--10757, nan--10780, nan--6157, nan--11014, MBNL1--8, MKLN1--1	nan--6653, nan--6062, nan--8711, nan--964, nan--814
Gene Cluster76	nan--8941, nan--4891, nan--11868, RUFY2--3, nan--6343, FBXL3--2, PIK3C2A--9, LOC646014	nan--8893, UGGT2--4, nan--1219, ZNF207--8, nan--4485
Gene Cluster77	nan--8797, nan--233, LOC100216546--1, nan--11233, nan--6720, nan--9117, nan--12057, SYK--4, EPB41L4B_weighted, nan--6395, nan--2684, nan--8796, AGXT2L2--2	ZCCHC7--3, nan--7060

Gene Cluster78	nan--8417, nan--3968	RREB1--7, nan--10626, nan--249, SCARB2--3, nan--6129, NCRNA00107, nan--9220
Gene Cluster82	nan--10141, nan--4030, nan--4888, nan--9861	CFDP1--2, nan--6162, C5orf56--4, nan--10694, nan--7376, nan--4664
Gene Cluster84	nan--5283, nan--1110, nan--5604, ARMC8--6, nan--9440, nan--6106, nan--4979, nan--6354, nan--12050, nan--5329, QRSL1--3, nan--8530, nan--11200, nan--7298, nan--1163, nan--8523, nan--8939, nan--10952, nan--7027	nan--7888, nan--9285, nan--11875, nan--6513, nan--11218, nan--9184
Gene Cluster85	PRPF38B--2, C2orf67--4, nan--6082, PDCD6--2, nan--5482, SOX6--8, nan--11286, ZNF224--1	nan--8945, nan--5892
Gene Cluster87	EML4--6, nan--4494, nan--6667, nan--11075	ZNF638--5, nan--10604, nan--10627, nan--8788, nan--504, FGFR1--5, nan--12111, DHX36--1, NPIPL3--1, nan--9460, nan--7541, FNDC3B--7, KIAA0368--4
Gene Cluster88	nan--8208, nan--5385, nan--5644, PBX4, nan--10578, TMEM219--2, WDR1--5	nan--8535, nan--11187, S100A6--2, nan--10478
Gene Cluster89	MICAL2--4, BCAS1--1, TRIM66_weighted	MUC20--3, MUC20--5, nan--8795, nan--10923, nan--9014
Gene Cluster91	JMJD1C--3, CFLAR--13, nan--10646, PCSK7--3	NT5C2--2, SPATS2L--4, LOC100131015
Gene Cluster92	nan--11905, nan--6026, nan--8829, nan--8925, nan--8862, nan--6560, nan--9024, RHOBTB3--6	LOC100131541, nan--8820, nan--8724, CD47--7, LRRFIP1--9, SORL1--3
Gene Cluster93	nan--8819, nan--10918, nan--3674, nan--6019, nan--11089, nan--4895, NF1--9, EXOC3--2	nan--7593, nan--11144, nan--3879, nan--7545, nan--6641
Gene Cluster94	nan--9229, nan--10476, nan--6371, nan--10798, DNAJC7--1	nan--10482, PMS2L2--1
Gene Cluster95	POU5F1P3, nan--9219, YARS--2, nan--10562, LOC100129637--4	
Gene Cluster98	nan--7620, nan--5128, nan--5127	
Gene Cluster105	SFXN3--1, nan--6209	C14orf106--3, nan--5197, nan--10910
Gene Cluster106	nan--10497, nan--9411, USP24--3	nan--7839, nan--7139
Gene Cluster110	MPZL1--4, nan--6259	
Gene Cluster114	nan--8890, LOC100129122--1	nan--9540, PLA2G4F
Gene Cluster115	NAPEPLD_weighted, CES2_weighted, TFDP2--4	B4GALNT3--3, nan--5483, MCTP2--4, HSDL2--4, SENP2--3

Gene Cluster116	NFIB--8, C1orf56--1, TMEM8B--1, nan--5170	nan--5695, LOC100507110
Gene Cluster118	nan--9019, nan--8752, KPNA1--3, ARHGEF7--2	DHRS12--1, RAB11FIP4--3
Gene Cluster122	GPSM2--4, SMYD2--1, LOC100505946, DAB2_weighted_1, nan--10444, PLXNA2--3	PLCB4--3, LOC339290_weighted, C11orf53, C11orf93--2, nan--9515, MYLIP--5, SLC25A27--4, FAM55A--1, nan--5360
Gene Cluster125	nan--6065, C20orf94, nan--5769, POLQ--1, nan--12073, CXCL14_weighted, nan--6501, LOC100288730	MSI2--6, nan--8997, nan--11300, SP3--4, nan--9183, nan--11931, RIMBP2--1
Gene Cluster133	nan--10533, nan--2381, nan--12043, TRIM2--4, nan--11054, nan--12117, nan--9523, nan--5656	nan--8922, nan--744, nan--11623, nan--2602, TTC17--5, nan--5214
Gene Cluster134	TOX4--5, ZNF277--3	
Gene Cluster135	ATP10B--2, nan--6672, nan--3135, nan--10286	nan--947, nan--8048, nan--9112, MECOM--1, nan--9042, nan--6163, nan--6044, nan--5476, nan--6297, nan--7757, nan--6391
Gene Cluster137	nan--7391, ZFAND6_weighted, nan--9268	nan--9497, C7orf54, nan--660, nan--546, PSME4--4, nan--12154, nan--11215, nan--6132, nan--7775, RRN3 /// RRN3P1 /// RRN3P2--2, nan--4951, nan--6145
Gene Cluster139	NCRNA00263, LOC100130987, nan--9178, CYCS--3	nan--6012, nan--8254, nan--10819, AGAP4, CTBP2--6, nan--226, nan--4099, nan--155, ZFP41--2, LMO7--3, nan--6081
Gene Cluster140	nan--11873, ZNF721--1, nan--10244, nan--7680, nan--5578, C4orf10--1, nan--2624, nan--8144, nan--10799, nan--5676, nan--10886, nan--661, nan--11605, nan--7747, KIAA1267 /// LOC100294337--3	TRAPPC10--4, nan--952, nan--8938, SFRS15--5, CCDC76--2, nan--8998, nan--9519, nan--5678, nan--7911
Gene Cluster142	C7orf28B, nan--884, nan--7347, ARHGEF1--1, nan--2490, nan--6569, nan--9058	NSF--2, nan--615, nan--4203
Gene Cluster145	nan--4207, nan--3263	
Gene Cluster147	nan--7316, nan--4017, nan--4113	nan--794, nan--6093, nan--7660, nan--5704
Gene Cluster148	PSMB7--2, FNBP1--4, CALU--5, nan--7523	nan--2462, nan--1698
Gene Cluster150	nan--4837, nan--998, BRD8--3, UBE2Z--3, nan--8503, nan--1878	nan--597, nan--207, nan--3975, HM13--2, CDC14B--6
Gene Cluster151	nan--11498, nan--9673, SAMD3--2, nan--7116, nan--11801, nan--4117, NAIP, C17orf104--5, nan--1216, nan--651, APBB2--6, C7orf41--2, nan--7537, nan--2749, nan--11353, SPPL3--2	nan--8072, nan--11281
Gene Cluster152	SFTA2, nan--11107, PARVA--2	RNF213_weighted, FLJ45340--2, nan--9638, MYOF--3, ZMYND8--5, CNIH4--2, LOC100505656, DOCK5--3, nan--8324, nan--

		944, WWC1--5, nan--3205, nan--11339, C5orf56--1, nan--7944
Gene Cluster153	ZNF254--1, nan--7411, USP34--5, HKR1--3, nan--707, PTPRF--4	SPTAN1--2, nan--8745
Gene Cluster154	FLJ44342--2, nan--6077, nan--3182, nan--10822, LOC728061	FAHD2A /// LOC729234--1, nan--8234, HOXA6
Gene Cluster155	nan--817, nan--6251, nan--5776	
Gene Cluster157	MACF1--7, nan--9312, nan--11426	
Gene Cluster158	C7orf53_weighted, nan--11092, KANK1--3, TNFRSF25--2, KRT38, nan--2945, TBX19, CEP290--3, FOXN1_weighted, NES--2, nan--1454	
Gene Cluster159	nan--3072, nan--10560	PGC--2, nan--6190, SYNPO--1, nan--8704
Gene Cluster160	nan--2041, CDC42BPA--4, nan--7356, nan--10882, nan--11581, SYNPO--3	nan--4222, GDA--1, KRTAP4-1, KRT34
Gene Cluster161	nan--11160, nan--10468, nan--9401, nan--8134, nan--6840, nan--6668, KIAA1609--3, nan--5777, nan--11357, MIR30C2, DLC1--3, CLDN9, PLA2G4C--2, RELN, BST1, nan--665, PHLDB2--3, nan--456, nan--5698	
Gene Cluster162	nan--10579	nan--348, TNKS1BP1, KIAA1407, nan--153, MFI2--2
Gene Cluster164	LOC100506897, SRGAP2--2	HLA-DOB--2, FLJ39051_weighted, nan--6621, nan--6031, nan--8591, RLIM--3, ZNF346--3
Gene Cluster167	LOC401522, ITGBL1_weighted, RGS7_weighted, nan--9572, ZBP1--2	nan--6345, RTP4, C8orf60, SLC15A3
Gene Cluster170	BYSL, UBR2--2, TAF8_weighted, MED20--2, FRS3, GRM8_weighted	MRPS10--1, NT5C--1, PPP4R1L--3, RIPK2--1
Gene Cluster171	nan--6160, nan--6562, nan--6620, nan--6856, nan--8751, nan--8845, LOC401320--5, USP49--4, nan--11749, POLH_weighted_2, CNPY3_weighted, C6orf108_weighted, POLR1C_weighted, CA4_weighted, MRPS18A_weighted, KLHDC3_weighted, TJAP1_weighted, nan--6064, nan--5699, AARS2--2, FLJ38717--2, nan--217, nan--572, VPS8--1, SCML4--5, TMPRSS2--1, nan--3633, MAD2L1BP, C6orf223--1, SLC13A2--1, CDC5L--2, ABCC10--2, XPO5--2, KIAA1267 /// LOC100294337--2, SLC35B2, LRCH3--4, nan--5359, AQP1--1, SUN1--1, TLCD2--1, nan--8963, nan--9599	

Gene Cluster172	nan--11262, nan--9715, CREB5--3, SULT2A1_weighted, TMEM63A--4, AK7--1, SRF--2, CD14	EGF, KIAA1267 /// LOC100294337--1, AP3M2--1
Gene Cluster173	IL6R--1, nan--2678	nan--5609, LOC650794--2, C21orf58--4
Gene Cluster175	KDM6B_weighted, TTLL3_weighted, FBXO9_weighted_1, SLC35E4--1, MBD1--4, CARS2--2, AQP1--2	TMEM43--3, nan--10838, FOSB, SULT1C2--4, LOC284801_weighted, PIP, MST1_weighted
Gene Cluster183	ATP2A3_weighted, nan--8519, LOC644246--2, PLCB1--2, ZFP3, C7orf41--1	nan--9274, CREBZF--6
Gene Cluster186	nan--351, SOS1--5, KIAA1841, nan--7929, RAPH1--5, nan--3940	
Gene Cluster187	C17orf42--2, nan--8914	nan--10469, CBR4--1, nan--11173, nan--470, FLJ31306--2
Gene Cluster192	ZNF493--1, ZNF605--2, ZNF493 /// ZNF738	
Gene Cluster197	nan--7465, nan--10440, BCL2L11--3	nan--12125, nan--7461, LOC100289333, LOC100132832, U2AF1--2
Gene Cluster202	nan--3499, nan--12172, nan--7389, SH2D3A--2, nan--3760, nan--11977, FARP2--3	nan--9041, nan--6806, nan--5928, nan--11001, nan--11271, nan--255, STAG3L1--1
Gene Cluster203	CUL9--1, CNPY3--4	CC2D1A--3, nan--11018, ZNF664--1, nan--573
Gene Cluster205	nan--6728, OR7E126P, nan--6463, ITPK1-AS--2, LOC595101	RPAIN--1, nan--5580, RNF34--3, nan--1897, OSBPL10--1, MYH14_weighted_1
Gene Cluster206	PPIA--6, DLX6-AS, ATG16L1--1, nan--2992, CASP1_weighted	DEFB1, nan--7989, nan--2390, nan--2601, CEACAM7--1, nan--4778, nan--11480, SNX16--2, PPIL4 /// ZC3H12D, nan--11181, GRM8--2, KIAA1984, nan--11828, ARHGAP44--2
Gene Cluster210	nan--5767, EPT1--2	LOC100127972, NUMA1--3, nan--1956, HIBADH--3
Gene Cluster212	nan--4849	C16orf55--1, nan--6033, nan--9407, nan--1475, PSMG4--4, nan--227, nan--10496, GSDMB--1, nan--11028, PCBD2--3
Gene Cluster214	nan--11714, nan--6167, ZBTB20--2, nan--8442, nan--2345, nan--3905, SLC25A16--3, PBRM1--6	nan--3580, nan--9696, nan--6690, nan--11112, nan--10378, POLR2J4_weighted, nan--5198, nan--6684, NUDT13
Gene Cluster216	CTTNBP2NL--3, STAMBP--2	PRKCE--2, ZNF561--2
Gene Cluster218	ARHGEF10--2, nan--476, CDC42BPB--1, FBXW4P1, nan--6365	nan--2695, ZNF502--1, nan--10834
Gene Cluster219	nan--7279, nan--8638, nan--8148, nan--5565, SSBP1--1	
Gene Cluster220	nan--5766, nan--8508, nan--5426, nan--10809, RORC--2, nan--7431, nan--10682, nan--10656	nan--8923, nan--7387, nan--10922

Gene Cluster222	NYNRIN, RAB3A, RNF208	SLC29A4, DAK--2, RCOR2, CPNE6--2, NEU4, GOLIM4--1, MT3, NAALADL1--2, nan--6498, NPC1L1--1, SPN_weighted
Gene Cluster223	nan--9628, ALDH1L1--2, QRICH2	SPATA8, KLK4--1, LTB4R--3, CYP2W1, LOC100134445 /// WASH1 /// WASH2P /// WASH3P /// WASH5P /// WASH7P_weighted, ZNF169--1, SMOX--1, nan--8967, ABCB8--2, LOC146880--1, CACNA1A--1
Gene Cluster224	nan--8068, MCOLN3_weighted_1, CACNG6--1, KCP--3, DIO3-OS_weighted, LOC283352--2, SLC2A9--1, FMO2--2, GAS7--2	STAT2_weighted, TNXA /// TNXB--2, nan--9650
Gene Cluster225	PI3_weighted, nan--3545, SFTPB--3, nan--115, NR4A1--3, PDXK--1	PML_weighted, RRP1--2, nan--10650
Gene Cluster226	nan--9898, GPR156--2, nan--7318, ARID5B--4, TMEM106A--3, ZNF71--4, PTCHD2--1, nan--5277, ABI3, nan--4675, GEMIN7--2, SYT2, LDB3--2, LOC100294020 /// LOC100507755 /// POM121L10P /// POM121L1P /// POM121L4P /// POM121L8P /// POM121L9P, ABHD11--2, LRRC43--3, DEFB124--2, MYOZ3--3, LOC730139, nan--1640, LOC388906, DNAJC4--1, CD300LG--3, MAPK12--1, PTH2, SLC5A9--2, UBXN10--1, MATK, ALDOB_weighted, FLJ20444	
Gene Cluster228	MEGF6--1, FOXJ1, SLC12A9--2, FLJ90757--1, OBP2A--3	MAP3K10, POFUT2--1, ARHGDIA--1, DYRK1A--3, ADAMTSL5, MAP2K7_weighted
Gene Cluster229	MOBKL2C--2, ZNF821--2, AFMID	
Gene Cluster230	SIK1--2, KIAA0664--2	
Gene Cluster231	JPH2--1, SFXN5--4, MAP6--2	
Gene Cluster232	CCDC120--1, NS3BP, C22orf26	
Gene Cluster233	TBKBP1, CD4--1, BCL2L1--4	
Gene Cluster234	nan--8520, PRSS36	
Gene Cluster237	nan--6931, GTSF1L--1	
Gene Cluster241	ZNF358--2, AGRN--1	
Gene Cluster248	SETDB1--2, DOM3Z--1	
Gene Cluster254	nan--9812, nan--2352	

Gene Cluster262	C17orf56--3, DYRK1B--2	C19orf6--1, JOSD2, GOLGA8A--1, TNK2--3, SPEN--1, INF2--2, RIN3--2
Gene Cluster264	PCGF2--1, TAF4--1, ATN1--2	SLX1A /// SLX1B /// SULT1A3 /// SULT1A4, TMEM184A--1, CC2D1A--1, NOTCH1--2, PGS1--2
Gene Cluster265	RAD23A--1, nan--3868	DOC2A, GATAD2B--1, SLC25A42--2, HPS4--1
Gene Cluster267	ATXN2L--1, KHSRP--3, FBRS--2, DMWD_weighted	nan--1380, REXO1--1, nan--6774, nan--6995, THRA--2, UBE3A--6
Gene Cluster268	ZBTB7B_weighted, RAB11B--2, DDA1--3, DCAF8--7, BCL3--1, VAMP2--1, HR--2	FKBP8_weighted, HDAC10--1, ATN1--4
Gene Cluster269	ARHGEF11--2, MZF1--1	
Gene Cluster272	LOC220980, PHLDB1--3	
Gene Cluster278	nan--617, nan--10854	
Gene Cluster279	nan--11574, nan--7806	
Gene Cluster280	nan--8405, nan--8543	
Gene Cluster281	nan--1411, CNTN2--1	
Gene Cluster285	GPR173--4, DUX1	
Gene Cluster286	IGSF9, ARHGAP8--2, MARVELD3--1, TET3--2	PVRL4, CAMK2G--3
Gene Cluster289	ATAT1--2, CISH--3	TBC1D16--4, CCDC71
Gene Cluster290	OSBPL7--2, CLDN15--1, LIPE--1, ATAT1--1, FAHD2A /// LOC285014--1	
Gene Cluster291	nan--4403, nan--297, ZNF524	
Gene Cluster293	nan--7103, FAM120A--2	NCRNA00174, ATP2A2--4, CASR--1
Gene Cluster294	nan--4764, nan--3659	
Gene Cluster295	nan--3519, nan--3044	
Gene Cluster296	LOC100287166, nan--9873, nan--4790	

Gene Cluster300	nan--10619, nan--5958	
Gene Cluster307	HCN3, ACSS1--2	PCYT2--2, LOC157627--1, C19orf42--4, URB1--1
Gene Cluster313	TRAPPC2L--2, nan--4344	
Gene Cluster322	SOLH--2, ATF6B--2	
Gene Cluster326	PRB3, C9orf98--2	
Gene Cluster328	KRT8--2, NTRK3--9	
Gene Cluster331	TMPRSS6--1, GLP1R--1	
Gene Cluster333	PROCA1--2, nan--9567, nan--9008	
Gene Cluster336	LOC100293559, GPR27--1	
Gene Cluster341	COMMD6--2, nan--7304	
Gene Cluster343	ZFHX2--1, PLEK--1	
Gene Cluster345	AQP7, nan--8741, PHPT1--3	
Gene Cluster348	TDRG1, COL9A2--2	
Gene Cluster353	ZSWIM3, CBFA2T2--3	nan--9326, SLC40A1--2
Gene Cluster358	nan--8340, DNAJB13	
Gene Cluster359	GOLGA6L1 /// GOLGA6L6 /// LOC100132025, BICD1--5	
Gene Cluster360	C11orf16, HSPA1L--2	
Gene Cluster368	LRRC25	CTSS--1, SLAMF6, RELA--3
Gene Cluster375	MAZ--3, CNP--1	
Gene Cluster379	nan--8761, PAFAH2--2, PAIP2B, FAM13A_weighted	LGR4--1, PEX1--1, nan--7115, nan--10573, VPS13B--4

