Myofibroblast Migration and Signaling in *in vitro* Colorectal Cancer Models

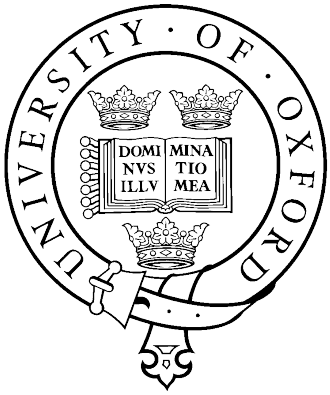
Peter Kalugin

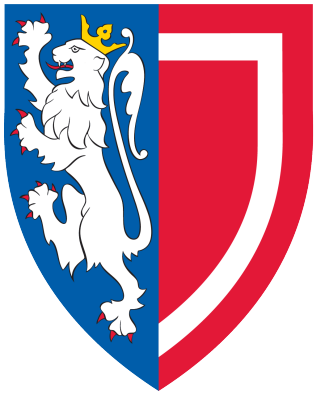
Balliol College

A thesis submitted to the Division of Medical Sciences, University of Oxford, in partial fulfillment of the requirements for the degree of Master of Science by Research in Oncology

Trinity Term, 2017

Weatherall Institute of Molecular Medicine
University of Oxford





I dedicate this thesis to my parents and my sisters.

Thank you for your unwavering support and encouragement, and for always challenging and inspiring me to do my best.

(Word Count: 24,909)

Abstract

Colorectal cancer is one of the most prevalent, and one of the deadliest, types of cancer in the world. The colonic epithelium is positioned atop a subepithelial structure consisting of extracellular matrix and stromal cells, and colorectal tumors closely interact with this tissue. Myofibroblasts make up the primary cellular component of the subepithelial stroma in the colon, and play important roles in the maintenance of the normal colonic stem cell niche. Increased recruitment of local myofibroblasts into the tumor bulk is a feature of some primary and metastatic colorectal cancers, and is associated with poor prognosis.

The first aim of this thesis was to probe the nature of the bidirectional interaction between colorectal cancer cells and colonic myofibroblasts. Two *in vitro* assays integrating myofibroblasts, colorectal cancer cells, and extracellular matrix were devised to mimic the *in vivo* arrangement and communication between these cells. Using these assays, it was shown that myofibroblast-derived hepatocyte growth factor reduces differentiation in a subset of colorectal cancer cell lines (**Chapter 3**). Additionally, colorectal cancer cells were shown to recruit myofibroblasts by secreting ligands of the platelet-derived growth factor family (**Chapter 4**).

A secondary aim of the project was to develop an analytical technique to robustly categorize colorectal cancer gene expression profiles in an unsupervised manner. Gaussian Mixture Modeling was used to generate categorized expression data for an in-house colorectal cancer cell line panel, as well as a publically available dataset from primary tumors (**Chapter 5**). 2x2 statistical analyses on the categorized data confirmed known genetic associations in the cancer cells, and also uncovered characteristic gene expression signatures corresponding to intratumoral heterogeneity and stromal contamination (**Chapter 6**).

This project identifies several mechanisms responsible for the heterogeneity observed in colorectal cancer behaviors, and provides tools to probe these characteristics further both experimentally and analytically. (297 words)

Acknowledgements

I would like to sincerely thank everyone who has helped me make this thesis a reality. First of all, none of my work would have been possible without the generosity and guidance of Professor Sir Walter Bodmer. Thank you for welcoming me into your group and providing me with the tools and the support that I needed.

I am also extremely thankful for all the members of the Cancer and Immunogenetics Laboratory. Thank you especially to Jenny Wilding and Djamila Ouaret for your close support and supervision, and for helping me make the right experimental decisions. Also huge thanks to Laura Colling, Haoyu Liu, Marahaini Musa, Tom Barnes, Shuhui Liu, Mustak Ibn Ayub, and Jens Puschhof. You are all extraordinary scientists and extraordinary people, and it was a privilege to work with you.

Additionally, I would like to thank Professor Peter McHugh for acting as my second supervisor, and to Professors Ruth Muschel and Eric O'Neill for a very fruitful discussion about my work during my transfer of status.

Thank you to Charles and Mary, and to the rest of the Rhodes community, for welcoming me into a family of brilliant and wonderful people that I will never forget. Thanks also to the Rhodes Trust for funding my work, and to Balliol College for giving me another home in Oxford.

Finally, I thank my friends, especially Anisha and Sarah, for their support, and of course Amrita for always being there for me.