Gene Cluster381	RBM7--2, THOC7--2	nan--7474, NDUFS1--4
Gene Cluster383	LOC100128822, MAN1A1--2, TACC2--3, PRKAR2A--5	ATP2B1--2, ANKH--5, MGRN1
Gene Cluster386	STRBP--3, nan--3917	
Gene Cluster389	nan--121, HCG18--4, WDR27--2	nan--120, MRPS25--3, FZD3_weighted, FAM134B_weighted, ADAT2--1
Gene Cluster391	COX16--1, SLC44A1--4, STK38L--1, ATP6V0A2--5	HIGD1A--3, nan--3860, LEPROT--3
Gene Cluster392	SORD--3, PDCD2_weighted, NEDD4L_weighted, nan--7468, MAN2A2--2, CYCS--2, nan--5150, LARS--1	
Gene Cluster395	nan--5180, MRPS14--2, CLN5_weighted, NAA38--4, LOC100506728, LOC286052	PL-5283--3, PAPOLG--3, nan--555, PDS5B--3, nan--7534, nan--5599, ARGLU1--3, RAP1A--4, TPD52--4, GLT8D1--3
Gene Cluster399	NFIA--3, nan--6048, NFIA--5, FAM59A--1, nan--10235	NFIA_weighted, ZDHHC11_weighted, PODXL2, ZDHHC11--1
Gene Cluster402	REG4_weighted, TFF1, SLC9A2, AKAP7--1, PDE3B--2	ATOH1, CLCA1
Gene Cluster403	RNASE4_weighted, ANG, GALNT4 /// POC1B, SMAD6--1	C2orf72_weighted, ALDH1A1, PTGER4--2, TSPAN7
Gene Cluster404	CLYBL--1, CNOT6L--4, CAMK2D_weighted, FOXP1_weighted_1	NUDT12--2, NUDT7--3, RPL13--5, PRKX, RPS27L--2
Gene Cluster406	RBM3--2, APEX2 /// LOC100507005, DTWD1--3	
Gene Cluster407	TMCC3_weighted, TLE4--1, CLMN_weighted, GCNT1--1, KLF7_weighted, GAB1--6, PLEKHG1, PTPN13--1, EIF4E3--2, TMTC2_weighted	ARID4A--2, RPS6KA6--4, KLF7--3, EIF4E3--3, ATP7A--2
Gene Cluster408	DDAH1--3, TMEM57--2, HMGCL, HBA1 /// HBA2_weighted, BTNL9--2, NRCAM--1	HNF4G--1, TNIK_weighted, MAML3--3, C13orf31--1, ST6GAL1--3, NAA38--2
Gene Cluster410	COX4I1--3, ATP5H--1	NDUFB8--2, UBQLN4--1, nan--3904, SIAE--4, ARAP2--3, nan--3937
Gene Cluster411	LOC400931--2, LOC285949	
Gene Cluster413	NFIB_weighted_1, TMEM144--4	TSPO, TXNL4A--1
Gene Cluster419	SARM1 /// TMEM199--2, nan--5135, NCRNA00182--1, GAPDHP62, RASGRF2--1	MOBKL2B--2, DCAF6 /// ND4, SLC22A23--3
Gene Cluster421	CASP2--5, PLCB4_weighted, nan--786, SGSH--1, C1orf213, SLC29A2--1, LOC442075, LOC100133660	nan--12168, CRIM1--1

Gene Cluster425	SLC6A7, C1orf159	
Gene Cluster428	nan--10681, nan--10855	ZFYVE16--1, nan--2344, SLC19A1--2
Gene Cluster429	SLC1A4--5, nan--7673, LOC100506171, nan--6798, nan--5996, G3BP1--4, C6orf208, nan--2376, nan--4707, NMUR2, nan--3212, nan--838, SERP2_weighted, nan--5257	
Gene Cluster430	KLK4--2, SLC29A2--2, nan--2667, nan--3111, CECR1, nan--5683, DENND1B--6, nan--11536, nan--7077, nan--7081, nan--9442	
Gene Cluster431	FREM1--1, XPNPEP1--3, CREB3L2--2, C1orf230, MYH7B--2, FGF20	TMEM180, nan--7451, nan--5473, LOC100129463, INTS6--4, PPP2R2C--3, nan--8311, CPA6 /// CPB1, PDS5B--5, TOR1AIP2--3
Gene Cluster432	ACPP--1, TMPRSS13--1, ARHGAP26--2, nan--6036	nan--1883, LOC400128, CACNA1D--3, nan--7981, ACPP--2, C9orf123--1, MTMR9LP, nan--11222, LOC641467--2, OR51E1, nan--6468, STARD13_weighted, C19orf59, C9orf68--2
Gene Cluster433	KLK12--2, nan--12174	nan--9291, KLK12_weighted, DLL1--1, EMP2--4, NPL--2, ATP6V1C2--2, ZBTB42, EPHB3_weighted, VPS13A_weighted
Gene Cluster434	PPP1R3E_weighted, GK_weighted, GK /// GK3P, PDK4--3, ONECUT2_weighted, RGS16_weighted, nan--5646, nan--7766, nan--5910	ZNF213, nan--7256, nan--3195, LOC338799--2, PCDH19, nan--90
Gene Cluster436	KIF12, GNG4--2, GNG4_weighted, MGC44328, RASA4	KCNH2--1, nan--8586, TMEM171_weighted, TPP1--2, HUNK--1, OTC, RRBP1--4, KLK4--3
Gene Cluster438	HOXA11-AS_weighted, NR0B2, C19orf45--2, nan--5492, CYP2C9_weighted_2	LOC646627--2, MYOM3_weighted, CHST5_weighted, nan--5581, WNK2_weighted, LOC100289632--1
Gene Cluster439	MAOA_weighted, TP53TG1_weighted, EPN3_weighted	METTL7A--1, TSPAN12_weighted, METTL7A--2, DOCK11--1, LOC100506966, FMO5--2, TP53TG1--2, H2AFJ_weighted_1, RNASE1--1, BLNK, GPCPD1--2
Gene Cluster440	LRP4, nan--3823, LOC100505730 /// LOC100509599, nan--5503, BCL11A_weighted_1, BCAN--5, LOC100508176--1, TMEM200A, nan--9158, TNFRSF25_weighted	nan--12166, PRLR--5, nan--3824, PTPRD--3, EHHADH, C2CD4A, CDHR1--1, CHP2, KIAA1324_weighted, TMEM63C
Gene Cluster442	SAMD5--1, ITPR2--3, HAAO, ACVR1C--1, TUBAL3, CYB561D2	ITPR2_weighted, nan--9603, nan--4328, nan--11930, nan--11929
Gene Cluster446	CLIP2, DNMBP--2, BCAN--3, nan--5431, PRR5L--1, nan--8988, FAT1, nan--3014, nan--11347, C22orf9--3, TRIM4--1, GJB2	INPP5D--3, BCAN--4
Gene Cluster450	PRR7, KLHL9--2, TSPAN33, DIAPH2--1	PIK3AP1--2, SATB2--1, BIVM--5, KCNAB2--1, PDZD8--3, WBSCR27, HOXB9--1, ANKRD16--2, GK--4, PTER--1, ZCCHC14--3
Gene Cluster454	NOP16--1, RPL7L1, MSX1--1	C12orf75, KIAA0564--2, XPO4--1, GCFC1--1, XPO4--2, SLC16A1_weighted, TSC22D1--2, SARS--2, DEF6_weighted

Gene Cluster456	LRP8--1, CCDC41--1, CCNH, MYEOV--1	CCDC85B--2, FABP5, ESD--3
Gene Cluster457	RPS6_weighted, AIFM2--2	PLK1S1--3, RPL18--1, DIS3--1
Gene Cluster461	LIMD1--1, HUWE1--1, SLC7A1--4, ST6GALNAC4--1	NUP98--2, WBP4--1
Gene Cluster465	AIM1L--1, nan--4313, TTC9--2, SYTL1, FTH1P5, nan--9763	
Gene Cluster466	KIAA0040--1, EPPK1_weighted, CDH3, PWWP2B--1, FUT1, FAM83F, FAM89A, RTKN--1, POLM	TEAD3, PLCD3--1, KCNK1_weighted, ZC3HAV1L--2
Gene Cluster468	SLC12A8, TFRC--2	TBC1D4_weighted, TPD52L1_weighted, ZNF165, TFB2M--4, CAMKK1--2, nan--9631
Gene Cluster469	SREBF1--1, GSR--2, ATAD3A--2	HTATIP2--3, BZW2, CCDC85C--1, BMP8B--3, HIST1H4C
Gene Cluster470	FBXO9_weighted_2, SRSF6--1, RABL3--3	HEATR7A--3, UGGT2--7
Gene Cluster471	NUP210--2, ACPL2--1	NUP210--3, WDR74--2, GNB1L--3, NUP210--4, GAL--2, EEF1G_weighted
Gene Cluster473	nan--4232, nan--4208	RPL29--2, ACCN2--1, C5orf30, TFDP1--2
Gene Cluster474	nan--11033, DUSP2, LPHN1_weighted, MLKL, ZAK--2, RNF40--3	C9orf140, ACCN2--2, ZBED3--2
Gene Cluster475	TMEM223--1, ZDHHC23, SNHG7--1, TRAF3--1, FH--4	
Gene Cluster476	MPV17L2	PPP6C--1, RWDD2B--1
Gene Cluster478	IGFBP2, ZNF124--1, MIR17HG--1	CDCA7--2, C9orf100--1, LTBP1_weighted, HMGB1_weighted, SFXN4_weighted, CDT1--2, DFFB
Gene Cluster480	ABHD6_weighted, TMEM68--1, nan--5295, RAVER2--2, nan--8714, HINT1--1, FARSA--1, PAK6_weighted, ABHD6--3, IL28RA--2, nan--1830, PASK_weighted	
Gene Cluster484	GOLIM4--2, CLASP2--5, CSRP2BP--1, SSBP1--3	SLC27A5--1, GRSF1--3, HECTD1--5, RSBN1--5, RANGRF--1, MARCKSL1
Gene Cluster487	EXOSC7--2, nan--7946, SP8--1, LOC285965--2, nan--3436, HORMAD1, USP38--3, DERL3_weighted	
Gene Cluster491	nan--8816	PTCHD1--1, nan--3218, RNF217
Gene Cluster492	INA, GPD1--3, PNLIPRP2--1, LOC100506589	

Gene Cluster493	PLA2G2A, CITED1, REG1A, ABCB1--2, RUNDC3B--2, SUPT4H1--2, RASL11A, BZRAP1, DTWD1_weighted	KCNV1, LOC100506125, LOC100289610, ZNF169--2, RUNDC3B--1, LCN15, TDRD1, ADH6--2, nan--2640
Gene Cluster497	FDX1--3, ZBTB39	
Gene Cluster498	CPXM2--2, C1orf88, IPMK--1, nan--8985, SLC10A4, nan--9404, AGAP11--2	
Gene Cluster499	VASH1--2, MOXD1--2, GAS7--5, PLXNC1--3, SORBS2--4, nan--5262, ITM2A_weighted	
Gene Cluster500	LOC644450--1, UST--1, GHR, CD33, OLFM2, PCDHB2, ZNF681--2, HES5	
Gene Cluster502	FBLN5, LRP8--2, APBA2--2, KLHDC8A--2, CD36--5	
Gene Cluster504	FCGR2A--2, ALDH8A1, PRKG1--3, SNX20--2, ANO5, L1TD1--2, nan--9648, nan--11806, CHD9_weighted, OR51E2_weighted, RBL2_weighted	
Gene Cluster506	ABAT_weighted, WFDC2--3, ATL1, ARHGEF9--1, KDM4C--4, LOC404266--1, VWA5A--2, SYCE1L--1	VWA5A--1, nan--7792
Gene Cluster507	KLF9--3, PDZRN3--2	nan--146, KLF9--2, DDX26B
Gene Cluster509	nan--9621	KCNK10--2, SERPINB6--1, C9orf130--2, nan--1162
Gene Cluster510	C8orf76--1, nan--4501, ADC--2, RASAL1--1	
Gene Cluster515	C11orf41--1, CFTR--3, GPR114--2, LGALS2	
Gene Cluster516	GIT2--3, PTPRN2--1, KLF2--1, C12orf66--1	nan--5533, SIPA1L1--1, UBR5--1, RAB2A--6, FAM125B_weighted, SENP6--3, RASA3_weighted, ATAD1--2
Gene Cluster517	PLA2G12B_weighted, MTTP, GPR128, PMS1--2, nan--2631, NAALADL1--1, CABP1--1, CABP1--3, TBX10--2	
Gene Cluster518	LOC100506848, FBXO10--2	SLC38A5, NAGS, TMLHE--2
Gene Cluster519	CYBA--1, TADA2A_weighted, ZNF20 /// ZNF625--1	BIN1_weighted, CTSC--1, CBR1
Gene Cluster520	IGF2BP2--1, HYLS1	HSPBAP1, LOC100310756, TGFBR3--2, SLC6A6--1, TACC1_weighted, TMEM80--1, TGFBR3--1
Gene Cluster521	nan--7702, nan--11803, PCSK6--4, TWIST1, CNNM1, TSIX, ZADH2--7	
Gene Cluster522	SLC35D3, INPP5D--2, nan--8582, C2orf88_weighted	TM4SF20, LPAR5--2, LOC730102--1, CYP2B6--2

Gene Cluster525	EPB41L5_weighted, XIST_weighted, nan--5085	PLSCR4, OBSCN--3, XIST--7, nan--9209, SLC22A15
Gene Cluster526	ORM1, TBRG4--2, SPESP1, VPS53_weighted	
Gene Cluster527	ELOVL5--2, PGPEP1_weighted, C8orf83_weighted	MLF1--2, ARHGAP30--2, MLF1--1, SLC38A10_weighted
Gene Cluster531	TTTY15, TRD@--10, nan--6755, SDPR_weighted	
Gene Cluster532	CAPN12, RPL31 /// TBC1D8, PTP4A3_weighted, nan--11501, CPPED1--3, CAPN13--2	SLC17A9_weighted, APOH--1, nan--7684
Gene Cluster535	SSX1--1, nan--3922, nan--8119	
Gene Cluster537	MYRIP--1, CXorf23, MED14--4, OGG1--1, NAPEPLD--5	GALK2--2, PRUNE2--1
Gene Cluster540	RAD52--2, TINAG--3, nan--10273, nan--5713, CYorf15A--1	
Gene Cluster541	SFTPA2, CALCA_weighted_2, C10orf81_weighted	
Gene Cluster542	nan--10754, CALCOCO2--3, nan--5568, PLIN2--1, NEU1, MR1_weighted	WLS--1, PRKCA--5
Gene Cluster543	SOCS5--2, SUCLG1, C14orf101--2, PRDM4--2, MAP3K1--2	
Gene Cluster544	CTH--2, TUBE1--1, C8orf59_weighted	METAP1D--4, MIER3--4, MEX3C--3
Gene Cluster545	ME1--1, KLF11--2, LRRFIP1--5	EEF1A1--5, nan--5417, SERAC1--2, PEX7--1, TULP4--3, NHLRC2--1, nan--5254, HSP90B1--4, TMEM181, TMED3, RCAN3--2
Gene Cluster549	nan--11251, nan--9255, nan--6849, nan--5740, ID4--4, LRRC36, nan--11756, NCOA2--4, SMARCA4--10, CA1--2, nan--2480, MOBKL2B--1, nan--1401, PTK2--1, NFATC2IP--2, nan--2760, PDXDC1--3	LOC100505875, nan--10978
Gene Cluster557	nan--3256, ONECUT2--1, SOX4--5, nan--6326, KBTBD12, BAAT, SLC17A4--2, C11orf1--4, ADAMTS6--3	
Gene Cluster558	RAPGEF2--5, SLC19A3--2, PTGS1_weighted, GJC2--1, KIAA0485--2, ELF5--2	
Gene Cluster559	nan--433, LOC100509223, DHRS9_weighted, HYAL1, NPAS2_weighted	SPSB1--2, MMP28_weighted, nan--216, GBP1--3, MMP1, GBP1--2, GPT2, CXCL11--1, XAF1--3
Gene Cluster560	LOC254057, EXPH5--1, PLEKHA7, C4orf34--2, STX17--1, ZNF532--1, KDM6A--3, FAM84B--1, BCL6--1, LOC100507239	MPZL2_weighted, CYB5R2, HS3ST1--2, LIMA1--2, MPZL2--1, HSPA2, SNRPA1--4, LSM14B_weighted
Gene Cluster561	WDR70, nan--5771, MTBP--4, DNAJC21--4, nan--8034, PRLR_weighted	PTGER4--1, PTGDR--1

Gene Cluster564	THRB--2, ALOX5--2, FRMD3--3, GREM1_weighted, SMPDL3B--1	PID1--1, PPBP, ALOX5--1, nan--5564, LDHC, C1GALT1--2, GLCCI1_weighted_2, ALOX5--4, ERAP2--5, JSRP1, nan--1030
Gene Cluster569	MME_weighted, PGBD5, nan--793	SLC30A2--2, nan--3919, CKMT2, nan--7776, nan--11072, EIF3C, nan--5232, CYP4F8, LOC286382
Gene Cluster570	LOC100129907, nan--9907, BEST3--3, nan--7761, nan--7029, nan--5707, DNAH10--3, nan--5379, MIAT--1, nan--4939, ZNF385A, PLA1A, nan--4028, XCL1 /// XCL2, ERAL1, C8G, NRL--1, PLIN1, APOM--1, APOC2--1, CARD9, SSX4 /// SSX4B_weighted, PPP1R14A, nan--6937, PART1--1, SSX2 /// SSX2B_weighted	
Gene Cluster571	nan--6491, nan--5841, CCDC88B, SMPX, MAGEA11, SERPINA7, C9orf71, nan--1123	nan--8486, RTN4RL1, LOC283352--1, C6orf97
Gene Cluster572	CES1, DLX5, TMEM37_weighted, APOA2_weighted, PAGE4, ANPEP--1, PRR9, XCL1--2, nan--3893	SLC6A4--1, IZUMO2, PREX1_weighted
Gene Cluster574	DBT--6, NCRNA00245--2, SGK269--1	
Gene Cluster578	APCDD1, NRXN3--4, CAB39L--1	SEMA3F_weighted, CAB39L--2, NRXN3_weighted, IGF2 /// INS-IGF2_weighted
Gene Cluster579	MPP7_weighted, DYNC1I1--2, DISC1 /// TSNAX-DISC1, GAL3ST4, SETD3--2, QRSL1--4, HSD3B1--1, ITGA9--4, L1TD1--1	TBX3--2, TBX3_weighted
Gene Cluster580	ISM2, ABCG2, TRPM7--6, CDADC1--3, GLDC	LOC100289373--2, LASS4, FREM2, SLC16A10--1, ATP6V1C2--1, UPK1A
Gene Cluster581	ATP11A--1, nan--11289	nan--6001, INPP5B_weighted
Gene Cluster586	NOTUM, 6-Sep--3, NRSN2, GRB14, nan--4822, GAMT--2	DNM1--1, GSTA4--2, FAM69B--3, NID1--2, nan--10711
Gene Cluster587	DLX3, LOC100130468, DLX4--1, DIO3, SLC2A8--2, PCP4, JARID2--1	ADRA2C, ZNF300, LOC388796--2, LEPREL2, PLTP, TUT1, WNT11
Gene Cluster591	LOC100506517--2, HDAC5--1	
Gene Cluster597	nan--7221, FAM127A, FAM127B, PHKB--3, MIDN--1, nan--5124	CBLB--2, XYLT2--3, PRKAB2--1
Gene Cluster598	FLJ45244--2, LOC389834--4, CXCL11--2, KIT, SEMA6A_weighted, MBTD1--3, CBS--3	PXK_weighted, THAP9--2, RBBP6--1
Gene Cluster599	C16orf5--2, MAP7D2, KAT2B--1	PSMD7_weighted, MARK1--2, PKIB--2, nan--9198
Gene Cluster601	ITGA3, FOSL1	C19orf33, DOCK5--8, ID1, NMU, CCDC68
Gene Cluster602	S100A2, STEAP1, STEAP2, LYSMD2, SPAG1, RNF207--1, ENOSF1_weighted	

Gene Cluster603	CXCL3, CXCL2--2, CXCL1, CD47--2	
Gene Cluster604	PION--2, TM4SF1_weighted, ITGB8--3, SYNJ2--2, nan--7236, SP110_weighted_1, C19orf66--3	PION--1, ARAP2--2, PARP14--1, TGFA--2, PARP9--1
Gene Cluster605	C2CD2, RNF130--2	PCGF5--3, MED17--2
Gene Cluster606	GALNT10--2, TNFRSF10A--2, LPP--4, ELOVL7, TBC1D9B--1	TRABD--2, GPCPD1--3, LOC728377, EREG_weighted
Gene Cluster609	ZNF862--1, LIF, PARP3, GDAP1--3	PTEN_weighted_1, PTEN /// PTENP1, CUL1--2
Gene Cluster610	LCOR--1, IAH1--1, C3orf58--1, C5orf25	
Gene Cluster613	GTF2IRD2--2, nan--5234, ZNF688--1, ATP1B1--2, GOLGA1--2, ME3--1, OPLAH, DOT1L--2	DAZAP2--3, RELA--1, STX10, AES, MBNL1_weighted, AK3_weighted
Gene Cluster614	SPRED1--2, SLC24A1--1, STK39, CASP4--1, SDSL, ZC3HAV1--3, RAB27A--2, ZBTB41	
Gene Cluster615	SLC25A16--4, C4orf33--1, FBXO41--3, ABHD15	HMGN5--2, CRYBG3, LOC388152 /// LOC727751 /// LOC727849 /// LOC80154, TAP2--1
Gene Cluster616	ALCAM_weighted, ANTXR2--2, GLS--1, KIAA0240--2, STAMBPL1--3, MID1--2, H2AFY--4, TRIM7--2	
Gene Cluster617	SH3KBP1--3, RGNEF--6	HIVEP2--2, TRIM38--3, REPIN1--2, nan--11621
Gene Cluster618	MAML2--2, C1orf116_weighted, INPP1, EPB41L1_weighted, SCNN1A--1, C1RL--1	LAMA3--4, OR7E14P, FOXQ1, nan--7725, EHF_weighted, OR7E47P
Gene Cluster620	SH3RF2--1, PRRG4_weighted, SERPINB5--2, nan--8801, GJB3--3, FAM110C	SORL1--2, HES1_weighted, NAAA--2, LOC100499467--1
Gene Cluster621	EVPL, SYK_weighted, ADAP1_weighted, CXADR--4	MAPK13_weighted, SEL1L3_weighted, SLC22A5, SHROOM3_weighted, PTGFRN_weighted
Gene Cluster623	MFSD6--2, ARHGEF5, MICA, RELB, FAM111A--1	DTX3L, APOL6--3, LGALS3BP, SPRY2, MST1R, MALL, PPAP2C, OCIAD2, TRIM56
Gene Cluster624	nan--5039, nan--8733, MTMR11--2, TNFRSF14, AREG--1, FAM47E /// STBD1, APOL6--1, PLP2, IRAK2--3, C9orf103, PARP10--1	IL4R--1, NIPAL2--3, CPT1A--1
Gene Cluster625	TUBA4A	SLCO4A1--1, SPATS2L--2, MGST1_weighted, PTPRJ--3, GPX1, BACE2_weighted, IL15RA, PSMB9, CD44_weighted_1, PSMB8, EGR1--1, AGFG2--5, GBP3
Gene Cluster626	C1orf93, EPS8, NMI	CD47--6, LFNG--2, ARHGAP32--3
Gene Cluster627	EMILIN2--2, SLC16A5--1	nan--5090, SNX19--4

Gene Cluster628	ATP6V0E1--4, SLC16A3--2, BAIAP2L1--3, CSNK1A1--1, IRF7	ZDHHC5--1, TTC12--3, RDH13--1
Gene Cluster629	MYO10--2, LTBR--1, TACC2--4, CDCP1--2, FAM83H--1, SFN_weighted, ITPR3_weighted, CD9--1, PATL1--1, VANGL1--4	CTNND1--2, CCNO, RASA2--1, PLBD1--1
Gene Cluster630	SPINT2, TPRG1L, SHROOM1--2, ANXA2_weighted, ANXA2P2, TGFBR2--2, TMSB10, TSTD1, SUMF1	
Gene Cluster631	HBS1L--1, ACTB_weighted, METRN--1, HES4, C16orf79	FAM173B--2, PHLDA2--2, TOM1L2--3
Gene Cluster632	TNS3--1, ME1--2, LMNA_weighted, OSBPL10--3, RIN2--1, GRAMD3--1, LAMB1_weighted	S100A10, PDLIM1, nan--8831, VSIG10--3
Gene Cluster633	LSR, MARVELD2--1	
Gene Cluster634	PFKFB2--3, STAP2, SH3BP4, nan--2614, nan--12084, TGIF1--2, CDK20	ETV6_weighted_2, VAMP8, S100A11_weighted, CXADR_weighted
Gene Cluster635	TMEM49, KCNH2--2, PFKP--1, PTPN3--1, IKBKE--2	ZC3HAV1--1, nan--5179, ARPC1B, HOXB7_weighted
Gene Cluster636	GPR39--3, SLC16A5_weighted, nan--5287, POM121	GAS6--2, C15orf52, NNT_weighted, TBC1D2
Gene Cluster638	nan--9470, SLC7A7, RSPH10B2, nan--3887, PTPRE--2, nan--10688, PAFAH2--1	
Gene Cluster639	C1orf59, AIG1--1, CLIC1, MGLL--2, ZFP36L1--2, STAT6--1	KLF4_weighted, MVP, S100A16, FOXO1_weighted
Gene Cluster640	KLHL24--4, YPEL2--2, RBM39--3, TRIOBP--5	GGPS1--1, CLN8--5, LOC100286909
Gene Cluster641	SLC1A7_weighted, SEMA5A--1, CHN2_weighted, ACE2_weighted, CLIC5--2, REN	nan--5653, EDC3--2, C14orf50, MXD1_weighted, CELSR3--1
Gene Cluster643	SH2D3A--1, DDR1--1	EPS8L1_weighted, C6orf132_weighted, COL17A1, nan--7717, nan--11321, KCNK6, nan--7486, N4BP2L2--3, AIG1--3
Gene Cluster644	nan--574, APOL6--4, CLDN23--3, TNFSF15, nan--10919, OCLN--4, BCL2L15_weighted, TMC5--4, CASP10--1, SDR16C5	nan--5608, nan--8854, nan--9877, CLDN1_weighted
Gene Cluster645	KLK3--1, CYB5B--3, ZCCHC6--3, nan--12091, BMP2K--5, LOC728392 /// NLRP1_weighted, TNFSF13B_weighted	BFSP1, SPATA18--1
Gene Cluster648	nan--8517, nan--5966, LTB4R--1, TRAF1--1, SETDB2--5, nan--11890, SEC16B_weighted	nan--8516, PLEKHG3--1, nan--7240, SLC44A1--1
Gene Cluster649	nan--8878, DIS3--2, FASTKD2--3, TYSND1--3, ATP1A1OS--2	LOC100288939_weighted, nan--11082, nan--9686, nan--6281, TRIM13--4, C2orf68--1
Gene Cluster650	PLK1S1_weighted, FAM113B, GOLGA7B, GATM--3, GNA14--1	CYP27A1, CCDC146, MAD1L1, KLK3--2

Gene Cluster652	CATSPER2 /// CATSPER2P1, nan--9378	
Gene Cluster653	WASH1 /// WASH2P /// WASH3P /// WASH7P, LOC344887, ZBTB7A--2, TMED9--3, LOC100507297, LOC401320--4, SH3BP2_weighted	GPATCH2--2, LOC572558--2, ATXN7L1--3
Gene Cluster657	AGXT2L2--3, nan--10557, FLJ41484, LOC339988_weighted, TGOLN2--1	SETDB2_weighted, MARVELD2--2, CDHR5--1, ZNF862--2, nan--128
Gene Cluster658	nan--5547, CST3--2, SARM1 /// TMEM199--1	
Gene Cluster662	nan--5463, OR2A20P /// OR2A9P--1, BEST1--1, UNC5B--1, LOC100507395, CBX4--1, CASP5, SLC9A3	GTPBP2_weighted, nan--5151, SLC6A9, KIAA1908, LOC100505547, ACE--1, MINK1--3, LOC644794--1
Gene Cluster663	LRRC37A3--2, ACE--2, BAIAP2L2--1, CDHR5--3, TNRC6C--2	
Gene Cluster664	nan--9959, FAM116B--1	
Gene Cluster665	nan--11025, nan--7303, LOC100505523, INTS10--3, SLC25A45--2, nan--10151	nan--6790, HNF1A--2, GRB10--4
Gene Cluster668	nan--1004, nan--1953, AP4S1--3, nan--3836, nan--10470, nan--7219, PTPN23--1, ARHGAP5_weighted, SCFD1--2, ACRBP--1, nan--3851	
Gene Cluster670	PGM2L1_weighted, nan--3675	GPR161_weighted, nan--11961, EPSTI1--2, nan--9318, FILIP1L_weighted, C6orf15, nan--10947, nan--2840, CD82
Gene Cluster671	RPS6KA2--3, IKBKG--2, PADI1_weighted, nan--10908, SERPINA1_weighted, LOC100508534	C2orf54, nan--10545
Gene Cluster673	ZBTB20_weighted, nan--3806	nan--3805, nan--5046, nan--5047, nan--5100, nan--5096
Gene Cluster675	SRGAP1_weighted_2, EDN1_weighted, MLL--5	nan--6017, nan--6088, MMP7, ZNF827--1
Gene Cluster677	nan--586, LEPROTL1--1, SMAD2--2, C11orf80--1, CLDN18--1, EPB41L4A--2, PHC3--4, FAM155A--2, SYNE2--5, DLEC1_weighted, BCR_weighted	
Gene Cluster680	CARD11--2, TNFSF10_weighted, GABBR1 /// UBD, RASSF10, NEDD9_weighted, LOC100499467--2	TNFSF10--3, TBXAS1, KLHL3--2
Gene Cluster682	nan--3827, nan--7257	
Gene Cluster683	nan--10664, CARD6, PLA2G16--2, nan--7262, RHBDL2_weighted, RABGAP1L--4	AQP3--2, KIAA1244--2, PON3, FAM198B--2
Gene Cluster684	HIST2H2BE, C11orf52	HIST1H2BI, HIST1H2BG_weighted, HIST1H2BH, nan--7293, nan--5012, HIST1H3H, nan--5588, LOC100507025

Gene Cluster686	ID4--1, ID4--3, ZNF654--3, VLDLR, nan--5294	BACE1_weighted, SESN3--2, nan--7235, SEMA4D--2
Gene Cluster687	EGR2, LY6G6C, ID4--2, PTCH1--4, HIST1H3E--1, SDK1--2, SESN3--3, nan--7608, nan--702, GPR155_weighted	
Gene Cluster689	MSI2_weighted, SHROOM2, MYLK_weighted, TLR4_weighted	NCALD--2, RAB30--3, EPHB6, SLCO2B1--1, nan--542, PTPRN2_weighted
Gene Cluster691	MAP1LC3A--3, KCNH8, EPM2AIP1--3, CPLX1	SHH_weighted, CHI3L1_weighted, nan--9251, MAVS_weighted, KHK--1
Gene Cluster692	KCNJ2, nan--5880, RASSF6_weighted, SIGLEC15, CHURC1--2, NEBL_weighted, SORBS2_weighted, SMOC1--2, FAM50B, TOLLIP_weighted	NTRK2_weighted, FBLIM1_weighted_1, ENPP3--1, GALNT10--3, IRF2, NTRK2--2
Gene Cluster696	IRF6--2, RDH10--1, SEMA3C--1, DSC2--1	
Gene Cluster697	SNTB1--2, CDC37L1--2, GAL3ST1, CYP4F3, RFX3--2, ZKSCAN1--3	LOC283788--4, SLC44A1--6, SULT1B1, nan--9156, REG3A--1
Gene Cluster698	POLG--2, SMPD3--1, KLK15--2, TGFBRAP1_weighted, C2orf49_weighted, FOXP2--7	
Gene Cluster700	DENND4C--3, nan--127, DUOX2, EDA2R, CNGA1	ACVRL1--2, PCDH20--2
Gene Cluster701	nan--10715, nan--11384, nan--9005, APOOL--5, NPEPPS--3, ZMYM2--2, nan--5787	nan--9507, nan--4683, nan--5658, PPP1CB--4, TXNDC9_weighted, NR1D2--1, RSF1--4
Gene Cluster702	CAB39L--3, nan--7315, FLJ35776--1, GLS--2, SHANK2--3	TMEM111--2, nan--5638, MAP3K12--2, IGFBP4
Gene Cluster703	CFLAR_weighted, ZBTB38--2, NAMPT--4, ISY1 /// RAB43, RBM33--1, CSRNP1	nan--7225, ZNF148--3, PPP1R12A--1, INO80D--4, SMAP1, TRIM13--3
Gene Cluster705	SPAG4, nan--9013, TAF9B--5, PDP2, GNL3L--4, PTCD2--2	nan--5630, PCMTD1--2, nan--5190, B4GALT1--5, PPP2R5E--5, LOC100507315 /// PPP2R5C, WTAP--5
Gene Cluster707	RBM6--3, nan--11005	MDM4--5, LOC100505937, SEC22B--3, COCH--1, nan--3804, IAH1--2, ANP32E--3, ATXN7--1, nan--1255, nan--5083, HNRNPH1--4, TOB1--2, nan--5227, nan--5387, C2orf68--2, CTTN--6, SLC38A1--1
Gene Cluster715	SMARCC1_weighted, OXSR1, IVD_weighted	
Gene Cluster716	nan--5138, DBF4B_weighted	ZNF468, ZNF260--2, ZNF600, ZNF432, ZNF765_weighted
Gene Cluster718	TUSC1, ZDHHC21_weighted, FAM179B, HSPB1, ARHGEF9--2, ZNF140--1, KLHL22_weighted, EPB41_weighted	
Gene Cluster719	COL4A2--2, TATDN3_weighted, C14orf181, RAPGEF2--1, GLCE	COL4A2--1, MLH1, EPM2AIP1_weighted, nan--5497, nan--5709, SEC63_weighted_1, TRMT12, CAPN5--2, C1orf103, QPRT--2, MGA-

		-3, SERPINE2--1, CXorf26_weighted, LOC100507636
Gene Cluster722	KRBA1, TPCN1--1, GALNT9	
Gene Cluster726	ZNF861P, AKR1E2--3	
Gene Cluster729	SRSF3--1, C4orf46--2, YWHAQ--3, C1orf135--2, TERF2--4, C16orf80, TUBB_weighted, ARL13B, CENPN--2, KLHL15, SPATS2--4, TUFM--2, COX4NB--1, RAD51L3--1	PGAM1, RQCD1--2, PRPS1--1, USP46--1, KIAA1524--2, CDKN2AIPNL--2, NEDD1--1, CCDC77, DNAJC8--2, TUBB3--1
Gene Cluster730	FAM161A--4, SIP1--1, MTF2_weighted, C14orf102--2	WDHD1_weighted, FANCM--3, GINS3--2, PLK4--1, C14orf145--5
Gene Cluster731	PPIL5--1, MTCH2--1, HSPA8_weighted, ITGB1BP1--2, KIAA0922--1, METT11D1--4, HSP90AA1_weighted	NDUFAB1, PTGES3
Gene Cluster732	CENPW, DEPDC1_weighted, ARNTL2_weighted, PRDX3, PBK, CCNE2_weighted, DUT--2, LDHA, TRA2B--1, PRMT6, SLC38A6, MPI--1, C18orf55--2	SNRNP25, BRIP1_weighted, TM7SF3--3, MLF1IP--3, POLR2D--2, SAMHD1--4, WDR61--2
Gene Cluster735	C10orf119--2, GINS4--2, MRPS5--2, PLA2G16--1, MAP3K3--2	LOC100506100, nan--5051, SBK1--1, C7orf30--4
Gene Cluster736	METT10D--3, nan--5252, ARHGAP10, LRTOMT--2, POLR2L--1, SNX8--1	
Gene Cluster737	INCENP, TOP3A--2, TSR1--4	GSG2, SPAG5, PSIP1--4, PCNA, HAUS8, MCM5--2, SLC19A1--4, nan--7571, XRCC3
Gene Cluster740	FAR1--1, ALG9 /// FDXACB1, GPR180--1, ARG2--2	WRB, FAM188A--1, CKLF--3, DBT--1
Gene Cluster742	SKP2_weighted, CDC25C--3, SMC1A--1, SLC25A15--2, NEK2--2, USP13--2, TFDP1--3	NAA16--1, IREB2--2
Gene Cluster746	FAIM, WDFY3--3, IFT57--2, HELQ--2, MTERFD2_weighted, GOLGA7--1, EZR--4	PAQR3--2, CGRRF1, RLIM--1, C15orf41--1, ITCH--1, ATL2--1
Gene Cluster748	nan--7263, CREB1--7, MORF4 /// MORF4L1, EHD4--5	CLIC4--3, CLIP1--4, PDIA6_weighted, AP3M1--1, MORF4L1--2
Gene Cluster749	SCAMP1_weighted, TMOD3_weighted, MIB1--4, RFX7--1, SNIP1, PHF17--2, CENPC1, ARID4B /// RBM34--2	KBTBD4--2, RC3H2--5, SPTY2D1--2, USP48--4
Gene Cluster750	ACTR2--5, CBLL1--1, MAPKAPK2--2	WAC--3, IL13RA1--2, DYNC2LI1--2, TMEM116
Gene Cluster753	IFI6, CSGALNACT2--1, DBN1--1, SGMS1, TUBA1A, TNFAIP8_weighted, ACER3_weighted, B3GALTL--2, CCNYL1_weighted, APC--1, TMEM65--2, CXorf40A /// CXorf40B--2, PANX1, ARL6IP6--3, HS2ST1_weighted	ALS2CR4--1, DEGS1_weighted, KIF3A--2, STX2--2, RAB8B--3, MAPK8--3, CWF19L2--3