Author Declaration

I, Peter Kalugin, 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

Title Page

Dedication i

Abstract ii

Acknowledgements iii

Author Declaration iv

Table of Contents v

List of Figures x

List of Tables xiii

List of Abbreviations xiv

Chapter 1: General Introduction 1

1.1 The Human Colonic Epithelium: Architecture and Physiology 2

1.2 The Normal State: Myofibroblast Roles in Healthy Tissue 5

1.2.1 The Definition of Myofibroblast 5

1.2.2 Myofibroblasts in Crypt Maintenance 6

1.2.3 Myofibroblasts in Wound Healing and Fibrosis 8

1.2.4 Myofibroblasts as Immune Mediators 12

1.3 The Deregulated State: Myofibroblasts in Colorectal Carcinoma 14

1.3.1 Malignant Transformation in Colorectal Carcinoma 15

1.3.2 Myofibroblast Roles in Precancerous Adenomas and Primary Tumors 17

1.3.3 Myofibroblast Roles in Tumor Invasion and Metastasis 19

1.4 Genome-wide Expression Analysis in Cell Line and Tumor Panels 22

1.4.1 Analysis and Clustering of Continuous Data 22

1.4.2 Discretization of Continuous Data 25

Chapter 2: Methods and Materials 28

2.1 Cell Culture Methods 29

2.1.1 Cell Lines 29

2.1.2 Cell Culture Conditions 31

2.1.3 Cell Line Cryostorage, Preservation and Recovery 32

2.1.4 Cell Counting 32

2.1.5 Conditioned Medium Harvesting and Usage 33

2.1.6 Transient siRNA Transfection 33

2.1.7 Matrigel Coculture 35

2.1.8 Drug Screening 37

2.2 RNA Methods 38

2.2.1 RNA Extraction 38

2.2.2 Reverse Transcription 39

2.2.3 Quantitative Real-Time Polymerase Chain Reaction (qRT-PCR) 39

2.2.4 Nucleic Acid Quantification 40

2.2.5 Microarray Gene Expression Analysis 41

2.3 Microscopy and Image Analysis 42

2.3.1 Fixation and Immunostaining 42

2.3.2 Epifluorescence Imaging 43

2.3.3 Image Analysis 44

2.4 Statistical Analysis 45

2.4.1 Continuous Gene Expression Analysis 45
2.4.2 Categorical Association Analysis 45

Chapter 3: Myofibroblast Effects on Colorectal Cancer Proliferation, Morphology, and Differentiation 46
3.1 Introduction 47
3.2 Results 49
3.2.1 Cocultures on a Flat Layer of Myo/fibroblasts 49
3.2.2 Immunostaining of Mono and Cocultures 57
3.2.3 Colorectal Cancer Cell Lines in Myo/fibroblast Conditioned Media 63
3.2.4 Drug and siRNA Testing of the Candidate Genes 73
3.3 Discussion 109

Chapter 4: Myofibroblast Recruitment by Colorectal Cancer Cells 113
4.1 Introduction 114
4.2 Results 116
4.2.1 Myofibroblast Recruitment Assay Design and Optimization 116
4.2.1.1 Assay Design 116
4.2.1.2 Assay Optimization 118
4.2.2 Myofibroblast Recruitment Assay Across Cell Lines 123
4.2.3 Myofibroblast Recruitment Drug Screen 133
4.3 Discussion 150

Chapter 5: Categorization of Gene Expression Using Gaussian Mixture Modeling 153

5.1 Introduction 154

5.2 Analytical Approach 156

5.2.1 Gaussian Mixture Modeling 157

5.2.2 Tail Trimming 157

5.3 Analytical Process 159

5.3.1 GMM Validation and Examples 161

5.3.2 GMM and Tail Trimming in Real Data 168

5.3.3 Example Outputs 170

5.4 Discussion 176

Chapter 6: Association Analysis of Categorical Gene Expression Data 179

6.1 Introduction 180

6.2 Analytical Approach 182

6.2.1 2x2 Analysis of Gene Expression and Mutation 182

6.2.2 Interactive Association Mining Interface 184

6.3 Analytical Outputs 185

6.3.1 Differentiation in Colorectal Cell Lines and Primary Tumors 187

6.3.2 Goblet Cell Formation 192

6.3.3 Epithelial-Mesenchymal Transition 197

6.3.4 Myofibroblast Contamination of Primary Tumors 202

6.4 Discussion 207

Chapter 7: General Discussion 210

7.1 Myofibroblast-derived HGF Modulates Differentiation in a Subset of Colorectal Cancers 211

7.2 PDGF Ligands from Colorectal Cancers Activate and Recruit Local Myofibroblasts 213

7.3 Gaussian Mixture Modeling Extracts Meaningful Heterogeneity in Colorectal Cancer Expression Data 216

7.4 Summary 219

References 220

Appendix 237

List of Figures

Figure 1.1 Structure of the colonic crypt and subepithelial lining 4

Figure 1.2 Myofibroblast function in wound healing of the skin.....11

Figure 1.3 The adenoma-carcinoma sequence 14

Figure 3.1 Outline of the layer coculture assay.....50

Figure 3.2 Brightfield images of six cell lines in the layer coculture assay 51-4

Figure 3.3 Zoomed brightfield images of LS180 in the layer coculture assay 56

Figure 3.4 Fluorescence images of five cell lines in the layer coculture assay 58-62

Figure 3.5 Fluorescence images of five cell lines in layer monoculture with conditioned media.....65-69

Figure 3.6 Quantification of conditioned media monoculture experiment.....70

Figure 3.7 Fluorescence images of drug responses of five cell lines in the layer coculture assay 74-83

Figure 3.8 Fluorescence images of drug responses of five cell lines in the layer coculture assay, zoomed 85-89

Figure 3.9 Confirmation of siRNA knockdowns in myofibroblasts 91

Figure 3.10 Fluorescence images of siRNA responses of five cell lines in the layer coculture assay 93-102

Figure 3.11 Fluorescence images of siRNA responses of five cell lines in the layer coculture assay, zoomed 103-107

Figure 3.12 *HGF* expression across myofibroblast and fibroblast cell lines 108

Figure 4.1 Outline of the myofibroblast recruitment assay 117

Figure 4.2 Myofibroblast recruitment assay – optimization 120-122

Figure 4.3 Myofibroblast recruitment assay – cancer cell line screen 125-128

Figure 4.4 Gene expression comparison between myofibroblast recruiting and non-recruiting cell lines 132

Figure 4.5 Myofibroblast recruitment assay – drug screen 135-146

Figure 4.6 *PDGFA* and *PDGFC* microarray expression in cell lines 149

Figure 5.1 Example of desired GMM output on simulated data 156

Figure 5.2 Example of tail trimming on simulated data 158

Figure 5.3 GMM analysis of unimodal simulated data 160

Figure 5.4 GMM detection of bimodality in simulated data 163

Figure 5.5 GMM detection of bimodality in simulated data by t-value .. 165

Figure 5.6 GMM detection of bimodality in simulated data by smaller mean subset size 166