Gene Cluster754	LOC100507345, MAN1A2--5, nan--4650, HINT3--2, ETNK1--3, ZNF77--2	
Gene Cluster758	TPR_weighted, FPGT, HFE_weighted, ASPM_weighted, TPM1--3, EAF2--3, ANKRD13C_weighted, GCLM_weighted, FAM119A--1	ITSN1--5, UST--2, SERTAD4--2
Gene Cluster762	HUWE1--4, RBM26--5, PCIF1--3, CREBZF--1	SPRED2--2, ARID1A--1, PRICKLE3
Gene Cluster764	SPRED1--3, nan--3831, BRAF--2, nan--5759, RWDD2B--2	ASB7--1, C3orf23--5, SORT1--1, USP25--2, RBBP9--4, NAB1--1, FBXO28--2, LOC96610--1, PIK3CB--2
Gene Cluster765	CTDSPL2--1, FAF1--2, TRUB1--5, IPO11--3, AHCTF1--1, RPAP3--1, GSTCD--1, GGA2--7, SENP1--1	GFM2--2, CCNF--2, SRSF1--1
Gene Cluster766	PIK3C2A--3, FCF1 /// LOC100507758, TP53BP1--2, MPHOSPH9_weighted, FBXO38--1, CBX5--5, JMJD1C--6	LARP4_weighted, SP1--2, KIAA1826--1
Gene Cluster767	IMPACT--2, SIKE1--5, SS18--4, RAB35--1, FEM1B_weighted, MTHFD2--2, EGLN1--4	LMAN1_weighted, GNA13--1, ZNF644--1, DDX3X--3, USP34--7, NUS1 /// NUS1P3, CCDC117--2, ZFX--1, CHD2_weighted, SPAST--1, C1orf55--1, KIAA1109--5, ZNF507--3, ZNF434--2, TCHP--2
Gene Cluster768	GRPEL2--2, B4GALT6_weighted, MYSM1--3, RPRD1A_weighted, FOXJ3--2, PBRM1--5, MGAT2--1	G3BP1--6, HIPK3--2, DHX34--1, ASAH2B--2, UTP15--1
Gene Cluster771	AP1G1--2, ABHD5--3, TXNL4B--2	FAM190B--2, GPD2--1, SGPP1--1, GOPC--3, RMND5A--1
Gene Cluster772	PTPN11--3, UBR1--2, ATF2--2, AMMECR1--1, C17orf80--3, MED16--2	WDR6--2, SLAIN2--4, RHOA--1, MFAP3--1, RAB23--1, PLEKHA3--2, CCNG2--4
Gene Cluster774	VKORC1L1--1, RAB6A--2, PSEN1--7, EIF4EBP2--2, MKLN1--5, LARP4B--5, SNX13--4, SNAP23--2, ZMYND11--2	ASXL2--1, CMTM4--1, CYB561--4, MLL--7, FIZ1--2
Gene Cluster775	HIPK1--2, ATP6V1C1--2, FAM105B--1, UBE3C--2, ENTPD7--1, MBNL1--1, GABPA--1	PAFAH1B1--4, POMGNT1--2, ABCD3--1, PSPC1--2
Gene Cluster776	NSMAF--1, DDAH1--1, CPSF1--1, PDE7A--4, LOC286109	MCM3AP--4, ZDHHC20--2
Gene Cluster777	ACP1--2, RALGPS2--2, TSN--3, PPIC--1	
Gene Cluster778	UPP1, SVIL--2, STAM2--3, BCAR3	GRB10_weighted, MYOF--2
Gene Cluster779	POMP--2, RAI14, C3orf59	
Gene Cluster780	SIK3--5, MEX3C--4	ZNF322B, CDYL--2, ATAD1--1
Gene Cluster781	TMEM14B--1, USP12--2, TMEM14B /// TMEM14C, KIAA1632--2, MAP3K2--3, PARP6_weighted	RNF38--2, CD276--3

Gene Cluster782	C11orf58--4, TMEM9B--2, PXN--2, AKIRIN2--1, STX6--4, GLYR1_weighted, SLC20A2	
Gene Cluster784	SLC25A5, SKP1--2, PEX13--4, PGK1_weighted	HNRNPH2--1, MTMR1--2, SLC39A6--1, SLC35A2--1, ZMAT2, NRF1--2, FUT8--1, WDR44--1, CUL4B--4
Gene Cluster785	CTTNBP2NL--2, FBXW11--2, SAR1A--3, GNG12--1, ZKSCAN5--1, DPY19L1--2	PRPF4B--3, MBNL2--4, IL13RA1--4, PSMA2--2, RAB1A--1
Gene Cluster786	C2orf47, C2orf69--2, SUMO1--1, OAZ1--2, POLR2E--1, AP2B1--2, RNPS1--1	
Gene Cluster787	LAPTM4B_weighted, GAB1--5, GPHN--3	EID1_weighted, SLC22A23--1
Gene Cluster791	SLMO2--2, HCCS--2, TXLNG--2, FAM83D, YME1L1--3, ACAD11 /// NPHP3	TUBGCP5--1, SLMO2--1
Gene Cluster793	FNDC3A--2, CPD--3, XRN2--2, PAPOLG--5, MORN2--2, WDR26--4, SAMD4B--1, COPS2--2, MOCS3, DSTN--1	nan--8683, ROCK1--2, ASRGL1--1, RPL41, TMED2--3, SELT--1, DENND4A--1, SS18--1
Gene Cluster794	RPL13A--3, HDLBP--4, WLS--2	RPL10A--1, TTC23--2, PROS1, GPATCH2--3
Gene Cluster795	C17orf69--1, LOC100131941--1	
Gene Cluster799	UBE3B--2, RXRB--3, SBF2--3, PIP4K2A--3, RHBDD1--2	
Gene Cluster800	GSPT1--5, PPM1A--6, ZNF398--1, CALM1--6	
Gene Cluster801	TOP3B--2, PTPRG--4, GALK1--1	CAMKK2--2, IGF1R--5
Gene Cluster803	TRAM1--3, SETX--3	CDK13--2, RPS15A--1
Gene Cluster804	CDC40--2, CNKSR3--1, TBL1X_weighted_2, TPD52--5, SNX13--1, C21orf57--3	SCOC--1, CASD1--1, AZI2--2, DLD--1, RPS23--2
Gene Cluster809	NT5E--4, ETV1--5, FAM20C--1, ZNF709--5	ANGPTL4--1, HOMER3--3, CTSC--4, TFPI_weighted_1, RGNEF--5, DPYD--3, ANKRD36BP2--1
Gene Cluster810	RBMS3_weighted, LOC203274, ANKRD50_weighted, RGS17, WNT5B_weighted, TLE4--5, PFKFB3, FBXL5--2, SLFN12, C8orf42--2, VEGFC, DOCK4--4, LOC729839, MAP9--2, HOXD3 /// HOXD4 /// MIR10B	ZNF573--2, C20orf12--2, APBA2--1, GALNT14--2, SREK1IP1--2, nan--10595, MCTP1_weighted, EVL--1, FHOD3, nan--9063, VPS41--3, ERGIC2--4, BICD1--4, NNT--3
Gene Cluster812	PDGFB--1, BIRC3, RASAL2_weighted, C3, UBASH3B--1, EMR1, OXTR, QKI--2, DENND3--1, QKI_weighted_1	INHBA_weighted, IGFBP7_weighted, nan--5590, MAMLD1, ALPK2, ADAMTS9--3, ADRB2
Gene Cluster813	SMAD3--1, EGFR--5, SRGAP3--2, ADORA1--1, ARHGAP29--3, DIRAS3	CATSPER1, THBS1--1, WTIP, WNT7B, nan--8528, TRIO_weighted, ENG--2

Gene Cluster815	nan--3863, SCARA5--3, PSMD9--2, PCDHB11	SHISA2, FGF2--1, nan--8971, CCNJL, C6orf153
Gene Cluster816	ZNF599--3, SERF1A /// SERF1B--1, DNAJC10--4, LOC100506783, COL4A1_weighted, SCLT1--3	APC_weighted, FER--3, NLRP1_weighted, ITGA10, GABRG2--2, FXYD2_weighted, LOC100132167, LOC285628, FBXL17--4, WWC2--3, NEGR1_weighted, TRERF1--2
Gene Cluster817	IL6, PTX3, MACROD2--4, RFTN1--2, PRR16--2, TPM1--6	PDGFB--3, CSGALNACT1--2, nan--5288
Gene Cluster818	LOC645638, BICC1--1, PADI2--2, ADHFE1, ADRB1--3, TMEM209--1	KCNIP3--3, TNF, SIRPA--4, IL15--1, FGF2--2
Gene Cluster819	NAV3--2, LOC100505562, LOC100130938, IL7, BATF3, MTBP--1, NKX2-5, ST6GALNAC4--3	IKBIP_weighted, CREB5--2, nan--4665, ERCC1--3, KLHL20--1, CUL2--1, LTBP2--2, TTC26--4
Gene Cluster820	ZC3H7B_weighted, CPNE8--2, nan--10268, FLJ45513, TM4SF1--5, nan--7456, GPR126--3, SLITRK4--2, SH2D5, LOC400043, C14orf105_weighted, ADAMTS1--2, HLX, ERAP1--4, DZIP1--1, NDST1--3, GABRG2--1, ZNF826P--1, ZNF677--1, FLJ35409--2, nan--712, nan--228, PHF17--1	
Gene Cluster821	PARD3_weighted, CDH2_weighted, MPDZ_weighted, SYDE1_weighted, PAX8_weighted, TRAM1L1, LOC643792, C18orf54--4, RIMS1--3, WSCD1_weighted, CNTNAP3--1, NF2--5, ZNF702P, PAX2--1, PRKD1--1, SNCA--2, nan--133, nan--132, CNRIP1, FOXC2--2, NRG1--3, SIRPA--1, LOC338620	
Gene Cluster828	API5--5, nan--3948, KIAA1549--2, nan--8726, nan--5512	ZNF182--2, LEPREL1--1, nan--727, CXCL6, nan--8965, LOC100505650_weighted
Gene Cluster829	CCDC132--2, nan--10523, LRCH3--1, nan--8953, nan--5982	nan--10574, CUL7_weighted
Gene Cluster831	AK4_weighted, TMEM22	RBL1_weighted, nan--7305, NFKBIE, nan--103, LTB--2, FAM111B--2, NIPAL4
Gene Cluster832	nan--6530	nan--7596, TM4SF1--3, LOC285835
Gene Cluster833	TMEM200B, RIMKLB--3	TMSB15B--2, SRR--3, ERC1--2, PCDHB3--1
Gene Cluster834	GCNT2--2, TRERF1--3, DFNA5, LPHN2--2, PRMT2--4, ARID2--4, FKBP15--6	ZNF420--1, RAB34_weighted, ZNF331--2, ZNF20 /// ZNF625--2
Gene Cluster835	SGCE, PEG10_weighted, FAM126A--1, SMARCA1_weighted, AMOTL1--1, FSTL1--1, ZNF691	RAB31_weighted, ANTXR1--4, PRKAA2--2, KLHDC8B, COL1A1--1
Gene Cluster836	SALL2, ZNF134, CHST10, HDGFRP3_weighted, WDFY3--1, CMTM3_weighted, ANTXR1--3, CREM_weighted, POP4--2	LHFP--1, FAM126A--3, KIAA0495--3, NR3C1--2, SDC2_weighted

Gene Cluster837	SMAD5--1, HOXC4, LEPR_weighted, nan--7721, SUMO1--2	ASAP1--5, MAP1B--2, NNMT_weighted, ZNF234--2, ZNF45--1, CPNE8--1, TRAM2--1, SNHG4--1, SCLY--3, TLE4--3, TMEM55A, FIGN_weighted
Gene Cluster838	PKIA--1, FBLN1_weighted, nan--954	MAGEE1--2, ATP8B2--2, SNCA_weighted, KIF5C--3, nan--9234, GNB4--3, C22orf39--2, FIP1L1--1, TXNRD2--2, MYH10--2, ADAMTS1--1, MCM8--2, CCDC102A
Gene Cluster840	CADM1_weighted, SIRPA_weighted, nan--9316, TRAIP	HMOX1, IL6ST--1, XPNPEP3--4, BCAT1_weighted, nan--5146, TPM3--4, PRICKLE1_weighted, RASSF2, SYNGR3, SOHLH2--2, FRMD4A--6, TTLL7--1, S1PR3--1
Gene Cluster841	TMEM136--1, WNT5A_weighted, FXR1--5, CHN1, FHL1_weighted_1, MSN, RASSF8--2, AKAP12_weighted, CEP170 /// CEP170P1, MTX3--2	GJC1_weighted, FSD1, RIMKLB_weighted, FBXO17 /// SARS2, GLI3_weighted, PIK3CD--1, HEG1_weighted, LOXL2--2, FERMT2_weighted, SLC23A2--4, MSRB3_weighted, SYNGR1--2, FAM55C--3, ARHGAP29--2, TUSC3_weighted, ATG10--1
Gene Cluster843	C10orf76--1, IFT52--2	SNX12--2, C10orf46--5, ZFP64--2
Gene Cluster845	EIF5A_weighted, H2AFY2, nan--5187, GALNTL4--1	AMFR--1, ROBO1, SMARCE1--4, C1orf53--2
Gene Cluster846	C8orf37, MPP2--1, ALS2CR4--2, IL27RA_weighted, EFCAB7	ZFP30, FECH--1, CLSPN--4, nan--7224, PAWR--5, nan--8684, FUT11, CDKL3--2, B3GALNT1_weighted, LOC84856--1
Gene Cluster849	EWSR1--3, DDX20--2, CLCC1--2, ACADSB--2, TMEM20--2	LRRC8C--4, nan--3787
Gene Cluster852	KCMF1--2, HNRNPD--8, C18orf54--3	nan--5389, NAP1L1--1, SLC35B4_weighted, ERC1--3, AMN1, SSX2IP_weighted, RNF138--2
Gene Cluster856	LOC100506844, ZEB1--3, VIM--2, IKBIP--1, C18orf54--5, MAPK12--2, CNTNAP1, NAP1L5_weighted, RGS4_weighted, C16orf45--2, nan--7260, FAM101B_weighted, nan--23	nan--711, E2F7--2, TMEM136--2, CHM--2, nan--5168, DNAJC18--2, KIAA0427--2
Gene Cluster857	NRG2--2, BEAN1, PCDHGC3--2, AKT3--5, PCDHGA1 /// PCDHGA10 /// PCDHGA11 /// PCDHGA12 /// PCDHGA2 /// PCDHGA3 /// PCDHGA4 /// PCDHGA5 /// PCDHGA6 /// PCDHGA7 /// PCDHGA8 /// PCDHGA9 /// PCDHGB1 /// PCDHGB2 /// PCDHGB3 /// PCDHGB4 /// PCDHGB5 /// PCDHGB6 /// PCDHGB7 /// PCDHGC3 /// PCDHGC4 /// PCDHGC5_weighted, VCAN_weighted	
Gene Cluster858	PTPRM--3, ITGB3--2, RBMS3--2, SPG20--1, TM6SF2--2, nan--11023, EDIL3_weighted, SLC14A1--1	
Gene Cluster859	LOC100506377, TNC--1, PDLIM7_weighted, NAV3_weighted, NRP2--2, ZNF655--1, DLC1--4, C5orf54	AP1S2_weighted, CALU--3, CDCP1--3, CDK17--1, LOC148189--3, DENND5B--4

Gene Cluster862	DUSP1--2, FOXL1--2, MRC2--2, YTHDC2--2, KPNA5--3, TRIM23_weighted, AZI2_weighted, TICAM2 /// TMED7-TICAM2--2, CHST7, BBS7--2, AGPAT4--1, DENND2A--1, MN1	TRIM37--1, OSTM1--2, SLC4A7_weighted, THSD4_weighted, PGBD1
Gene Cluster863	LGALS1, ZNF510, LNPEP--1, TCP11L1--1, SEC22C_weighted, nan--7447	ERCC1_weighted, ELK3--1, MAP3K12--1, CORO2B, ZEB1--2, SRGN_weighted, GFPT2, nan--5442, SFRS11, NTAN1--1, NRG2--3, DNER
Gene Cluster864	C20orf194_weighted, C3orf55--2, SLC22A4, ANK2--1	GNG11, AFAP1L2, TRPC1_weighted, SYNJ2--1, BCAP29--5
Gene Cluster865	OSTM1--3, SPIRE1_weighted, SHC1--1, KLC1--4	SHC1--2, TK2--4, GATSL3 /// TBC1D10A--2
Gene Cluster866	SLC2A10, nan--5064, FRAS1--4, FAM114A1--2, PPP2R3A--1, TRIB2_weighted, nan--11985	ZNF618_weighted, CDRT4--1, FAM107B_weighted, LOXL1, CALHM2_weighted, F8, nan--713, ARHGEF4--1, LTBP3--2, PLAT, TGFB2_weighted, ULK2--2, ARHGAP24--2, SYTL3--4, TBCEL, NRIP3_weighted, F3, RAP2B--3
Gene Cluster869	C6orf145--1, EEF1A2, C10orf47--2, MAP1B_weighted, FBXL2, PAX6--1	UBE2E2, NXN, RBMS1_weighted, ALDH1A3, SCRN1, nan--5664, PHLDB2--4, TEAD1--2, EXOC6B--2, PPFIBP1--4, PTPN14--4
Gene Cluster870	UGCG--2, CUEDC1--1	KLHL5_weighted, UBASH3B_weighted, MITF_weighted, ICAM1_weighted, APOBEC3C, TMEM132A--2, NKX3-1--1, BDNF_weighted, LPAR1_weighted, RBMS2--4, LOC730755, NFKB2--2, TRIM6, TLR2, CLIP4_weighted, ROR1--1, ICAM1--3
Gene Cluster871	GADD45B_weighted, IGFBP6, IDS_weighted, ZBED2, RASSF5--2, PLAU_weighted, MYBL1--1, HOMER3_weighted, DCBLD2_weighted, FGD6_weighted, C8orf58--1, RBMS2--5, FN1_weighted, KIFC3--2, UBTD1	CYR61_weighted, AKAP12--2, ETS1--3, SMURF2_weighted, IDS--1, OLFML2A, CST6--1, KLC3--3, PLCXD2--2
Gene Cluster873	HOXB2, EFNA5_weighted, nan--5498, FAM129A_weighted	SEC14L1_weighted, LOC153684, COL4A5--1, nan--8846
Gene Cluster874	EHD2--3, LACTB--1, SNX21_weighted, BHLHE40--1, TMEM17--1, HIVEP2--1, LARP6--1, CHST15, TFAP2A--2, FAS_weighted, PLAUR--3, TFAP2A--1	STK17A--3, DRAP1, LOC554202, EHD2--2, MET--3
Gene Cluster879	NCDN_weighted, AP2A1--4, FCHSD1--1	WASF2--3, ATXN2L--2, SNAPC2
Gene Cluster880	REEP2, CAMK2N2--2, SEMA4F--1	SMARCD3--1, NAB2--2
Gene Cluster881	nan--5382, PTPRS--3, RTEL1 /// TNFRSF6B, APLP1	RRAGD--1, OSBPL6--1, SLC16A2
Gene Cluster882	PARD6B--2, C3orf34_weighted	MAPK11--2, FMNL1--2, PFKFB2--1, PRSS53, RBM33--7
Gene Cluster888	ZNF416--1, ZNF585A--2, ZNF71--1	ZNF211, ZNF567--1, nan--5434, CBX6--1

Gene Cluster889	IGF2BP3--4, PVR_weighted, DHX35--3, TRNP1--1, SOCS2--2, DUSP5	ABCC4_weighted, CDKN2A_weighted, SOCS2--1, CAPRIN1_weighted_1
Gene Cluster890	PHF21A--1, ARNTL--2, KLF12--5, TICAM2 /// TMED7-TICAM2--1	GTDC1--1, SARNP--2
Gene Cluster891	SEC14L1--2, SLC23A2--3, nan--7599, PFDN4--2, PGBD3, CYTH3--1	LTBP2--1, PI4K2A--2, PHLDB2--5, ETS1--1, ATL2--3, SERPINE1_weighted, BTRC--3
Gene Cluster893	GPC1--1, SCARA3_weighted, PTRF_weighted, ITGA5, TUBB2A /// TUBB2B, FBXO2--1, MXRA7--2, SLC4A3, CHST3_weighted, PACSIN3--1, RAB6B--3	MYL9, CLU_weighted, PANX2--2, MAP4K2, C1orf43--1
Gene Cluster894	FAM176B--2, ANKRD33B, QSER1--1, PHTF2_weighted, PRIM2_weighted	
Gene Cluster895	OBSL1_weighted, TGM2--1, HOXA1, CHRNB1, ZNF512B--2, KIRREL--2	LOC100288911--1, ARL4C--3, COL18A1_weighted, C10orf114--2, PRICKLE1--1
Gene Cluster896	CAV1_weighted, WWC2_weighted, GNB5_weighted, SCD5--4, KLHL5--1, SIX4--1, SNX24_weighted, WWTR1_weighted, SOAT1_weighted, CALD1_weighted, FSCN1_weighted, CAV2_weighted, KIAA1467--3, FAM171B--1, CITED2_weighted, IL17RD--1, DLC1--1, MDFIC--2, ZNF226--1, nan--5116, RAB3B_weighted, SLC36A4	FRMD6_weighted, nan--5082, RDX--1, EMP3, IQCJ-SCHIP1 /// SCHIP1, DSE--3, IL6ST_weighted, NUAK1, PHTF1_weighted, FAM63B--1, LOC100509683, MYO5A_weighted
Gene Cluster897	ZNF227_weighted, TANC2--2, EFNA5--2, DZIP3_weighted, CEP70_weighted, HOXC6, MDP1_weighted	TSPAN4--2, CEP170_weighted, PDGFC--1, OGFRL1_weighted, C6orf225, ATP10D, TBC1D9--1, HCFC1R1_weighted, RAB38, DYNC2H1--3, OGFRL1--3, GJA1
Gene Cluster898	GPX8--1, KATNAL1--2, FZD2, AMZ2P1, MEIS2, ZC3H12C--2	LIX1L--1, CFL2--3, GPX8--2, MAP4K4_weighted, FAM195B, MT1E, AKT3_weighted, TMEM45A, SLC25A37_weighted_2, CLDN11_weighted, GPR176--2, ROBO3
Gene Cluster900	ZFR--4, C10orf88--1, GIN1--1	PHC1_weighted, USP46--2, SGTB--1, CSGALNACT2--2, C2orf76, CSRP2_weighted, TNNT1, MAP4K5--3, HNRPLL--2, DNAJC18--1, ZNF2, GTDC1--4
Gene Cluster901	NUP188--3, KLHL7--1, MPP6, CSNK1G1--2, MTMR12--1, GTF3C4--3	PLK4--3, C4orf46--1, RTTN--2, DOCK1--5
Gene Cluster903	nan--8770, NKIRAS1, CEP135--1, TWSG1--1, BLVRA--1, PKD2, ZNF180, SNX10, SNAPC1, BICD1--1	
Gene Cluster904	PALLD--3, HDAC10 /// LOC100509694 /// MAPK12, ENO2, CCDC15--1, RSF1--1, YTHDC2--1	CFL2_weighted, TPST1, MAP3K6--2, PALLD_weighted
Gene Cluster905	nan--8763, RAB12--2	
Gene Cluster906	NBPF3--2, HACE1	ENO1--3, ARSK--3, C16orf87--1, WDR91--1, FGFR1OP2--2, C21orf91--1, LOC387647--1

Gene Cluster907	11-Sep--2, EIF4E--1, EIF2C1--2	H2AFZ--1, FOXRED2_weighted, TET2--2, GTSE1--2
Gene Cluster908	BAG2--1, TGFB1--2, LOC100507448	MCPH1--1, CNOT7--1, NEIL2, PPP3CC--1, SBK1--2
Gene Cluster909	CASP2_weighted, DCAKD--1, LETM1--1, HNRNPH1_weighted, NAA15--5, WDR76--1, BBS7--1, NIPA1--1, nan--8757, PSTK--1, MIER1--1, ZMYM6--6, ING5--3	PGGT1B--1, SKA1, FDXACB1
Gene Cluster910	TOR1AIP2--6, PAFAH1B2--2, C1orf58--2, THRAP3--1, nan--8689, EEF1A1--2	FUT10--2, TMTC3--1
Gene Cluster913	MBP--6, ATP9B--6	
Gene Cluster915	NBPF1--2, TTC33--1, EXD2_weighted, SEC24D--1	PPP1R9A--2, ETV5_weighted, BBS4_weighted, WDR19--2, EFHC1_weighted, ELL2--3, PQLC2_weighted
Gene Cluster916	DUSP3--3, TAF9B--3, MAPK1IP1L--1, LEPRE1	KPNA6--2, STYX--4, GIT2--4, ZNF561--1, KLHL20--4, ZNF555--1, nan--5120
Gene Cluster921	ARMC8--2, MAP6D1--1, ZBTB49	FGFR1_weighted, FGFR1--2
Gene Cluster923	TMEM19--4, KLHL17--1	METT10D--1, RPL23--2, nan--4616
Gene Cluster927	LOC90499, NLRP7, nan--8409, nan--8930, ZNF418, ITSN1--8, PTN_weighted, D4S234E_weighted, LOC100127980_weighted, SOX15_weighted, EFS_weighted, CLDN6_weighted, PRTG_weighted, SALL4_weighted, ERVWE1_weighted, EDNRB_weighted, LRP2_weighted, C21orf105, PEG3_weighted, GBGT1, LOC440173, CYP11A1, MEOX1, MAGEA8, MAGEA10, CCKBR--1, MXRA8--1, C1orf105, TNNT2--2, ZNF813, ACSS3--1, MYL7, TP53TG3 /// TP53TG3B, CCDC8--1, ARHGAP28--3, USP51--1, LIN28B, XAGE2 /// XAGE2B, nan--5800	
Gene Cluster928	GAGE1 /// GAGE12C /// GAGE12D /// GAGE12E /// GAGE12F /// GAGE12G /// GAGE12H /// GAGE12I /// GAGE12J /// GAGE2A /// GAGE2C /// GAGE2D /// GAGE2E /// GAGE4 /// GAGE5 /// GAGE6 /// GAGE7 /// GAGE8, GAGE3, GAGE1 /// GAGE12F /// GAGE12G /// GAGE12I /// GAGE12J /// GAGE2A /// GAGE2B /// GAGE2C /// GAGE2D /// GAGE2E /// GAGE3 /// GAGE4 /// GAGE5 /// GAGE6 /// GAGE7 /// GAGE8, GAGE1 /// GAGE12F /// GAGE12G /// GAGE12I /// GAGE12J /// GAGE4 /// GAGE5 /// GAGE6 /// GAGE7, GAGE12F /// GAGE12G /// GAGE12I /// GAGE5 /// GAGE7	

Gene Cluster929	MBNL3_weighted, SMTNL2, KANK4, LOC253842 /// NR6A1--1, SLC19A3--1, DNMT3B, MOCS1--3, LOC339260--4, ACTG2--1, MYCN--1, ABCA4--2	C10orf82, NTF3, LAMA1--2, MAMDC2, C6orf99, RYBP--1, LOC100507263, AFAP1L1--2, ZSWIM7--1, DCP2--2, FES, ZNF114_weighted, ACVR2B--2, LOC728392 /// NLRP1--3, nan--7360
Gene Cluster930	ARID3B, FLNC, RASL11B--1, PRSS35, APOE_weighted, FABP3_weighted	RGS9, PDZD2--1, WNT6--2, nan--9363, DOCK9--3, MYO7A_weighted
Gene Cluster931	CCDC21--1, ATP8A2--1, STAG3, ZNF750, SLC22A11, NXF2 /// NXF2B--1, POMT2--1, LUZP4, NXF2, NLGN4X--2, TPK1--2, C17orf103--2, PACSIN1--2, LOC253842 /// NR6A1--2, S1PR2--2, PPM1N--1, PRIMA1, TRPM6_weighted, IGDCC3_weighted, ZNF589_weighted, GATA3_weighted, CXCL14--3, nan--7564, TNS1--2, ACVR2B--3, XAGE3, nan--6424, STON1--2, LOC100505712, FXYD4, PLEKHG4B--1, SPRR2G, MTUS2_weighted, ALAD--2, AZI1, CSF3R--1, SLC25A48, ZFP42--2, C10orf140--2, nan--3008, nan--3051, DACT2, CD53--1, CYP19A1--3, KIF1A--1, FBLN2, GAS1--2, ANKRD6--2, ENPEP--1, MYLPF, CGB /// CGB5 /// CGB7, FOXO4, MAP4K1--3, TKTL1--1, PHACTR1--1, JAM3--1, CPZ /// GPR78, WNT7A, nan--3882, NRXN2--1, CXCL12--2, GIP, SLC7A2--1, FAM65B--1, SLC7A4--1, ZNF415, FKBP1B /// MFSD2B	
Gene Cluster932	MFAP5_weighted, GABRP, KIF1A--3, C14orf149, SEMA7A--2	C10orf10--1, F11R--2
Gene Cluster933	LOC728449, IGF2AS	KRT9, LOC644192--2, nan--11604, CSF2RA_weighted, GREB1L--1
Gene Cluster934	SYCP2--1, CDKN1C--2, EMR2--1, BCL2L10--1, C11orf70--2	TCAM1P, nan--8735
Gene Cluster935	nan--9106, CRYBA2, LOC100509629	C15orf41--2, PARP11--3, XPNPEP3--2
Gene Cluster936	KCNG1--2, ZBTB46--2, FAM133A--2, LOC100288985, MMP14_weighted, STC1_weighted	
Gene Cluster937	DOCK9--5, nan--5736, APBB1IP--3, LOC100288617, MAGED4 /// MAGED4B_weighted, IRX4, PAX7--1, NOS2, ASGR2, TBX2--1, SLC46A1--3, HAND1, FAM78A	CRLF1, RBPMS2, TMOD1--1
Gene Cluster939	FAM178B, LOC285191--2	
Gene Cluster940	KIAA1462_weighted, SUSD4_weighted, SNRPN_weighted, ANTXR1_weighted, ZNF667--2, RASSF8--3, GNAS-AS, CXCR7--3, COL12A1--2, LHFP--2, ZNF445--1, SERPINE2--2, TGFBR1_weighted, ADAMTS12--2, TCF7L1,	

	SLC24A3--1, ITIH5--2, GJA5--1, AURKC, nan--3777, VCX2--1, LRRC17--1, IL2RB, PCSK2--2, COL1A2--1, nan--481, nan--4988	
Gene Cluster942	BOC--2, nan--8732, PSG5_weighted, DNAH3--3, PASD1, PSG6_weighted	PSG9--1, DOCK9--2, nan--9071, nan--852, MSX1--2, DOCK9--4, DKK3_weighted
Gene Cluster943	NFKBIB_weighted, ATP11A--2, MBNL2_weighted, nan--114, LOC642852--4, GNL1--2, ATP11A--3, nan--7848, TNNI3	CALB1_weighted, RGS5--3
Gene Cluster945	SPARC--1, ENTPD1--2, SPP1--2, MYO10--5	CXCR7--2, GM2A_weighted_1, TSPAN18--1, SUSD2--1
Gene Cluster947	nan--11634, nan--10685, MAP3K1--1, COL5A1_weighted, NDN, CTCFL, MAGEA1, GABRA3, MAGEC1, SLCO2A1, BAGE--4, DAAM2	
Gene Cluster949	LOC100288413, ATXN3_weighted, CTHRC1, LOC283104--2, nan--7716, MED13L--5	KLRG2_weighted, IL20RB
Gene Cluster950	NMD3--2, SYT17--2, CLEC2L	ISM1--2, PRKCI--1, PNO1--3
Gene Cluster951	C6orf35--1, ZNF805--2, nan--5108, C2orf49--1, FLJ36116	CDYL2, LOC339260--3
Gene Cluster952	CCDC150--2, ZNF222, ZNF235--2, KIAA0586--2, SUSD4--1	ZFP82--1, SGSM2--2, ZNF320--1, PDE4DIP--9
Gene Cluster953	BRWD1--3	FBXO36--2, ZYG11B--3, HELB--2, BRWD1--2, LRIG3
Gene Cluster956	KCTD15_weighted, MFAP3L--2, ZSCAN30--2	DNAJB4_weighted, PPP3R1_weighted, SERPINB8--2, ZBTB40--2, LOC221710--3
Gene Cluster957	ISYNA1--1, SLC7A2--2, TKT--3, TLE3_weighted, LYPD6B, NHS--1, KCNMB4--2	PRKAR2B, PACSIN2--1
Gene Cluster958	LOC100507198, SERTAD4--1, ZNF532--2, CLN8_weighted, FLVCR2--2, AHDC1--1, ZNF79--1, PCDHA1 /// PCDHA10 /// PCDHA11 /// PCDHA12 /// PCDHA13 /// PCDHA2 /// PCDHA3 /// PCDHA4 /// PCDHA5 /// PCDHA6 /// PCDHA7 /// PCDHA8 /// PCDHA9 /// PCDHAC1 /// PCDHAC2--2, BMPR2_weighted	TRIM45_weighted, GRHL1--1
Gene Cluster959	CH25H, ARMCX1, CD99L2--3, RAB27B_weighted, LIFR_weighted	IL1R1--1, LOC26172, ARMC9--1, NAV2--5
Gene Cluster961	TSPYL5, NPFFR2, nan--6747, ZNF542--1, ATP8B3--3, PDGFD_weighted	
Gene Cluster962	LOC100128252--2, ANKRD6--1, DYSF, CRISPLD1, nan--7529, HIC2_weighted, ODZ4	LCP1, SGK494--1, ZBTB34, ANK3--2, AP4S1_weighted, ZNF606--1
Gene Cluster963	nan--9123, nan--8768, GRAMD2--2, SEMA4C--1, FRS2--2, PHYHD1, ZYG11A, ANKRD23 /// ANKRD39--2, MUC15_weighted, NMNAT2--1, VWA1--2, TMOD1--2, nan--298, LAYN--3	TP63--3, B3GNT7_weighted, TMEM150C--1, C10orf11, MLLT4--5, KSR1--4, nan--873

Gene Cluster965	VANGL2, RASGRP1, nan--7722, LOC100288911--2, NR4A3--2, GAB2, ZNF175--1	PDE4D_weighted, SAMD4A--3, ADAMTSL4_weighted, SOCS1--3, LIG3_weighted, GJA5--2, SIGLEC6_weighted, C21orf129, SILV, FLVCR2--3
Gene Cluster966	PPP2R3B--2, ACTL8, SVEP1--4, SP6, MED22, CAPN6--1	LOC645722--1, TBX2_weighted, SLC46A1--1, MRPS14--1, DRG2_weighted
Gene Cluster967	nan--6164, NFRKB--3	EPAS1--2, PLAC2--1, ADAP2--2, EFCAB2_weighted, SPEG--2, UPK2, HIP1--3, DUSP9
Gene Cluster968	SLC2A14 /// SLC2A3_weighted, NANOS1, SLC2A3_weighted, SLC1A3--2, LBH, NFATC2_weighted, IGF2BP1_weighted, LOC100509961, nan--7375, TNFAIP2--2, SPECC1--2	GABARAPL1_weighted, ODZ3--2, VGLL1_weighted, JPH2--3, nan--7700, GABARAPL1 /// GABARAPL3
Gene Cluster969	ALPP, TNNC1, C4BPB, nan--10956	ALPP /// ALPPL2, TFAP2C--2, ALPPL2_weighted
Gene Cluster970	GATA3--3, GM2A_weighted_2, PTPN14_weighted, FAM46B, ST3GAL1_weighted, C1QTNF6--1	IGDCC4, DAPK1--2, SLC1A3--1, C12orf59
Gene Cluster971	ARRB1_weighted, STS_weighted	PTGES_weighted, NOTCH3--2, nan--5221, P2RY2, HES2--3, STS--4, GALNT2--3
Gene Cluster973	ARHGDIB--3, ST3GAL4--1	ARHGDIB--2, ERP27, ADAM8--2, RECQL5--3, KLK8--1
Gene Cluster974	TNFAIP3--1, COBL--1	KIAA0040--2, OPN3--2, PCSK6--3, FNDC3B--4, nan--11380
Gene Cluster975	NFE2, NAV1--5, BTBD19--1, DOCK8_weighted, SH3GL3--2, RAB3B--2	nan--1559, OLFML3, MYCN--2, HIST1H4E
Gene Cluster976	WNT3--3, nan--8621, SNRPN--5, TAGLN_weighted, ROR1--3	CRYAB, FGF18--4, PP14571, nan--6182, CLIC5--5, nan--8850, ARHGAP28_weighted
Gene Cluster979	GINS2--2, nan--11034, CSTF3--2, FAM111B--1, RDM1, ELOVL6_weighted, NUCKS1--2, MKI67--1	HELLS_weighted, BRCA1--1, PPIA_weighted, C13orf27, HMMR_weighted, C16orf75, HNRNPD--6
Gene Cluster981	CDC20, LMNB1, TRIP13, CCNF--1, CASC5--4, CENPJ--2, CDCA8, TOP2A_weighted, FAM72A /// FAM72B /// FAM72C /// FAM72D, KIF15, BIRC5--2, MKI67--4, TKT--2, TTK, TACC3, NUSAP1--1	POLQ--2, FANCI_weighted, HMGB2--1
Gene Cluster983	TNFAIP8L1, ZNF558, CDCA7L, CEP110--1	MKI67--3, NEK2--1, PLCG1--1, MDC1--1, CDCA3--1
Gene Cluster985	SUV39H1, FTSJ2--2, RNASEH2C--1, AP3S1, CHD2--4, PGGT1B--2	NDUFB8--1, PRICKLE4 /// TOMM6--1, SDCCAG3--2, EHMT1--2, KIAA1731--2
Gene Cluster986	NCAPG2, NUSAP1--2, SSR1--5, NME1-NME2 /// NME2, LSM5--3	
Gene Cluster988	TUBA3D, TUBA1B_weighted, UBE2T, ZWINT, MELK, PRC1, MCM3, RACGAP1, CDCA5, HNRNPA2B1--1, GTSE1_weighted, KIF18B, ATAD5, ESCO2--3, SEPN1	RRM2_weighted, SPC25, BUB1B, MCM10_weighted, TK1--2, TUBA1C_weighted, TK1--1, MCM2,