Figure 5.7 Concordance of simulated and predicted subgroup membership 168

Figure 5.8 GMM application to real microarray data from cell lines 171

Figure 5.9 GMM tail trimming in real microarray data from cell lines .. 172

Figure 5.10 GMM application to real RNA-Seq data from TCGA 174

Figure 5.11 GMM tail trimming in real RNA-Seq data from TCGA 175

Figure 6.1 Example 2x2 table, with r and χ^2 calculations 183

Figure 6.2 Example screenshots of GMM query interface 186

Figure 6.3 Differentiation in colorectal cancer cell lines, genetic associations....188
Figure 6.4 Differentiation in primary colorectal tumors, genetic associations....189
Figure 6.5 *TGFB2* and *PDGFC* expression in cell lines and primary tumors....191
Figure 6.6 Goblet cells in colorectal cancer cell lines, genetic associations....193
Figure 6.7 Goblet cells in primary colorectal tumors, genetic associations....194
Figure 6.8 *ATOH1* and *SPDEF* expression in cell lines and primary tumors....196
Figure 6.9 EMT in colorectal cancer cell lines, genetic associations....198
Figure 6.10 EMT in primary colorectal tumors, genetic associations....199
Figure 6.11 EMT markers and transcription factors in cell lines and primary tumors, genetic associations....201
Figure 6.12 *NKX2-3* expression in primary tumors....204
Figure 6.13 *AOC3* expression in primary tumors....205
Figure 6.14 Genes co-varying with *AOC3* in primary tumors....206

List of Tables

Table 2.1 Cell lines used in this thesis 30

Table 2.2 siRNAs used in this thesis 34

Table 2.3 Drugs used in this thesis 37

Table 2.4 TaqMan Gene Expression Assays used in this thesis 40

Table 2.5 Reagents used for sample staining in this thesis 43

Table 3.1 Cell lines used in the layer coculture assay 50

Table 3.2 Top fifty genes upregulated in myofibroblasts over skin fibroblasts 72

Table 3.3 Drugs tested in layer coculture assay 73

Table 4.1 Cell lines used in the myofibroblast recruitment assay 124

Table 4.2 Summary of the myofibroblast recruitment screen across cell lines 129

Table 4.3 Drugs tested in myofibroblast recruitment assay 134

Table 5.1 Cell line data used in GMM analysis 160

List of Abbreviations

ACF	Aberrant Crypt Foci
BIC	Bayesian Information Criterion
CAF	Cancer-Associated Fibroblast
CM	Conditioned Medium
CSC	Cancer Stem Cell
DAPI	4',6-diamidino-2-phenylindole
DMEM	Dulbecco's Modified Eagle's Medium
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
ECM	Extracellular Matrix
EMT	Epithelial-Mesenchymal Transition
FAP	Familial Adenomatous Polyposis
FBS	Fetal Bovine Serum
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GMM	Gaussian Mixture Modeling
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
MF	Myofibroblast
mL	Milliliter
mm	Millimeter
μg	Microgram
μL	Microliter
μm	Micrometer
μM	Micromolar
nM	Nanomolar

PCA	Principal Component Analysis
qRT-PCR	Quantitative Real-Time Polymerase Chain Reaction
RNA	Ribonucleic Acid
RNA-Seq	Ribonucleic Acid Sequencing
RT	Reverse Transcription
SF	Serum Free
siRNA	Small Interfering Ribonucleic Acid
TA Cells	Transit-Amplifying Cells
TCGA	The Cancer Genome Atlas

Chapter One

Introduction

1.1 The Human Colonic Epithelium: Architecture and Physiology

The surface of the human colon is composed of a single layer of epithelial cells, renewed every 4-5 days. This high rate of cell turnover is determined by the activity of the colonic stem cells residing at the bases of crypts, invaginations that punctuate the epithelial layer and extend into the lamina propria (**Figure 1.1**) (Humphries & Wright 2008). Homeostasis of the colonic epithelium depends strongly on the appropriate regulation of stem cell proliferation, as well as subsequent fate and differentiation (Medema & Vermeulen 2011). Disruptions of this homeostasis due to accumulating mutations, with sequentially more severe phenotypes, constitute the paradigm adenoma-carcinoma sequence of colorectal oncogenesis (Walther et al. 2009).

The crypt is considered to be the functional unit of the colonic epithelium. At the base of the crypt lies the stem cell niche, a privileged environment that preserves the stem cells in their undifferentiated state (Humphries & Wright 2008). As the stem cells proliferate, at a frequency of up to one division per day, their progeny move up the sides of the crypt and partially differentiate into transit-amplifying (TA) progenitor cells. After 4-5 rapid rounds of division, TA cells terminally differentiate into the three primary cellular components of the colonic epithelium: the enterocytes, which perform absorptive functions; the goblet cells, which secrete mucin into the colonic lumen; and the enteroendocrine cells, which secrete various hormones that modulate functions of the epithelium and the intestinal smooth muscle (Snippert et al. 2010; Marshman et al. 2002). The epithelial surface of the crypt is surrounded by the heterogenous mesenchyme of the lamina propria, consisting of pericryptal

myofibroblasts and pericytes, as well as immune cells and bone marrow-derived circulating cells; the crypts are bounded immediately below by the muscularis mucosa (Powell et al. 2005). The pericryptal myofibroblasts, traditionally defined as such by their fibroblastic appearance and expression of α-smooth muscle actin (αSMA), and more recently shown to express distinguishing markers AOC3 and NKX2-3, form a structure variously described as a sheath or syncytium immediately below the basement membrane of the crypt epithelial cells, extending along the full height of the crypt (**Figure 1.1**) (Hsia et al. 2016; Pascal et al. 1968; Kaye et al. 1968; Richman et al. 1987; Sappino et al. 1989). A range of studies has implicated these myofibroblasts in the maintenance of the crypt stem cell niche and the regulation of TA cell differentiation, as well as additional roles in wound healing, fibrosis, immunity, and oncogenesis (Humphries & Wright 2008; Klingberg et al. 2013; Owens & Simmons 2013).

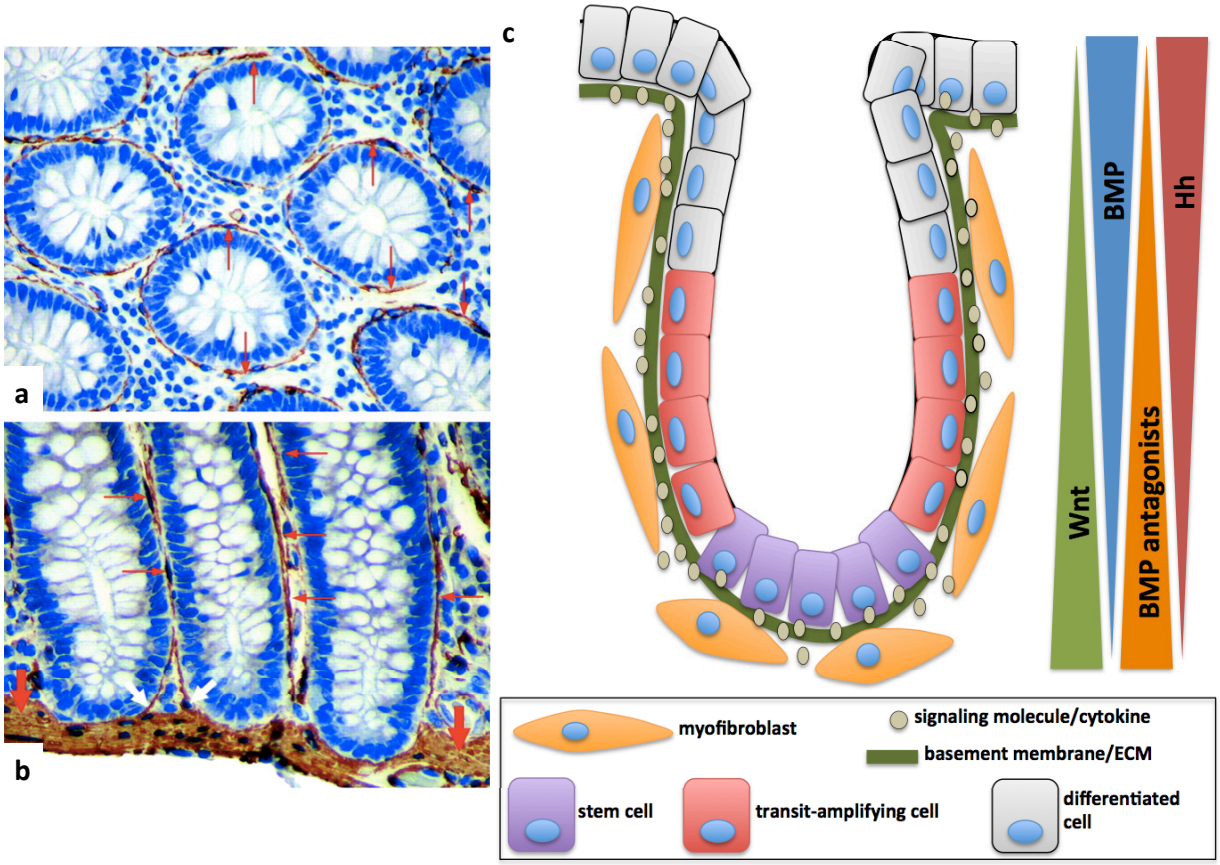


Figure 1.1: Structure of the colonic crypt and subepithelial lining. (**a**) Cross section and (**b**) longitudinal section of a colonic crypt from healthy colon, with nuclei stained with DAPI and pericryptal myofibroblasts (thin arrows) and muscularis mucosa (thick arrows) stained for α-smooth muscle actin. (**c**) Schematic of the colonic crypt, depicting the differentiation of the epithelial cells and the location of the myofibroblasts below the basement membrane. Gradients of the activity of several signaling pathways involved in colonic cell differentiation are shown. (**a** and **b**) adapted from (Powell et al. 2005), (**c**) adapted from (Powell et al. 2011).

1.2 The Normal State: Myofibroblast Roles in Healthy Tissue

1.2.1 The Definition of Myofibroblast

Before a thorough discussion of the varied roles of colonic myofibroblasts can be pursued, it is important to address the uncertainty around the definition of this cell type. Myofibroblasts were originally defined as fibroblasts possessing certain characteristics of smooth muscle cells, responsible for the contraction of wounds (Majno 1979). However, it was found early on that myofibroblasts vary substantially in their patterns of protein expression, so much so that it was proposed that a myofibroblast should be defined based on ultrastructural rather than immunohistochemical characteristics (Schürch et al. 1998). Later studies suggested αSMA as a general myofibroblast marker, but significant heterogeneity remained within this definition. Myofibroblasts as defined by αSMA expression are found not only in reactive settings such as wound healing, but also in the intestinal mucosa, lymph nodes, liver, spleen, and a number of other tissues in normal steady state. Recent work has shown that resident tissue myofibroblasts are somewhat heterogenous from tissue to tissue and clearly distinct from activated fibroblasts in both their expression patterns of certain marker proteins and in their responses to the key cytokine Transforming Growth Factor-β1 (TGF-β1), a primary fibroblast activator. The following sections distinguish between myofibroblasts in general, defined broadly as αSMA^{+}, and pericryptal myofibroblasts, most recently defined as αSMA^{+}, $AOC3^{+}$, $NKX2.3^{+}$, $LRRC17^{+}$, $EDNRB^{+}$, VIM^{+}, $EPCAM^{-}$, $SHOX2^{-}$, $TBX5^{-}$ (Hsia et al. 2016).

1.2.2 Myofibroblasts in Crypt Maintenance

Pericryptal myofibroblasts play a crucial role in the maintenance of the colonic stem cell niche in the crypt, predominantly via classical morphogenetic pathways (**Figure 1.1c**) (Medema & Vermeulen 2011). The WNT pathway has long been implicated in intestinal morphogenesis, and a majority of sporadic colorectal cancers exhibit inactivating mutations in adenomatous polyposis coli (*APC*) or activating mutations in β-catenin (*CTNNB1*) (Kinzler & Vogelstein 1996; Reya & Clevers 2005). More recently, it was found that nuclear β-catenin, signifying an activation of the WNT signaling pathway, is found in cells at the base of the crypt and not further above (Batlle et al. 2002). Mice with deletions of the WNT target transcription factor TCF4, or with overexpression of WNT inhibitor DKK1, show reduced or absent stem cell niches in crypts of colon and small intestine (Clevers et al. 1998; Pinto et al. 2003; Kuhnert et al. 2004). Although Paneth cells are reported to secrete WNT ligands in small intestine, colonic crypts lack these cells, and pericryptal myofibroblasts are believed to be important sources of WNT stimulation to maintain colonic stem cells (Farin et al. 2012; Lei et al. 2014; Valenta et al. 2016).