		FEN1_weighted, TUBA3C, PIF1--2, H2AFZ--2, KIF18A, BIRC5--1
Gene Cluster989	ORC1, VRK1, MND1, CHEK1_weighted, NCAPH, GSTO1--1, DNA2, CHRNA5, CEP55	NEIL3, POLE2, SRSF2_weighted, MAD2L1--2, FANCD2--3, nan--8510, RAD18--2, TMEM48--2, SLBP, CDK1_weighted, ANKRD32--2, RNASEH2A, SKA2--2, ERCC6L
Gene Cluster990	ZNF367, FANCA--2, ASF1B, WDR34, MLF1IP--2	E2F1--1, CENPM, MCM7_weighted, C15orf42_weighted
Gene Cluster993	SSU72--3, COQ4--2, MSH2	MSH6--3, C1orf57--1, SPIN4
Gene Cluster994	CCNA2--2, CENPA--1, CCNB2--2, C12orf24, POC1A--2	KIF2C--2, AURKB--1, BUB1--1, FOXM1, CDKN3_weighted, CENPF--1, DLGAP5, TPX2, ASPM--1, PLK1, ANKH_weighted
Gene Cluster995	TUBB2C--2, C1orf112, LIG1, E2F8, FANCG, HAUS7 /// TREX2, CENPL--1, ITGB3BP, KIAA0101--1, RBBP7, ALDH7A1_weighted	KIF20A, PTTG1, PLK4--2, SNRPG, nan--5194, DHCR7--2, RAD54L, CDCA3--2, KIF4A, MSH6--1, DYNLRB1--4
Gene Cluster996	LPCAT1, ZMYM1, GSTO1--2, NUF2, SFXN1--2, GGCX--2, KIF14--2, MASTL, nan--7435, SMC2--2, CDR2L, APOOL--3, SAMHD1--3	KIAA1549--1, C16orf55--2
Gene Cluster997	nan--7269, WWP1--1	CNN3, nan--5210, nan--7410, SASS6, SLC25A23--1
Gene Cluster999	SHCBP1, NCAPD3, PSMC3IP--2, RNFT2--3, SDR39U1, OAT, ABCC4--3, USP37--3, KIF1B--3, FAM119A--3	FBXO5--1, TAF5L--1, TMED7--2
Gene Cluster1000	SCPEP1--2, C17orf75, YTHDC2_weighted, ZCCHC10--1, KIAA1033--1, nan--3136	RELL1--2, APLP2--8, ANXA1--1, KLHL7--2, EME1--1, CEP57--3, FZD6, PICALM--3
Gene Cluster1001	NME1, URB2, MEF2A--2, ADSS--1, KPNB1_weighted, 2-Sep--3, HDLBP--1, STK25, DDOST--1, ZNHIT6--1	
Gene Cluster1002	B9D1--2, E2F7--1, PIGK--3, 5-Mar--2, nan--5397, nan--5050, FXN--1, HP1BP3--4, PAFAH1B2--3, SRSF3--3, RBM12--2	CDC6_weighted, PEMT, DNAJC9--1, SLC20A1, PPHLN1--4, TFAM--2, PDIA6--2, LSM4--2, TUBG1, FOXRED1
Gene Cluster1006	RFC4, PGM1, QSOX2--2	RANBP1--2, ACLY--2, HPRT1--2, NDUFS8--2, PGP--2, SLC15A4--2, TMEM85--1
Gene Cluster1007	PTPLA--2, SSRP1_weighted	LOC100509231 /// LOC100509288 /// OVOS2, NCS1--2, CCDC109B, NUDT5--2, NLN--4, RGS10--2
Gene Cluster1008	KLRG1, RND2--1, NRK--1, EPB41L3_weighted, 6-Sep--4, LPAR3--1, SEMA6D_weighted	CDKN1C_weighted, TEX19, ITPRIPL1, TUB--3, RIMKLB--2, STON1--1, TGFB3--2, EFHD1, APOC1--1, FAM184A--2, LOC642533, GNAL_weighted, RARS2--1, FAM175A--2
Gene Cluster1009	GGA2--3, CAND2--1, HOXC13, nan--6172, LYPD6--3, LOC100271840, nan--11876, nan--11965	ZNF519--3, CADM1--3, nan--7540, nan--11731, ANGPT1_weighted
Gene Cluster1010	ANO4--2, FOXG1--1, RLBP1, AASS--1, FXYD6, COLEC11, LOC100505880	HOXD10--2, FOXN3--7, nan--11762
Gene Cluster1011	GNAL--2, RABEP1--1	HOTAIR, nan--8736, CUL1--1, FYN--3, ERO1LB--2, nan--5171, MTSS1--2, TNFRSF19--3

Gene Cluster1012	EYA4--3, LDB2, GRPR, RUFY3--4	NIPSNAP1--1, YPEL1--1, VRK2--2, OXR1--3, RTN4R
Gene Cluster1013	TBL1X_weighted_1, PPM1M, KCNQ4--2, TRPS1_weighted, RAC3	ACOX3--1, HSD17B14--2
Gene Cluster1016	C1orf54	GNG7--2, nan--3750, MORN4--2, nan--4506
Gene Cluster1017	LRRC49, NARG2--2, THAP6_weighted	LOC399491--2, PDCD5--2
Gene Cluster1019	PKIA--2, LOC440905, FRAS1--1, nan--5173	PLAC1, BCL2L11--6, nan--4901, C3orf17--1
Gene Cluster1020	HOXD13--1, HOXD13--2, nan--8363	ZNF280C, TRDMT1--1, SOCS4--1, DLX2--1, ANKRD26--2, DLX1--2
Gene Cluster1021	C13orf33--3, SCARF2--1, FAM69B--2, DCLK2--3, LOC100127983--2, LOC646509--1, WDR33--5, WDFY4--2, nan--5509, GALNTL1--2, nan--5696, nan--5826, NEXN--2, ODZ2, nan--6358, nan--7355, nan--7583, LOC100505470 /// LOC646576, HHIP--4, TMEM136--3, RBM20, LOC645323--2, POTEM, nan--11496, PARP8--2, ENSA_weighted, nan--6224, JAZF1--1, nan--4818, NEFL--1, EPHA8--1, UCHL1--1, nan--464, FRAS1--2, nan--2552, nan--2740, MEGF11--3, FARS2--1, THBS4, POU6F1--1, ADAM23--2, GDF1 /// LASS1, GNG7--1, SOX5--2, RUFY3--2, EMID1--2, SLITRK5, CTAG2--2, GABRA2--3, PLEKHO1, GOLPH3L, PKNOX2--1, ENOX1, ZFHX4--2, TARS2, BACH2--1, LEF1--2, DPYSL5_weighted	
Gene Cluster1022	EMX2, PGM5--3, ADAMTSL1--4, GABRA2--2, HSPC159--1	nan--5228, FAM120C--1, EDA--1, nan--7566
Gene Cluster1023	C5orf13--3	HBQ1, CD200--1, FLJ42627--2, KCTD6--1, CD200--2
Gene Cluster1024	nan--6483, FAM81A--3, nan--48, FBXO45--3, KIAA0368--1, NDUFC1--2, nan--5586, NUFIP1--3, nan--8802, nan--5997	LOC729970--1, GNAL--1, MED4--2, KIAA0494--1
Gene Cluster1025	nan--5579, TCF4_weighted, KREMEN1_weighted, nan--5340, SLC38A4_weighted, C19orf39, DUXAP10--2, nan--5348	WIF1, FOXP1--3, SDK2--5
Gene Cluster1027	METTL2B, NRP2--1, FAM13A--2, MCL1_weighted, GPC6_weighted	CCDC136--2, nan--9770
Gene Cluster1028	AR_weighted, PPM1L--3, nan--9388, DIRAS1--1, LIX1--2, nan--10549, nan--7857, SEMA3G, LHX2--1, FHL1_weighted_2, nan--785	KCNJ12--2, ADCY1--5, C22orf34--2, NUDT11, LOC644450--3
Gene Cluster1029	SULT1E1--1, HHIPL2, CCR10, FERMT3, C19orf30--1, BARHL1, ACAP3--1, PRRT2, NTM--2, ACMSD, ST8SIA4--4, CSRNP3--6, WDR25, nan--8282, TRIB1--3, nan--10417, nan--11046,	

	CABP7--2, KLB--2, MTSS1_weighted, PKD1_weighted, NCAM1_weighted, RBFOX1_weighted, ELOVL2_weighted, ELN_weighted, GNG2_weighted, ARX--2, ROGDI, BEX1, FAM5C, SLC44A5--1, ZNF572, SYN1--1, nan--258, HAR1A, TPCN1--2, GNG4--3, GSTA1--1, MFNG--2, NNAT, CHGA, CACNA1H--1, FOXF1, ARSB--3, CDH15--1, RENBP, SYNGR4, CBFA2T3, FGFR2--6, RASA4 /// RASA4P, ARHGEF6, UGT2B28, RASA4 /// RASA4B /// RASA4P, LOC100291393 /// SHC2, PRPH, CELF3--2, PRODH2--1, CAMK4_weighted	
Gene Cluster1030	SQLE--1, TTLL6	PAX5--2, SRRT--4, nan--5793
Gene Cluster1031	ASCC2, C17orf28_weighted, LOC729021 /// LOC729218, FAM131B	PKD1--2, SLC22A3--1
Gene Cluster1032	GUCA2A, GGT7_weighted, CLCN5_weighted	HOXA11--1, TTR, C5orf4--1, ONECUT2--5, CBLN1, KCNJ3--3, AKR1C4
Gene Cluster1033	GCG, CMIP_weighted	DLX2--2, SRRM2--4, LRRC16B
Gene Cluster1036	SPDEF--1, SPDEF--5, NCRNA00261	CACNA2D2, SPDEF--4, RAP1GAP--1, ANXA10, F5_weighted, HEPACAM2, LTK_weighted, MRAP2, INSM1, RAB26_weighted
Gene Cluster1037	CSMD3, TAGLN3, TNFRSF11B--1, MYT1--2	ABCG1--2, ABCG1--1
Gene Cluster1038	ALDH1A2_weighted, RTN1_weighted, COL28A1--2, HS3ST5, F5--3, nan--5157, ATP1A3, SI, SYTL5--1, OLFM1--1, NPB, LHX9--2, MGP--1, CRMP1, RHOH--1, BEGAIN, CLCN4--2	SMAD9--1, FSTL5, CDH11_weighted, GLRA2, AQP3--1
Gene Cluster1039	CXXC4_weighted, ENPP2_weighted, COL3A1_weighted, SOX6_weighted, GLCCI1_weighted_1, NXPH1, LPHN3_weighted, nan--5271, RPRM, MEP1B, NPTX1, PVT1--2, PCDH9--1, DLL3--1, TRIM9--1, CD36_weighted, SNAP25_weighted, DLX1--1	WDR33--1, ABCB4--2, LOC100507778
Gene Cluster1040	ALX1, SLAIN1, NTS, IRX5	BMPR1B--2, KLHL13, SLITRK6_weighted
Gene Cluster1043	PCDH18--1, PKD1L2--2	
Gene Cluster1044	FAM49A--1, AMBN, nan--5078, ASAM, GLDN--1	SV2A, LOC100128098, EFNB3--1, UCHL1--2, TBC1D24--1
Gene Cluster1046	ST3GAL6_weighted, FRMD4A--5, TRPV2--1, CNPY4--2, GRK5--2, BMP7--4, TMSB15A	DSCR8, C12orf56, TUBB2B, HSPB8--1, PCOLCE, LOC100216479, DMD--1, ECHDC3, STX17_weighted, CPVL, PLXND1_weighted

Gene Cluster1047	GPR37--1, NEURL--2, PBX1--1	
Gene Cluster1048	ENO3--1, ELMO1, GPR19	HEY2--2, KCNMA1--3, C1orf74
Gene Cluster1049	THSD7A--4	SMA4, C16orf7
Gene Cluster1050	PDGFA_weighted, GAD1_weighted, CDK14--1	ITGB6_weighted, PTGER2, TNFRSF11B--2, SHF, SPOCK2_weighted, TMEM27
Gene Cluster1054	AOX1_weighted, DPF1--2	LOC220930, TWISTNB--1, NHEDC2--1, NOL6_weighted
Gene Cluster1055	nan--2103, nan--11356, nan--11893, CD9--2	nan--8155, AKAP13--8, nan--9107, nan--5321, SPINK5, nan--5679, TTLL4--1
Gene Cluster1056	AKAP9--3, nan--7284, SCAPER--3, SCAPER--1, SHANK2--2, nan--7930, MCTP2--5, C5, ZNRF1--5, RHBDL2--1	nan--299, AQP6--2, RPS6KA2_weighted
Gene Cluster1057	EVL--2	MB, XKRX, TBCD--1, HPS4--3, RNASE6, SLCO3A1_weighted, DTNB--1, RFX2--2
Gene Cluster1060	APBA1--2, PHEX--2, LUM, SHC4_weighted, ST3GAL3_weighted	KRT4--1, GLT8D2_weighted, SIX1--1, OLFM1--3, SHC3--2, MAP2--2, EPR1, FAM174B--3, SIX1--2
Gene Cluster1062	nan--1267	SLC4A4--2, INPP4A_weighted, GKAP1_weighted, MFGE8--3, SSBP2_weighted, ANXA1--2
Gene Cluster1063	INPP4A--6, TIGD7--1	DOCK3, PCSK1N, SUPT7L--3, TECPR2--1, FRY--1, nan--5089, LOC100508546--2, DCAF5--3, FAM200B--3, ZSCAN18--2, FGF13, DTX3--3
Gene Cluster1064	ZNF85--2, KCNJ15_weighted	LOC100289049, USP15--3, PMS2P1 /// PMS2P5, CSTA, ENPP5--1, KISS1R, PTGS2_weighted
Gene Cluster1066	GUCY1B3_weighted, CX3CL1_weighted, IL1F7_weighted, CPM_weighted, GUCY1A3_weighted, MFSD4_weighted, nan--10787, IL1RL1_weighted, nan--9111, nan--1073, PRKACB--3, FLJ38109, nan--4640, CLEC2B--2, SERPINB7, WDFY3--4, nan--9185	SLC38A11, FAM83A_weighted
Gene Cluster1069	nan--12041, nan--11382, CFH, CFH /// CFHR1, ABI3BP--3, nan--8408, nan--7446, SLFN13_weighted, IKZF2--3, SEMA3A--2, nan--10981, VPS53--5	nan--888, SPRY1--2, nan--926, nan--5555, ZFP28--3, CCNG2--5, FOXA1--2, HOPX--3, DPH1 /// OVCA2--2, KIAA0513--2
Gene Cluster1070	C6orf52--3, CASP4--2, ZNF839--1, TMF1--3, PNPLA4--2, TLR3	RIT1--4, MCTP2--1, ARHGAP20--2, SHISA3, LOC643733, nan--11674, DNAH11--1
Gene Cluster1071	ATG2B--2, SPATA7--1, PCLO--2, SUPT7L--1	ADSSL1--1, HIST1H2AM, PELI2, AQP3--3, SPATA7--2
Gene Cluster1072	HLA-DMA, B3GNT9, PYROXD2--1, HIST1H1C, RAMP1	

Gene Cluster1073	SOSTDC1, PLCL2--1, NR3C2, EFR3B--3, PPA2--4	NCOA1_weighted, KAZALD1--1, EIF2C4--4, ZNF30, SYNE1--1, FAM189A2
Gene Cluster1074	nan--9046, nan--8058, TBX1--4, nan--7444, RGS8 /// SDHAP3, nan--6722, CLEC2D--3, nan--6100, DNAH5--1, RBPMS--6, TMEM150C--2, nan--5827, nan--5306, BCL6--3, SCUBE3--2, nan--9159, RBM43--1, RNF144B--4, nan--9674, NKX2-1_weighted, TBC1D15_weighted, FBN1_weighted, GLRB_weighted, IKZF4_weighted, CNOT2_weighted, TMEM19_weighted, CP_weighted, BANK1_weighted, MAP3K13_weighted, nan--11690, nan--11606, nan--11115, CLIC6--2, nan--10546, TPBG--2, nan--5193, NCRNA00086 /// NCRNA00087, LOC100506621, CAND1--2, ALK--2, TP63--2, FOXE1--1, STAT4, GPR109B, LMO3--3, THBD--1, APOD, RBM14 /// RBM4, ZNF826P--2, nan--1256, nan--1130, nan--949, KRBA2, STK16--1, IDS--4, MLL--3, MEGF9--2, CXCL17, MTMR3--4, TMTC1--3, NBEA--2, RAB21--2, LONRF2--1, CAPNS2, CLEC7A_weighted, NAPSA, MDM1--2, EPHX3, KCNMB4--1, SLC35E3, SRD5A3--1, LRRC23--2, HSPA6--2, LRRC4	
Gene Cluster1075	MUC1--2, LOC642587--2, VTCN1, MTM1--2, LYNX1--1, SSPN--4	IL7R_weighted, RNF24--1, nan--8067, LOC340340, INPP4B--2, TUSC3--1, SLC4A11--2
Gene Cluster1076	ZNF641_weighted, KLHDC7A--2, SOX4--2, LOC254896 /// TNFRSF10C, TRAFD1_weighted, LRMP--1, IVL, LOC374443--2, nan--8771, ZNF319	SYNRG_weighted, GPR87, CLPTM1L_weighted, PRRT3, RASSF5--1, VASN, RHOBTB2--1, NPC2
Gene Cluster1077	IDO1, ARHGDIB--1, PIP4K2B--3	MUC4_weighted, MUC16, LRRN1, PHACTR3, TRAFD1--3, LOC402778_weighted, TRIM22
Gene Cluster1078	LOC643529, nan--10489, SLC15A2--1, nan--6029, nan--6143, RNF213--6	ZBTB43--2, ANKRD13A_weighted, nan--664, nan--7085
Gene Cluster1079	QKI--10, nan--10929, nan--5834, ZNF853--2	TMEM40_weighted, LOC100289775 /// WNT7B, LOC100506119 /// LOC100508342, nan--937, GLIS3--2, GPNMB--2, ELFN2_weighted, LOC100499466--2
Gene Cluster1080	C2orf74, PPP1R3C--1, LOC339803--2, CST6--2, nan--10285, ITGB8_weighted, ITGB2_weighted, EFEMP1_weighted, LOC441869--2	FST_weighted, nan--3627
Gene Cluster1081	SLC6A14, KIAA1245 /// NBPF1 /// NBPF10 /// NBPF11 /// NBPF12 /// NBPF24 /// NBPF8 /// NBPF9--3	KRT15, DST--10, CCDC80_weighted, nan--6069
Gene Cluster1082	CYP1A2--1, HHAT	NEDD9--1, L1CAM--2, MBOAT7--2, PRKCG--2, GJB5, LY6D
Gene Cluster1083	TGM2_weighted, C2CD3--3, CADM4--2, ARHGAP23--2, TRIOBP--1, CYTH3--3, SLC48A1_weighted	PPME1_weighted, CST5

Gene Cluster1084	TEAD1--1, PPFIBP1--1, OLR1, RNF182, PCYT1A--2, LRP12_weighted, nan--8910	PRKCQ--1, UROS--2, PTPN1--2
Gene Cluster1085	LGR6, nan--5511, SIK2--1, TGM2--4, BTG4--2, LRRC34--1	nan--45, nan--10938, TH, KCNJ4--1, ANKRD2, nan--7735
Gene Cluster1088	TCHH, LOC100505490	ANXA8 /// ANXA8L1 /// ANXA8L2, DLG1--5, RNF207--4
Gene Cluster1089	FN1--5, MIA, ARTN_weighted, SNAI2, nan--3797, CCDC123_weighted	SERPINA5, SERPINA3, CCL2, MCC--1, RARB_weighted, NDP, COL6A2--1
Gene Cluster1090	SORBS1--3, ADAM19--1, LHX6--1, CST7, SLC43A3_weighted, LAT2--2, LOC100287428, TOX2, UNC5A--1, FAIM3_weighted, PAQR5--1	ZCCHC24_weighted, LOC145783--3, DNTTIP1--1, C7orf10, C9orf70, DUOX1--4
Gene Cluster1091	ALOX5AP, BID--1, CCK, AURKA--3, CDKAL1--2, KATNAL1--1, RAB7A--5, LOC100287616, PAGE1, CNTD1, NPEPL1_weighted	
Gene Cluster1092	ARPM1, LRRC34--2, NGFR, CYP24A1, KLRC1 /// KLRC2, AP1S2--1, LOC404266_weighted_2	PCSK5_weighted, ITGA1--1, PPFIBP1--5, ANXA2P3, ATXN7L1--2, VRK3--6
Gene Cluster1098	nan--3877, RAB43	
Gene Cluster1100	C18orf25--4, ATAD3B--3, CPSF3L--4, TLR6_weighted, nan--4680, STEAP4--2, TRPV4, NUP160--3, CFLAR--5, IL1A--2, HLA-DQB1--2, PPP2R3A--2, nan--8504, nan--9481, CADM4_weighted, IKZF1--5, EDN1--1, RASSF4_weighted, LOC441869--1, LOC148189--2, nan--5552, C5orf46, PPP2R2C--2, KRT14, ABL2--1, HS3ST1--1, SERPINA1--4, ST6GALNAC5--2	CLIC6--1, CD274--1, CXorf61
Gene Cluster1103	IFIT1, DDX60, CMPK2, SP100--3	IFIT3--2, IFIT2--2, OASL_weighted, IFI44--2, MX1, IFIT3--1, GBP1--1, DHX58, MX2, DDX60L--2, RSAD2_weighted, XAF1_weighted, IFIT2--1, IFI44--1
Gene Cluster1107	SLFN5--2, ABLIM3, DUSP4_weighted, HDAC9--3, ANO1_weighted, TGFA--1, C9orf169, HLA-DRB1 /// HLA-DRB4, EMP1--1, MAML2--1, TGFA--3, PRKCDBP, PRDM1--2, IFI16_weighted, CD163L1, PDE4B_weighted, HLA-DRB1, KRTAP3-1, CD274--2, SAMD9_weighted	HLA-DRB1 /// HLA-DRB3 /// HLA-DRB4, HLA-DRA_weighted, HLA-DRB1 /// HLA-DRB4 /// HLA-DRB5, nan--7437, IL23A--2
Gene Cluster1109	SAMD9L--1, CXCL5_weighted, TRANK1, SCEL_weighted, LYST--2, nan--5219, TTC9--1, KYNU--3, KYNU--4, RCAN1--1, RAGE, KYNU--1	DAPP1_weighted, RNF144B--1, VNN1--2, THBS1_weighted, IL1B_weighted
Gene Cluster1110	RASGRP3--1, EDN2, TRIO--2, SDCCAG8--3, ATF3--1, FAM167A--1	GLI2--3, NID2, HRASLS2--3, NSD1--3, KLK10--2, C18orf25--2, PTPRR_weighted
Gene Cluster1111	GPR110--4, MED15, RARRES3, IGFL1, SH3RF2--2	FSTL3, PTAFR_weighted, nan--5302, INPP4B_weighted, GPR110_weighted, CACNG4--1, nan--11314, STEAP4--3, UPK3B, KLK5, C10orf10--2, LAMC2--2, KRT81, EMP1--4, HSPG2--1, RCAN1--2, SERPINB2

Gene Cluster1112	CCDC69--2, IFRD1--3, NCF2, COL16A1	CDKN2B--2, SPNS3, CD74--3, MSLN, TRIM29--3
Gene Cluster1113	CDKN1C--3, MSX2_weighted, ZNF280B--2, SLC2A1--2	PSTPIP2, FGF18_weighted, SLC2A1--1, CDKN1C--6, DCUN1D3, RORA_weighted, nan--5111, PCSK5--1
Gene Cluster1115	TIMP2_weighted, MOSPD1--2, KLK6, AIF1L_weighted, FOLR1--1	TIMP2--3, OBFC2A--3, DOK7, GADD45B--1
Gene Cluster1116	BAIAP2_weighted, IER5L--2, TIMP3_weighted, GRAMD2--1, TRAF4--4, CLIC3	CERCAM--1, DYNLL2, LYPD3, TSPAN14--1, LOXL4--2, S100A3, TMEM25--1, PTGDS--3
Gene Cluster1117	ANXA3, MYOF--1, EMP1_weighted, AHNAK--2, BCL2L1_weighted, RBMS2--2, RAPH1--2	KRT80, PPL, FRMD5_weighted, IL18, ARHGAP42--2, JUB--3, TMEM133, C13orf15--1, SH3TC2--1
Gene Cluster1119	LOC727820_weighted, SYT1--2	TMPRSS3--2, INHBB, EPB41L4A--4, TMPRSS3--3, TMPRSS3--1, SYT1--1, DNAH12--2
Gene Cluster1120	PLEKHB1--1, RAC2--2	MFI2--3, SKAP1, nan--9328, LOC404266_weighted_1, nan--5529, NES--1, SLC2A12--1
Gene Cluster1121	nan--10772, BCAR1--2, ARHGAP23--4, LOC100286993, ZNF488, OBSL1--5, JAZF1--2, LRDD--2, WNT3--1, PIP4K2B_weighted, PVRL1--3, SNCG--2, MMRN2--1, 9-Sep--2, DPYD--1, nan--1766, NUDT17--1, ADPRH--1, RNF207--2, OTUB1--2, NF1--2, VAMP1--1, RAB9B, FBXO22--1, MMP24--2, nan--945, CAPN6--2	FGD3--1, RASL10B--1, SCARNA17, FRY_weighted, VWA1--5
Gene Cluster1122	ZDHHC8P1, HSD17B2, LOC154761, ARHGAP44--1	SULF2_weighted, ROR1--2, ANK1_weighted
Gene Cluster1124	ZNF467--2	SAP30--1, NEB--1, C1orf53--1, FAM189A1, NFATC4_weighted, PRR14--1, nan--2100, ATF5_weighted, GRIP2, ZXDA
Gene Cluster1126	SMO	MAPT--2, TM7SF3_weighted, ATP8B2--1, MEX3B, ENSA--3, RAB28--1, NRGN, SLC2A11--3, FGD1, LAS1L--2
Gene Cluster1128	TCP11L2--6, nan--2558	
Gene Cluster1129	OTUD7B--4, nan--9052	
Gene Cluster1131	RNF121, MBD6--2	
Gene Cluster1135	LOC100129550, CLSTN3, SPTBN1--6	CYLD--1, SLFN5_weighted, PHF1--1, nan--5071, nan--5174, MR1--3
Gene Cluster1139	U2AF1L4, FHL3	
Gene Cluster1140	CMYA5--1, ZKSCAN3--3	

Gene Cluster1144	MERTK_weighted, GRAMD1C, ABCA1_weighted, BASP1, NMNAT2--4, BMF	
Gene Cluster1145	nan--9585, TNFRSF10C_weighted	ZMAT3--4, nan--5021, PROM2--3, BCL6--2
Gene Cluster1147	FOSL2_weighted, RIPK4--2, ZNF654--1, SMPD1--2, nan--7313, BCL2L2--1, ERBB2--1, FLRT3--2, C22orf25, ZNF557--1	CNOT4--4, SMARCE1--3
Gene Cluster1150	DDX58--1, TAP2--2, MALT1--2	USP18, IFIT5_weighted, HERC6--3, FBXO6, nan--3108, C9orf91, PMAIP1--2
Gene Cluster1153	AP1S3_weighted, MOSC2_weighted, SGMS2_weighted	SGMS2--2, CAND1--4, nan--7662, TCN1, PRKCH--1, CDKN2C--2, FAM149A_weighted, AP1S3--5, NFATC1--4, HHEX_weighted, LYSMD1
Gene Cluster1156	KLC2, SGPP1--2, EIF1--4, PNMA1, PQLC3--2, PLD6, HOMER1--1, PPM1L_weighted	AADAT, LMO4--2
Gene Cluster1157	TPCN2--1, TUBB6, TSGA14_weighted, FZD7_weighted, SNX18--2	DNAH14--4, DENND5A, LDHB_weighted, H3F3B--2, SNX6--2, nan--5139
Gene Cluster1158	FEM1C	ZCCHC11--1, MED13--3
Gene Cluster1161	SLC30A4--2, ADCY7--2, C16orf74, SERPINB9--2	
Gene Cluster1163	BAG4_weighted, B4GALT1_weighted	COL6A1--6, ULBP2_weighted, BTC--2, SOCS6_weighted, SLC7A11--1, COL6A1_weighted, C5orf33--3, DOCK5--5, BTC--1
Gene Cluster1164	GLIPR1--3, GOLGA2B_weighted, EPS15L1--4, MICALL1--1	GLIPR1_weighted, TNFRSF10D--1
Gene Cluster1165	PCNX--1, ZNF609_weighted, UNC119, PNMA2--2	RHOF_weighted, SPHK1, TP53_weighted, MT1M, PCDH7_weighted, MRAS--2, TYMP--1, ALDH1L1--1
Gene Cluster1169	ZNF512--2, ZNF655_weighted, COL27A1_weighted, MLLT11, LOC100506676_weighted, NRD1--2	GM2A--5, FRMD4A_weighted, EEPD1_weighted, nan--1041, PHLDB3--1, BMP6--1, TUBD1--3
Gene Cluster1173	COCH--3, C20orf117--3	SYNM--1, RRAGD--2, NR3C1_weighted, COCH--2, HTRA1
Gene Cluster1176	NCRNA00173--2, PALM, CCDC3, GNAZ, PRAME, nan--5366, CHST11_weighted	GABBR1, NCRNA00173--1, CCDC46--2, NDRG4
Gene Cluster1177	ZNF268--1, ZNF850, ZNF264--3, ZNF329, nan--10741	ZNF284, ZNF264--2, ZNF585B--1, ANKRD20A1 /// ANKRD20A2 /// ANKRD20A3 /// ANKRD20A4 /// C21orf81 /// LOC100132733 /// LOC284232
Gene Cluster1178	ZNF226--4, ZNF268--2, RSF1--5	
Gene Cluster1179	ZNF420--2, ZNF529--2, ZNF615, ZNF419--1, NINL, ZNF302--3, BEX2, SPAG16--2, ZNF550--2	ZNF559, ZNF75A--1, ZNF331--1, GPRASP2

Gene Cluster1180	LARP6--2, ARMCX3_weighted, PRKACB--1, MYH10--1, NGFRAP1, ARMCX6, ZNF673, HLTF, STC2_weighted, KDSR_weighted, TDRKH--3	ZNF274--1, PRKACB--2, GNAQ_weighted, IFT57--1
Gene Cluster1181	FRMD4B--1, KDM5A--5, ZBTB10_weighted, ZNF43--1	GYG2_weighted, LOC283357, nan--10698, RHOV
Gene Cluster1185	DEK, CEBPZ, CEP57--4, OPA1--1, LOC100507328, MAP3K2--5, FMR1_weighted, UBR1--1	
Gene Cluster1189	RPL13P5, SLC2A6, ANXA6, KLHL29--4, MED7--2	CRIM1--2, FKBP1B, NDUFS6, DCP2--4, nan--5950
Gene Cluster1194	GATA2--2, POT1--2, SAV1--1, DDX59--2, SUPT3H	FBXO3--5, FAM178A--2
Gene Cluster1196	TFPI_weighted_2, ARSJ, KCTD12_weighted, GNAI1_weighted, MIB1--3, CACHD1--1, PSD3_weighted, EFCAB4B--3, ZNF136	DOCK4--2, ING3--3, MUM1L1, EPHA4_weighted, FAM132B
Gene Cluster1197	SEPP1--1, PPAP2B--1, C17orf108--2	GATS--1, LOC100507376--1, FBXL20--3, DDHD1_weighted, DNMT3A--1
Gene Cluster1198	FAM91A2 /// FLJ39739 /// LOC100286793 /// LOC728855 /// LOC728875, FYN_weighted, PARP8--1, CDC42EP3_weighted, SRGAP2P1--3, CHD6--2, PLA2G15	BNIP2--2, SRGAP2P1--4
Gene Cluster1200	B2M_weighted, C1GALT1--1, nan--5363, STIM1, ASPH--6	ZNF827--2, ZNF827--5, TMX4_weighted, ITPKA
Gene Cluster1201	ELK3--2, RABGAP1L--3, RAB27B--3, PAM_weighted, IFIH1--3, RHOQ--6, ZXDB--2, NR1D2--2, FKBP14--1	
Gene Cluster1203	ZNF449--1, RNF146--1	CYLD--3, ARMCX5--1, TTLL1--2
Gene Cluster1208	MDGA1--1	ADIG--1, PDGFA--2, LAT /// SPNS1_weighted
Gene Cluster1210	KIAA0196, FAP, NSMCE2	SIPA1L2--1, CPLX2--2, FLYWCH2, STMN3--1, LOC644172 /// MAPK8IP1, SCN1B, MAN1C1--2
Gene Cluster1211	ITGA1--2, TENC1, ITGA7--1, VAMP5--1, TIAM1--1	LAPTM5_weighted, PRKCA_weighted, CTSB--1, CD58--5, HDAC4_weighted
Gene Cluster1212	SPON2, ITGA7--2, MAOB--3	nan--5561, MAFB_weighted, CHRNA3--1, nan--5906, MGC16075, nan--9320
Gene Cluster1213	LOC100505592, EFEMP2--2	S100A5, MEF2A--1, LETM2--1, RAB6B--1, KIAA1632--5
Gene Cluster1215	ZNF704--6, nan--7992, HIVEP3--4, nan--5733, FAM20C--2, IGFN1--2, SLC35F3--1, nan--5237, nan--5022, nan--10913, CHST6, BIRC7, PCDHGA3--3, PCDHGA10 /// PCDHGA11 /// PCDHGA12 /// PCDHGA3 /// PCDHGA5 /// PCDHGA6, SYT11--1, KCNB2--1, MEF2C--1, CDON--1, KCNMA1--2, EFHC2--2, FBXO27, GRK6--1	

Gene Cluster1217	RSPH3--2, METTL9--3, nan--5522	AK5--2, KSR2--2, LOC399959, MCC--2, C16orf52 /// LOC100509359, C19orf51
Gene Cluster1218	WNT5A--3, C5orf41_weighted	PPARD--1, nan--5673, PAX6--2, HLA-DPA1_weighted, KRT86 /// LOC100509764, KLHL24_weighted, FLJ32255_weighted, nan--5361, ABTB1--1, SLC27A1
Gene Cluster1219	FAM40B--1, LOC728192--1, IGKC--3, DUS4L--2, LOC648987, OXGR1	OTUB2--3, SYNPR, nan--8059
Gene Cluster1220	nan--5481, HAS2_weighted, DSC3_weighted	RRAGB, LIN7B--2, MUDENG--1, LAMA3--3, MYNN--3, FBLL1, nan--2456, DKFZP434K028--1
Gene Cluster1221	ITPR1--1, XRCC4--3, IGSF5, C16orf52--3	
Gene Cluster1222	nan--630, FAM175B--2, ZNF280A, HRASLS--2, IL21R--2, ANUBL1--1, HINT3--1, nan--8130, nan--8811, nan--10247, RASGEF1A--2, nan--11729, DFNB31--2, UTS2_weighted	
Gene Cluster1224	KAP2.1B /// KRTAP2-2	PPARD--3, LOC100131067--1, EIF1--6, nan--8738, LOC100506090, MLL--11, C19orf76
Gene Cluster1226	nan--11946, DSG4, RIMBP2--4, RGS10--1, PHC2--2, TACR2, GPR15	
Gene Cluster1227	RCN3--2, nan--8540, nan--7658, ATP6V0A2--6, HRASLS5_weighted, LOC100124692 /// LOC93432, ZNF556, KCNJ3--1, HOXD9--1, SERPIND1, CLDN10--2, STK17B--2, C1orf21--1, PCDHB10, PDE10A--1, DNAH2--2, THBS2, CYP2E1--2	
Gene Cluster1228	nan--11907, nan--10300, nan--10209, nan--9786, nan--9644, nan--7612, HAPLN1--4, nan--5261, nan--5243, GABRB3_weighted_1, MEAF6--2, LOC282997, MUTED /// TXNDC5, FGGY--1, TRPA1--2, HERC2P3, NR2F2--1, GFI1, ACADL--1, LRRC8C--1, HECTD2--3	
Gene Cluster1229	PLCL2--3, FGFBP2	PTH2R, NRTN, GPX7, MUC6, FOXL2, TMEM107--2, GALNT13--1
Gene Cluster1230	FGFR2_weighted, DAPK1--1	PDE5A--7, EGLN3--2, FGFR2--2, EGLN3--1, nan--11644
Gene Cluster1232	FAM46C--2, PLAGL1_weighted, SOX2_weighted	PLAGL1--2, FAM46C--1, nan--11631, nan--11024, PIK3AP1--1
Gene Cluster1233	FLCN--1, CA8--2, FAM46A--2, PRDM13, GPR27--2, NPL--1, NR2F2--5	RWDD2A, ADCY2--1, nan--5175
Gene Cluster1235	CEPT1--2, PCDHB18, KCNRG--1, REG1B, SH3GL2, CPS1--2, PFKFB2--2, TRIM7--1, TUBBP5	DCDC2--1, LOC401074_weighted, TRIM36--3
Gene Cluster1236	CCDC138, L2HGDH--1, SNHG5 /// SNORD50A /// SNORD50B, SOD2--4, TRIM59_weighted	OIP5, HEATR3--1, C21orf45--3

Gene Cluster1237	SETMAR--3, CDC25A_weighted, CDC25A--1, UBE2N--2, CASC5--1, VPRBP--5, NFYB--2, HAUS6--1	PRDX2--4, TAF5--1, MFSD8--2
Gene Cluster1240	ATP5B, LEO1, EXOC5--5, MBIP	RGS2, PIGA, TMEM70--2
Gene Cluster1242	TRIM25--2, ASH1L--3, ADAL--2, EIF2AK2--3, ACVR1C--2	
Gene Cluster1243	ENPP1--2, HAUS5--2, LAMA2--2	ARHGEF26--2, GABRB3_weighted_2, CACNB2_weighted, ZNF131--4, nan--475
Gene Cluster1246	MGA--4, KDELC1	ZNF721--2, ANKRD37--1, MLLT10--1, GDAP2--1, CMBL--1, nan--12194, nan--12188
Gene Cluster1249	MDH1--1, NMRAL1_weighted	
Gene Cluster1251	GLS2--2, TFR2--2, C9orf125--1, KCNAB2--2, ABHD11--1, TMEFF2--2, METTL12	C9orf125--2, TFR2--1
Gene Cluster1252	FKBP5_weighted, FANCA--1, BEND7--2	FKBP5--2, LRP3--1, MTERFD3--1, MESP1--1, APOOL--2
Gene Cluster1253	SLC1A1_weighted, FOXA1--1, TBC1D30_weighted, C1orf115--1, CYP1B1_weighted	SLC16A14, SMPDL3B--2, IMMP1L--2, SH3RF3
Gene Cluster1254	GDI2--2, MLLT10_weighted, PITX2, STAT5B--3	CASC5--2, CPT1A_weighted, LOC401022--1
Gene Cluster1256	MTMR4--1, RPS2--2, REST--2, AMFR--2, RNFT1_weighted, EPC2, SFT2D1--3, DNAJC16--2, USP31--4	TNFSF9, CA2, ECH1
Gene Cluster1257	STOX1--2, GXYLT2--1, MBD4--4	GXYLT2--2, nan--4850
Gene Cluster1258	GTF3C6, UBXN8, KPNA5--2, FDXR, USP48--2, ME2_weighted, NDUFA6--2	HDDC2--1, TMEM69, THEM4_weighted, KIF26A--2
Gene Cluster1259	CSTF2T--1, SAR1A--1, CSTF2T--2, RPL37A--2, AP3M1--2, TFDP1--4	NT5C3L, TIMM13--1, TMEM183A, nan--6170
Gene Cluster1260	WDR7--1, TMED5--6, ZNF271--1, C6orf162--1	PGM3--2, VAMP7
Gene Cluster1262	EIF3L--1, MAPKSP1--2, LOC100507620--1, ATP6V1E2	C3orf23--3, nan--10719, IMPACT--1, LOC100510093, DYNC2LI1--3, MBD1--5
Gene Cluster1263	RIC8B--2, nan--5454	
Gene Cluster1268	nan--7493, nan--2800, RRP15--1, RPL27A--3	AGPAT5--2, nan--7891
Gene Cluster1269	ATF7IP2, KREMEN1--2, nan--5502	LOC100287628, SLC16A10--2, LYPD6--2
Gene Cluster1271	ZADH2--5, HAUS2--2, IKZF5--2	nan--10994, KIAA1715--1, SLC2A13--3, MATR3--5, TSFM--2, DIXDC1--3, nan--5109, nan--5336