Bone morphogenetic protein (BMP) growth factors belong to the TGF-β superfamily of morphogenetic proteins, and signal via shared SMAD intracellular transducers (Hardwick et al. 2008). Originally identified in the study of juvenile polyposis, this signaling axis has since been found to play a critical role in the spatial organization of the crypt and the control of TA cell proliferation and differentiation. Conditional inactivation of BMPR1A, one half of the dimeric BMP receptor complex, as well as the overexpression of BMP inhibitor noggin, leads to a substantial expansion of the

proliferative compartment of mouse crypts, presumably by abrogating the effect of pericryptal myofibroblast-derived BMP4 ligand (He et al. 2004; Haramis et al. 2004). Gene expression comparisons between the tops and bottoms of human colonic crypts also show an increased expression of BMP antagonists *GREM1* and *GREM2* by pericryptal myofibroblasts at the crypt base; this is thought to restrict the activity of BMP signaling to the upper region of the crypt, where it induces and/or modulates differentiation (Kosinski et al. 2007; Auclair et al. 2007). It therefore seems that pericryptal myofibroblasts act as sources of both BMP ligands and inhibitors, depending on their location along the longitudinal axis of the crypt, indicating an underlying heterogeneity in the pericryptal myofibroblast population or a context dependence of its behavior.

A further layer of complexity is provided by the Hedgehog pathway, which has also been implicated in epithelial-mesenchymal signaling in the crypt (Powell et al. 2011). Indian hedgehog (IHH) is a ligand of the Hedgehog pathway expressed by differentiated colonic epithelial cells, and is believed to antagonize WNT signaling, thus restricting WNT activation to the bottom of the colonic crypt (van den Brink et al. 2004). Loss of IHH signaling in mouse small intestine also results in reduced stimulation of SMAD proteins downstream of BMP and Activin receptors, as well as a concomitant wound healing-like increase in epithelial proliferation, coupled with mesenchymal and immune infiltration (van Dop et al. 2010). It is suggested that IHH stimulates the secretion of certain BMPs or Activins from pericryptal myofibroblasts, which in turn modulate the balance between proliferation and differentiation of the epithelial cells, while loss of IHH disrupts the integrity of the pericryptal myofibroblast layer and the muscularis mucosa below (Kosinski et al. 2010).

1.2.3 Myofibroblasts in Wound Healing and Fibrosis

Wound healing is an adaptive process, dependent on the tissue as well as the extent and type of the wound. The mildest form of epithelial injury is one that leaves the basement membrane intact. Such damage can occur by mechanical scraping or by mild epithelial exfoliation, a possible response of intestinal or stomach mucosa to a toxin or pathogen (Wallace & Granger 1996; Moore et al. 1989). An epithelium denuded in this way is rapidly repaired by restitution, a process in which intact epithelial cells on the edges of the wound become motile and move along the basement membrane to join in the middle and reform tight junctions; this requires no inflammation and may be followed by cell proliferation if necessary to replenish the epithelial cell density (Silen & Ito 1985). Restitution of the colonic epithelium is reportedly promoted by pericryptal myofibroblasts residing directly below the basement membrane through the secretion of a range of different factors, including TGF-β3, hepatocyte growth factor (HGF), stromal derived factor-1 (SDF-1/CXCL12) and others, although no clear model for this process exists so far (Powell et al. 1999; Iizuka & Konno 2011; McKaig et al. 1998; Moyer et al. 2007).

More detailed roles for myofibroblasts have been described in the wound healing of the skin. In the case of a deep wound that pierces the basement membrane and damages subepithelial tissues, a more involved process involving inflammation, contraction, and scarring occurs to repair the tissue architecture (**Figure 1.2**) (Tomasek et al. 2002). During the initial inflammatory phase of healing, a fibrin clot forms a provisional matrix filling the space of the wound and providing an anchor for cells to enter it. Next, new stroma, termed granulation tissue, begins invading the

wound space from below. The prevalent model suggests that resident tissue fibroblasts are initially stimulated by ligands including TGF-β1 and platelet-derived growth factor (PDGF) to invade the provisional matrix of the clot, where they begin producing the ED-A isoform of fibronectin (Martin 1997; Ffrench-Constant et al. 1989; Serini et al. 1998). The fibroblasts begin reorganizing local collagen fibers and placing stress on them, which in turn promotes the transformation of fibroblasts into proto-myofibroblasts that exhibit actin stress fibers (Tomasek et al. 1992). Proto-myofibroblasts are then induced to express αSMA in response to TGF-β1, either in an autocrine fashion or due to secretion from other nearby cells (Desmoulière et al. 1993; Kim et al. 1990). At this point, they exert increased contractile forces on the wound, along with secreting a range of collagens and other extracellular matrix (ECM) components and matrix metalloproteinases (MMPs), which remodel the provisionial matrix of the clot and close the wound (Hinz et al. 2001; Klingberg et al. 2013; Prajapati et al. 2000). Widespread apoptosis of myofibroblasts follows, leaving a scar. Notably, the contribution of resident tissue myofibroblasts akin to pericryptal myofibroblasts to this process, in contrast to activated fibroblasts induced to express αSMA as described above, remains controversial.

Fibrosis is a pathological outcome of the scarring process described above, in which appropriate apoptosis of myofibroblasts is not completed upon wound closure, and ECM remodeling continues to function in an uncontrolled fashion and compromises the structure of the tissue. There are various causes of fibrosis, including aberrations in innate and adaptive immunity as well as biophysical cues and possibly even epithelial-mesenchymal transition (EMT) of local tissue cells into myofibroblasts (Wynn & Ramalingam 2012). The signaling bases of continued fibroblast activation

into myofibroblasts in fibrosis are often the same as in normal wound healing, and the myriad of pathways to fibrosis betrays the variety of different cell types that are able to secrete these signals. Changes in myofibroblasts themselves, namely the prevention of their normal apoptosis at the conclusion of wound healing, have also been implicated in fibrosis. Specifically, the transcription factor RELB of the NF-κB family has been reported to control the downregulation of chemokine secretion from myofibroblasts, and its loss is associated with chronic inflammation and fibrosis (Xia et al. 1997; McMillan et al. 2011).