Gene Cluster1275	ADRBK2_weighted, C5orf44--2, BARD1--3, G2E3--4	GTF2H2B--2, RIOK2--2, nan--5137
Gene Cluster1276	SRI--1, PLA2G4A, ELOVL6--3, KIAA1632--3	ELP2--1, nan--7325, nan--7229, MAPRE2--3, TPR--5, RGMB--4
Gene Cluster1279	HOXD13--3, MCPH1--4, BMP2K_weighted, TMEM38A, nan--5458	SYCE2, WISP3--1
Gene Cluster1280	CDK6_weighted, nan--8853, CTDSPL2--4, TMEM182--2	CDK6--6, MRPL19--3, LACTB--2
Gene Cluster1281	C12orf43, TAF4B--1, SCLT1--1, MDM4--4, HS6ST2--1, GNPTAB--3, nan--1974	
Gene Cluster1285	C3orf64, C4orf41--1, VPS37A--2, PLAA--3, IER3IP1--1	ENPP4--2, ACTR8--1, MOCOS--2
Gene Cluster1288	ARL5B--2, PUS7L--2	METTL10--3, ESF1--2, CCDC42B--1, URB1--2
Gene Cluster1290	FBXL4--1, NBN--4, SHPRH	UHRF1BP1L--3, METAP2--2, ATP2C1--2
Gene Cluster1291	CCDC75--2, MAPK8--5, BNIP1--3	GMCL1 /// GMCL1L, SPAST--2
Gene Cluster1294	KRTAP9-8, KNDC1--1, nan--5461, ZMYND12, nan--4887, KRTAP1-1, C10orf110, C14orf45--1, MT4, nan--3954, SELPLG--1, NDEL1, ALOX12B, CD180, AVIL--2, LOC283404, ZBTB26--1, MFSD2A--1, LPO, nan--9253, TRPV1, PCDH7--2, ZNF354B, NKX6-3--2	
Gene Cluster1295	C1orf110, COMMD3, RAB7B--2	DDX3Y--4, KRT6A /// KRT6B /// KRT6C, nan--9148
Gene Cluster1296	LOC100510507, LOC219347, KCNQ1OT1_weighted, ZDHHC21--9	nan--7498, KBTBD8, CD99P1
Gene Cluster1298	KRTAP3-3, KRTAP1-3--2, SLC6A13--2, SEMA3F--1, MICALCL	PSORS1C2, DENND2C--1, TRIM11--3, C3orf62--2, SNCAIP--1
Gene Cluster1300	ZNF224--3, nan--6166, CTSK, ZBTB3--1, CRISPLD2--2, RASD1	nan--8809, KRT33A, LOC100272228--3, ZFYVE20--1
Gene Cluster1302	TPRG1, CYP4B1_weighted, USP9Y--2, nan--11742	
Gene Cluster1304	MGC21881--1, MST1P9--1, DNAL1--3	NEK11--3, FAM134A--4, DNAL4
Gene Cluster1306	ZNF187, LOC100287515 /// ZNF26--1, ELL2_weighted, nan--10712, ZNF260--1	ARRDC3, C16orf52--1, CARD8_weighted, LOC100289294--1, B4GALNT3_weighted, TLE1_weighted
Gene Cluster1308	FBXO16, nan--5136	PIK3R3--2, CSAD--2, POLI--1, SLC30A7--3, LOC400931--3
Gene Cluster1312	C11orf35, LOC645722--2, nan--5186, HGSNAT--4, STARD9, PDHX, SLC1A2--4, LOC692247_weighted, USP7--4, CARD14--6,	LOC338651--1, LOC100129637--2

	ANKRD19, nan--898, MKX_weighted, ZC3H6--1	
Gene Cluster1314	LOC440028--1, PYGM, PRDM5--2, C3P1, nan--12187, nan--12189, CYP2R1--1, MYH4, SCAND3	
Gene Cluster1315	ERICH1 /// FLJ00290	MIB2--2, RPAP2--1, nan--12159, LARP1B--5
Gene Cluster1317	NRN1, SCG5, MIA3--3, SMAD9--2, UGT2B15--1, DHRS2_weighted, FAM43A--1, CPE_weighted, BCOR--4, MMAB--2	TOX--1, SAMD10_weighted
Gene Cluster1318	KLHL28--2, ZKSCAN4, SYNJ2BP--3, nan--2382, FBXW7--3	
Gene Cluster1319	FLJ27352, LOC100506719	CCPG1--2, CCPG1_weighted, DNAJB9_weighted, nan--7455, CCPG1--5
Gene Cluster1320	ZNF323, MIB2--1, CAMK2D--5, nan--6101, C15orf37	PRKAB2--2, RNF125_weighted
Gene Cluster1321	CYP2C9_weighted_1, nan--9205, DPY19L2--2, nan--7969, DNALI1_weighted, RAB37--1, nan--1746, nan--960, nan--749, CTAGE5--2, LARP1B--4, LOC338758, TRIL_weighted, MARK1--3	BTBD9--4, GPD1--4
Gene Cluster1325	DSG3--2, WIPF2--4, XAB2, NUDT16--1, ATHL1, LOC389831--3, RFFL--2, DCT--2	DSG3--1, ZCWPW1--2
Gene Cluster1327	FAM164C, MXD1--1, PLEKHA4	
Gene Cluster1328	ZNF277--2, CEP63--2, DDX17--1, FAM27E3, LPP_weighted, C3orf19--1, CDC42SE2--4, TMEM154, G0S2	HBEGF--4, DDX17--3, CHRFAM7A /// CHRNA7
Gene Cluster1329	SPRR1A_weighted, PSCA, LOC100506979, C8orf4, RND1, LOC100287896	C9orf24, KLHDC7B--1, HDAC8--2, SPRR1B, FICD, ERLEC1--2, SULT1C2--2, PRSS1, NUMA1--1
Gene Cluster1330	UNC5CL, BTBD9--2, ATP6V0A4, CXCR3--2, LPXN--1, LRCH1--1, GUCA1B--2, C6orf64_weighted, C6orf130--1, CALML3--2, JAKMIP3, TBPL1, nan--7250, TPSB2, CALHM3, PDE8B--1, MYT1--1, TNFSF4, SPN--1, LOC728192--2, HCN1_weighted, MED20--1, NFYA_weighted, GMFG, LOC643201--2, LOC100507253, ZBTB7C--1, SYT13--1, DGCR5, JAG1--1, LSAMP_weighted	FIBCD1--2, SERPINA4, GPER--2, RARRES2, S1PR4, HGD--2
Gene Cluster1331	nan--7801, nan--7838, NFRKB--4, MEOX2--1, FECH--3, GJB6	
Gene Cluster1335	ZAP70_weighted, PSG3_weighted, nan--9176, BTBD11--2, nan--8251, nan--8095, FAM179A, nan--5467, FYTTD1--1, IGF1--4, PSG9--2, ITGB5--4, PSG4--2, PSG7, GLI3--1, KIAA1755--1, LOC643763--2, nan--362, C17orf52, PSG1--1, KRT74_weighted, CLSTN2, CD8B--1, FOXF2,	

	C11orf31--2, PRF1_weighted, LIMS3 /// LIMS3-LOC440895 /// LOC440895_weighted	
Gene Cluster1336	LMO2, FAM122C--4, nan--5093, VSIG10L, EN2, MBLAC1	SCN5A, MRVI1--2, ZBTB46--1, KRT71, ULK4--3
Gene Cluster1339	nan--5300, KIAA1549--3, MEX3A--1	
Gene Cluster1340	APH1B--2, PXMP4--2, ETV5--4, HSPA1A	TGIF2--2, FKBP10, LOC100131199_weighted
Gene Cluster1341	nan--9576, nan--979	ARHGEF7--6, HNRNPU--7
Gene Cluster1342	ZXDC--3, SLC7A8--5, LOC100506057 /// STK32C, LOC100506922	SLC7A8--3, SNAI1, DACT1, SERP1--4, FZD10
Gene Cluster1343	RGNEF_weighted, NAP1L4--2, RNASEL_weighted, API5--4	ALKBH3, QSER1--2
Gene Cluster1344	OGDH--2, NCRNA00116, WLS--3, nan--7280	THADA--2, DLG1--3, PHOSPHO2--1, PTPN12--3
Gene Cluster1345	IGF2R--1, HSPH1--1, CRIM1--3, DYNC1LI2--1, TAF15--3, VEGFB, MDM4--6, nan--5973, NFKB2--1, SEC31A_weighted, CTSB--5	HDAC11--2, PRR12
Gene Cluster1346	IWS1--2, FBXO21_weighted, PPP1R10--1, ORC2--1	
Gene Cluster1347	GPATCH2_weighted, FMNL2--1, CAMK1D /// LOC283070_weighted, DOCK1--1, KIAA1958--1, LOC440944--4	COMMD7--1, RPS24--1
Gene Cluster1348	CLP1--3, SLC35F2, ZBTB38--3, BAIAP2L1--2, SNX9_weighted	
Gene Cluster1354	nan--3510, NPY1R, CNTN1--2, THSD7A--1, PRKAA2--4, VWDE--2, nan--10924, ATXN1--1, MAP2--1, FAM49B--3, nan--8679	
Gene Cluster1355	CRABP1--2, COBLL1--4, L3MBTL4--2, TCF12--3, SPINK7	ACHE--1, DDIT4L, LOC100506797
Gene Cluster1357	RHOBTB3--4, PDS5A--4, TNFRSF18--1, ZDHHC2--4, LOC100506638, RPP30--1	SNAP29--1, ATL3--2, FGF19, ADK--2, ARAP3--1
Gene Cluster1358	SDS--2, CDC42--3, ZXDA /// ZXDB--2, HSD17B6_weighted	
Gene Cluster1359	NDUFAF3, PRTFDC1, C10orf125, LOC100128881	RAB3IP_weighted, KCTD14_weighted, ROCK2--1
Gene Cluster1362	OSR2, LIMCH1_weighted, KCNS3	FNDC3B--6, SYTL4--2, PTK2--4, TSPAN2--3, PCGF5--5, nan--1281
Gene Cluster1363	ZIC1--1, CDH12, CALCRL--2, MDFIC--3, LOC100287275 /// PCOTH, SCGB3A1, SYT8, nan--9131, BMPR1B--3, GALNT13_weighted, MIA3_weighted, DLG3--4, SLITRK6--3	

Gene Cluster1364	LPAR3--2, HEYL--2, KCNH3, SEMA6B--2, ABHD10--1, TUBB4, SYN2--2, VCAN--1, KCNQ2--2, PCDH7--3, PRRX2, EML1--2, GTSF1, ERI1--2, RBM24, F2R--2	
Gene Cluster1367	MTX3--1, OTUD5--3, BRD4--2	ZNF503--1, ZNF784
Gene Cluster1371	NUDT17--2, TMCC2--2	
Gene Cluster1373	nan--11472, EID2B, SRFBP1, nan--9217, PL-5283 /// SLC13A4, SPOCD1--2, LOC728485--1, ZNF438--1, B3GAT1, DUSP7--2, HGF--2, ANGPTL4--2, DMRT2_weighted, WDR85, IQCG--2, NKX3-1--2, CD68, COBRA1	
Gene Cluster1374	PRDM11--2, ZMIZ2--3	
Gene Cluster1376	nan--5402, LOC284454--2	
Gene Cluster1379	BTN3A2 /// BTN3A3, CHKB-CPT1B /// CPT1B--1	BTN3A2_weighted, BTN3A1--2
Gene Cluster1386	WBP5, HLA-A_weighted, NEAT1--1, SEPSECS_weighted	RHBDD1--3, LOC100132288 /// LOC100233156 /// MAFIP, THSD4--2, CAMK1D--2, PML--1
Gene Cluster1388	ZNF540--3, C21orf81, nan--9870, nan--8832, nan--8702, PID1--3, nan--8066, nan--7319, GSTT1--2, nan--5491, TM4SF18--2, KHDC1, ZNF527--2, LOC100132815--4, OBSCN--2, C11orf68, FLJ31715, TLE6--2, nan--576, C21orf15--2, ZNF519--2, ALS2--2, TGFB2--1, FOSL2--1, SLC4A4--5, nan--4091, nan--4606, OBFC2A--1, KCNN2, TLR5, nan--7275, nan--9099, SEMA3A--3	
Gene Cluster1389	LOC285735	CNTNAP2--4, IQGAP3--1, HPSE--2, TXNDC9--1, FUT10--1, TLK1_weighted
Gene Cluster1392	APOL1, ZC3HAV1--2	nan--5824, FZD4--1, CCL5_weighted, LOC100289361, MIR155HG, HIST1H2BC, MTCP1 /// MTCP1NB, EPM2A--2
Gene Cluster1395	YTHDC1--3, AGAP10 /// AGAP4 /// AGAP9 /// BMS1P1 /// BMS1P5 /// LOC399753--2	
Gene Cluster1396	LOC100128071, IDH1--4	AHI1--5, nan--10693
Gene Cluster1398	C1orf51, TMEM220	CHST4, nan--5110, ATP8A1--2, nan--5864
Gene Cluster1399	TMEM231_weighted, nan--5575, FAM120C--2, BEX5, ZNF280B--3, C7orf63--2	LOC100289098, nan--182
Gene Cluster1400	IFNAR1--4, KLHDC7B--2, LOC100127891, nan--7310, nan--5626, nan--5607, C17orf44, LOC100288954, ITLN1, GPR63, PHF7--1,	

	HNRPLL--1, C11orf21--2, NXF3--1, CYTIP, TSPAN32_weighted	
Gene Cluster1402	C19orf23--1, LOXL3--1	
Gene Cluster1407	DNAJB12--1, SH3BP2--3	SLC24A6--2, APAF1--3
Gene Cluster1408	LOC100506661, FAM201A--1, RAD51L3--2	nan--6261, FAM76B--2, LOC389906--2, C6orf52--2, nan--4595
Gene Cluster1409	PCGF3--3, LOC100287482	RAD52--1, FZD3--1, ZNHIT2--2, COL7A1--2, ALDH5A1--2
Gene Cluster1410	NSD1--2, VAMP2--3	C17orf65, WDR61--1, C17orf70--2
Gene Cluster1414	IGFL2, DMRTA2, nan--5322	ZDHHC3--4, nan--8799, ODAM, PLA2G3, SPTLC3--2, NFIB--7, CDH7--2, CIDEB
Gene Cluster1416	STRN3--2, CDC42SE1--3	HNRNPL_weighted, MTA2, SMU1--1, TMEM33--5, USF1, BOD1L--3, RAB12--1, TOR1AIP2--4, CTNNB1--1, LYG1

Supplementary Table S3 Full List of Gene Cluster Assigned to Each Cluster

	Gene Cluster ID Associated That Are Not Inclusions	Gene Cluster IDs That Are Inclusions
Cluster3	15, 65	1, 25, 4, 71, 72, 69, 70
Cluster4	16, 13, 17, 22	12, 5, 18, 20, 19, 439, 8, 440
Cluster6	28, 696	27, 408
Cluster9	14, 51, 473	44, 37, 474
Cluster13	13, 601	
Cluster15	53, 469	
Cluster19	2, 3, 39	
Cluster22	22, 48	
Cluster24	60, 286	55, 61
Cluster32	597, 391	640, 122, 116, 705
Cluster34	114, 434, 649	
Cluster36	75, 87, 76, 92	74, 145, 77, 89, 84, 202, 137, 214, 125
Cluster38	93, 110, 94	220, 218, 98
Cluster42	148, 91	150, 88, 152, 153
Cluster73	265, 269, 1410, 1129	
Cluster82	1291, 1290	852, 754, 730, 1288, 1281
Cluster83	1416, 1251	1153, 1007
sssCluster87	765, 742	729, 907, 732, 737

Cluster88	781, 782, 1344	
Cluster91	91, 813, 815, 818	
Cluster92	1196, 1203	881, 873, 1135
Cluster93	809, 856, 863, 880, 862	
Cluster94	837, 812	888, 1083
Cluster97	871, 874, 891	
Cluster98	890, 1116, 835	869, 893, 895, 841, 968, 1198, 894
Cluster101	900, 889	901, 903, 746, 753, 915, 908, 916
Cluster102	1163, 1150, 1161, 1139	1107, 1165, 1164
Cluster103	910, 845	909, 923, 736, 846
Cluster104	913, 956	1013, 816
Cluster114	395, 404	
Cluster115	402, 103, 131	
Cluster118	1237, 1236, 1252	1246, 1269, 1268
Cluster119	1271, 1253, 1254	804, 615, 1263
Cluster133	1077, 1079, 1081, 1078, 1080	
Cluster134	134, 1110	
Cluster151	151, 1040	
Cluster152	1011, 1008, 1009	
Cluster159	762, 343	

Clusasster163	865, 1340, 949	951, 843
Cluster169	1072, 1071, 1319	1304, 1197, 1073, 1070
Cluster172	172, 1062	
Cluster175	175, 961, 963	
Cluster178	178, 1044, 1048	
Cluster179	929, 179, 572	

Supplementary Code S1 Code for the GMMchi simulation tests

```
#validation, simulate data to test #1

%run GMMchiV4.ipynb

import random
two_groups = []
tails = []

#retrieve categories
categories_all = []

for spread in tqdm(range(2,15)):
    mean = 10
    std = 2
    size = 40
    spread_factor = spread #how many std is the spread of tail
    mult_factor = 10 #so we can use more simulated data between two numbers


    z = np.random.normal(mean, std, size)

    tail_size = int(size*0.1) #how many samples in the tail
    tail_spread_lower_limit = mean+3*std #multiple by 10 and divide it back down
below

    tail_spread_upper_limit = tail_spread_lower_limit + std*spread_factor

    #create simulated data
    simulate_df = pd.DataFrame()


    iterations = 5000
    for i in range(0, 5000):
        #generating a random unimodal distribution with a tail
        tail = [x/mult_factor for x in random.sample(range(tail_spread_lower_limi
t*mult_factor,tail_spread_upper_limit*mult_factor), tail_size)]
        data_tail = np.concatenate((z, tail))

        simulate_df['simulate_{}'.format(i)] = data_tail

    #transform the dataframe
    simulate_df = simulate_df.T
```

```
    categories_ = []

    for i in tqdm(simulate_df.index):
        info, classif, categories, chi, bins, f = GMM_modelingt(i, simulate_df, l
og2transform=False,
                                                                graphs=False,verb
osity=False, unimodal_categories =False)
        categories_.append(classif)
        del info, classif, categories, chi, bins, f #free memory

    categories_all.append(categories_)
    #how many ones with two groups
    two_groups.append(sum([1 for x in categories_ if x =='unimodal' ])*100/len(ca
tegories_))
    #how many with tails
    tails.append(sum([1 for x in categories_ if 'tail' in x ])*100/len(categories
_))


#validation, simulate data to test #2

%run GMMchiV4.ipynb

import random
two_groups_bi = []
tails_bi = []

#retrieve categories
categories_all_bi = []

for spread in tqdm(range(2,15)):
    mean = [5,10]
    std1 = [2,0]
    std2 = [0,2]
    size = 40
    spread_factor = spread #how many std is the spread of tail
    mult_factor = 10 #so we can use more simulated data between two numbers

    x,y = np.random.multivariate_normal(mean, [std1, std2], size).T
    data_bi = np.concatenate((x,y))

    tail_size = int(size*0.1) #how many samples in the tail
    tail_spread_lower_limit = mean[1]+2*std2[1]
    tail_spread_upper_limit = tail_spread_lower_limit + std2[1]*spread_factor
```

```
    #create simulated data
    simulate_df = pd.DataFrame()


    iterations = 5000
    for i in range(0, 5000):
        #generating a random unimodal distribution with a tail
        tail = [x/mult_factor for x in random.sample(range(tail_spread_lower_limi
t*mult_factor,tail_spread_upper_limit*mult_factor), tail_size)]
        data_tail = np.concatenate((data_bi, tail))
        simulate_df['simulate_{}'.format(i)] = data_tail

    #transform the dataframe
    simulate_df = simulate_df.T

    categories_ = []

    for i in tqdm(simulate_df.index):
        info, classif, categories, chi, bins, f = GMM_modelingt(i, simulate_df, l
og2transform=False,
                                                                 graphs=False,verb
osity=False, unimodal_categories =False)
        categories_.append(classif)
        del info, classif, categories, chi, bins, f #free memory

    categories_all_bi.append(categories_)
    #how many ones with two groups
    two_groups_bi.append(sum([1 for x in categories_ if x =='unimodal' ])*100/len
(categories_))
    #how many with tails
    tails_bi.append(sum([1 for x in categories_ if 'tail' in x ])*100/len(categor
ies_))
```