Although not completely analogous to what happens in the skin, colonic wound healing is also known to involve an engagement of myofibroblasts (Rieder et al. 2007). While the colonic stroma is not known to contain cells directly corresponding to skin fibroblasts, pericryptal myofibroblasts are thought to exhibit an activated state upon epithelial damage. In this context, they exhibit increased migration into the wound site, and are thought to contribute to ECM synthesis (Leeb et al. 2002; Stallmach et al. 1992). More substantial damage due to injury or disease of the colonic epithelium, such as in intestinal bowel disease, radiation enteritis, cystic fibrosis, and other conditions, is known to result in scarring and fibrosis. Further study is required to fully characterize the similarities and differences between wound healing in the skin and the intestine, and the distinct roles that myofibroblasts play in these processes.

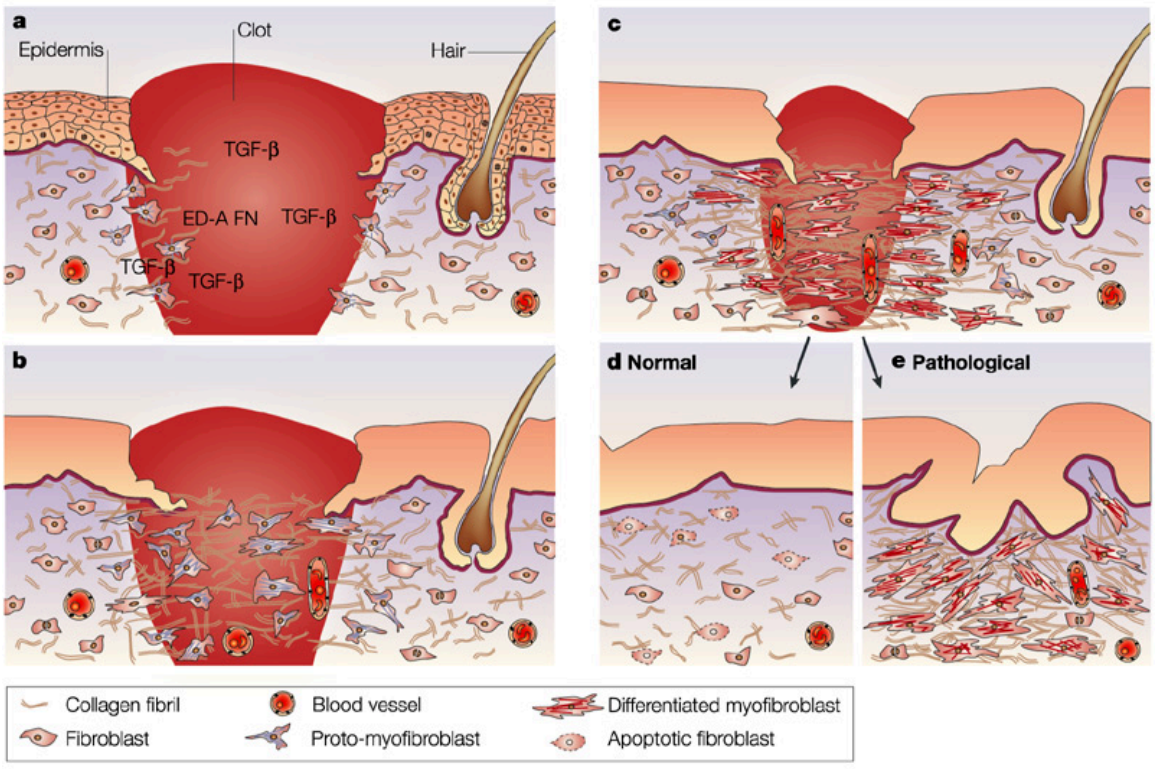


Figure 1.2: Myofibroblast function in wound healing of the skin. (**a**) When a deep wound forms in the epithelium, it is filled with a fibrin clot, and TGF-β ligands recruit fibroblasts into it, which secrete ED-A fibronectin. (**b**) Invading fibroblasts secrete and rearrange collagen to modify the clot ECM. (**c**) Fibroblasts transform into contractile myofibroblasts exhibiting actin stress fibers, which contract the wound. (**d**) Finally, the wound closes, and myofibroblasts undergo apoptosis to form a scar. (**e**) In pathological conditions like fibrosis, myofibroblasts are not appropriately culled and continue to deform local ECM. Figure from (Tomasek et al. 2002).

1.2.4 Myofibroblasts as Immune Mediators

As hinted in the previous section, one critical interaction not yet discussed is the one between myofibroblasts and the immune system. In the skin, myofibroblasts and immune cells act in close proximity during wound healing, and invade the provisional matrix of the fibrin clot as soon as it forms. In the intestine, myofibroblast-immune interactions are most evident in Inflammatory Bowel Disease (IBD) and Crohn's Disease (CD), and may involve both pericryptal myofibroblasts and other components of the mesenchyme. The simplest mouse models of chronic intestinal inflammation involve the activation of mesenchymal cells by tumor necrosis factor α (TNFα) secreted by the epithelium (Armaka et al. 2008; Roulis et al. 2011; Theiss et al. 2005). Pericryptal myofibroblasts are also known to respond to a range of other cytokines, including IL-1α, IL-1β, IL-4, IL-17A, IL-17F, IL-21, IL-22, and interferon γ (IFNγ), among others, with a wide spectrum of secretory responses (Di Mari et al. 2003; Okuno et al. 2002; Hata et al. 2002; Yagi et al. 2007; Andoh et al. 2005; Andoh et al. 2002; Bamba et al. 2003; De L. Karlson et al. 2006). These secretions, in turn, can modulate the survival, accumulation, and differentiation of many different cell types from the lymphoid and myeloid lineages (Owens & Simmons 2013). Intriguingly, recent reports have also implicated intestinal mesenchymal cells in direct sensing of pathogens. Pericryptal myofibroblasts have been reported to express a range of NOD-like and toll-like receptors, which recognize structural signatures of bacterial pathogens and stimulate innate immune responses, and to secrete the immunostimulatory chemokines CCL2 and IL-8 in response to bacteria (Kim et al. 2011; Normand et al. 2011; Otte et al. 2003; Pinchuk et al. 2007). Isolated bacterial lipopolysaccharides have also been reported to induce the production of a wide range

of secreted chemokines and adhesion proteins in pericryptal myofibroblasts (Pang et al. 1994; Burke et al. 2010; Walton et al. 2009). Moreover, it has been shown that pericryptal myofibroblasts possess the ability to process and present antigens to $CD4^+$ T cells in a major histocompatibility complex (MHC) class II-dependent manner, implicating them as possible non-immune antigen-presenting cells (Saada et al. 2006). The spectrum of myofibroblast-immune interactions is evidently quite varied and requires much more thorough study.

1.3 The Deregulated State: Myofibroblasts in Colorectal Carcinoma

Across tissues, myofibroblasts exhibit a wide spectrum of activities, both at steady state and in inflammation and wound repair. A further layer of complexity is introduced in carcinomas, when mutated epithelial cells begin aberrantly expressing and secreting receptors and signaling molecules. The microenvironment has an important influence on the evolving cancer, and the interrelationship between the tumor and its surroundings changes as this evolution progresses. As primary components of the colon crypt stem cell niche, pericryptal myofibroblasts play a key role in responding to an incipient tumor and determining the signaling milieu that it is subjected to, from early oncogenesis to progression and metastasis.

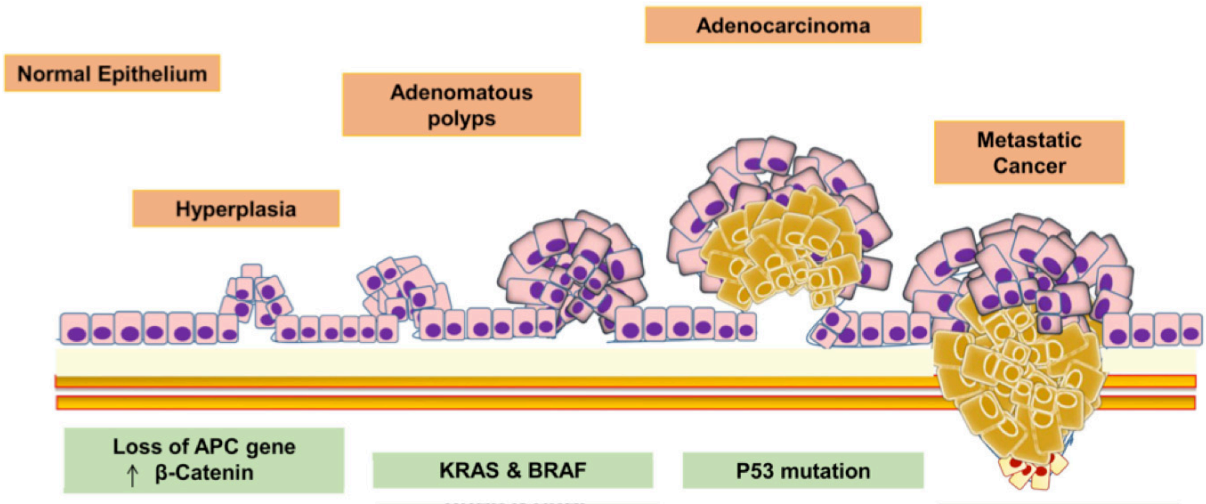


Figure 1.3: The adenoma-carcinoma sequence, including several stages of adenoma development and eventual malignant transformation and metastatic spread. As shown below, changes in the WNT pathway initiate the process, with subsequent mutations common in *KRAS*, *BRAF*, and *TP53*. Figure from http://cancer-research-frontiers.org/2016-2-1/

1.3.1 Malignant Transformation in Colorectal Carcinoma

Colorectal tumors begin as benign local growths and subsequently undergo malignant transformation, following the classical adenoma-carcinoma sequence (**Figure 1.3**) (Hill et al. 1978). Extensive studies of hereditary and sporadic colorectal cancers have identified a set of mutational signatures that characterize the various pathways to malignancy along this sequence (Fearon & Vogelstein 1990). The most prevalent initial mutation inducing transition from normal epithelium to adenomatous tissue occurs in the WNT pathway, specifically in the *APC* gene. Inherited loss-of-function mutations in this gene give rise to familial adenomatous polyposis (FAP), a syndrome exhibiting the appearance of hundreds to thousands of colonic adenomatous polyps by adolescence, with nearly 100% of affected patients developing colorectal cancer (Bodmer et al. 1987; Markowitz & Bertagnolli 2009). Various studies estimate that 60-80% of sporadic colorectal cancers exhibit *APC* mutations (Powell et al. 1992; Kinzler & Vogelstein 1996). Additionally, and not in a mutually exclusive fashion, around 15% of sporadic colorectal tumors are found to exhibit DNA mismatch repair deficiency, a form of genetic instability involving defects in the mismatch repair pathway (Poulogiannis et al. 2010). These cancers exhibit an increased mutation rate due to unresolved DNA replication errors throughout the genome, and especially in regions of repetitive DNA such as microsatellites. Inherited mismatch repair deficiencies characterize hereditary non-polyposis colorectal cancer (HNPCC or Lynch syndrome), another familial disease accounting for 3-6% of all colorectal cancers (Konishi et al. 1996). Notably, there is some controversy over the prevalence of *APC* mutations in mismatch repair deficient cancers, with estimates ranging between 10% and over 50%, and only 25% are reported to carry gain-of-function

mutations in intracellular WNT signal mediator β-catenin (*CTNNB1*), which was proposed to compensate for the lack of *APC* mutation in those tumors (Huang et al. 1996; Olschwang et al. 1997; Mirabelli-Primdahl et al. 1999). Thus while an overactive WNT pathway seems to be the major first hit along the mutational sequence from normal to malignant colonic epithelium, alternative pathways may also exist (Jass et al. 2003). Beyond the first mutation, further changes must accumulate for the growth to transition from precancerous lesion to carcinoma. The rate of this process is affected by various forms of genomic instability, namely mismatch repair deficiency as discussed above, chromosomal instability as a consequence of WNT pathway dysfunction, aberrant methylation, and possibly other rare DNA repair changes, although none of these characteristics are required for malignancy to develop (Fodde et al. 2001). Malignant colorectal cancers often exhibit inactivation of the p53 pathway and the TGF-β pathway, as well as activating mutations in *KRAS*, *BRAF*, and *PIK3CA* (Markowitz & Bertagnolli 2009).

The accumulation of all of these mutations, and the selection process through which mutated cells outgrow their wild-type neighbors to form a sizable tumor, requires many rounds of cell division. Terminally differentiated, post-mitotic cells make up the majority of the colonic epithelium, as well as nearly all tissues, and it has therefore been postulated that the cells-of-origin of cancer must be either tissue stem cells or at most partially differentiated progenitor cells (Visvader 2011; Bach et al. 2000). In recent years, the "bottom-up" theory of colorectal carcinogenesis, involving the growth of early adenomas from a single mutated stem cell at the bottom of the crypt, has been supported with several approaches (Humphries & Wright 2008). FAP polyps begin as monoclonal, monocryptal aberrant crypt foci (ACFs), crypts completely

filled with adenomatous cells (Preston et al. 2003). Induced deletion of *APC* in a single stem cell of a mouse intestinal crypt leads to rapid bottom-up clonal expansion of the stem cell and its differentiated progeny, with monoclonal transformed crypts visible in days and large adenomas within a few weeks; subsequent top-down invasion into nearby normal crypts is also observed (Barker et al. 2009). Identical deletions in progenitor cells away from the bottom of the crypt produce small adenomas developing at a much slower rate, suggesting that while progenitor cells may be the cells-of-origin of some colorectal tumors, the stem cell-derived adenomas may be more aggressive, at least initially. Intriguingly, although initial FAP ACFs are monoclonal, macroscopic adenomas are generally polyclonal in origin, suggesting that they arise as fusions of several ACFs, likely with distinct mutational profiles (Novelli et al. 1996).

1.3.2 Myofibroblast Roles in Precancerous Adenomas and Primary Tumors

To varying extents, the stroma of both precancerous adenomas and malignant carcinomas of the colon accumulates myofibroblasts (Adegboyega et al. 2002). In fact, increased αSMA in the stroma of a primary colorectal tumor is a strong predictor of disease recurrence after treatment (Tsujino et al. 2007). As with normal colonic epithelium, colorectal cancers are significantly responsive to their microenvironment, and many mutations critical to colorectal tumorigenesis involve precisely the signaling pathways implicated in the maintenance of normal crypt homeostasis, resulting in misregulation of cell proliferation and differentiation in the crypt. For instance, pericryptal myofibroblasts are the primary source of WNT ligands in the

colon, and addition of WNT activator R-spondin 1 is required for the establishment of *in vitro* intestinal organoids derived from primary wild-type mouse tissue in the absence of mesenchymal cells (Sato et al. 2009). Similarly, pericryptal myofibroblasts provide BMP ligands to regulate the differentiation of colonic epithelial cells, and mice combining mutations in *Apc* and the mouse ortholog of *SMAD4*, a mediator of signals downstream of TGF-β and BMP receptors, develop aggressive poorly-differentiated tumors with considerable stromal invasion (Takaku et al. 1998). Similar results were obtained through the ectopic expression of BMP inhibitor noggin in mouse enterocytes (Haramis et al. 2004).

Other pericryptal myofibroblast-derived signals with less clear functions in normal colon have also been implicated in colorectal cancer. As hinted above, while colorectal tumors (and many, if not all, other solid tumors) exhibit progressive deterioration of normal tissue architecture, they often do not completely lose characteristics of differentiation. In fact, it is now broadly accepted that tumors consist of a population of self-renewing cancer stem cells (CSCs) along with their differentiated progeny; these CSCs may be distinct from the stem/progenitor cells-of-origin of the tumor, but exhibit similar characteristics to normal stem cells (Vermeulen et al. 2008). Although the general assumption is that CSCs make up a small fraction of the overall tumor, reports suggest that this is not an absolute limitation and that the CSC proportion can in fact fluctuate significantly as the tumor evolves or responds to treatment (Johnston et al. 2007; Johnston et al. 2010; Werner et al. 2016). Akin to normal stem cells, CSCs reside in distinct niches that supply them with various factors necessary to maintain their stemness (Plaks et al. 2015). Colorectal tumors exhibit heterogeneity in WNT activation despite *APC* mutations,

with increased signaling through this pathway observed in regions dense with myofibroblasts (Fodde & Brabletz 2007). It has been suggested that hepatocyte growth factor (HGF) secreted from myofibroblasts promotes this increased WNT activity and may thus constitute a key signal in the colorectal CSC niche (Vermeulen et al. 2010). Treatment with prostaglandin E_2 (PGE_2) is known to induce increased HGF secretion from pericryptal myofibroblasts, an intriguing observation in light of the fact that over 20% of inflammatory bowel disease patients develop colorectal cancer (Shao et al. 2006; Lakatos & Lakatos 2008). Additionally, inhibition of Notch signaling in colorectal tumors by either antibody against Notch ligand DLL4 or by various pharmacological treatments promotes CSC differentiation, reduces the CSC population, and reduces tumor growth, although it is unclear if the ligand is presented by the myofibroblasts or the cancer cells themselves (Hoey et al. 2009; Yuan et al. 2015).

1.3.3 Myofibroblast Roles in Tumor Invasion and Metastasis

Aside from the signaling effects described above, myofibroblasts have also been implicated in the remodeling of the tumor ECM, and myofibroblast-derived ligands have been found to influence tumor invasion and metastasis (Liang et al. 2005; Takatsuna et al. 2016; Tommelein et al. 2015). TGF-β ligands produced by the colorectal cancer cells are reportedly the primary inducers of pro-metastatic transformations in the stroma, inducing IL-11 secretion from pericryptal myofibroblasts, which in turn promotes cancer cell metastasis (Calon et al. 2012). TGF-β1 is also known to induce the production of a range of MMPs and tissue

inhibitors of matrix metalloproteinases (TIMPs), as well as several ECM components including collagen I, tenascin-C, and laminin-B1, in myofibroblasts from primary colorectal tumors (Hawinkels et al. 2014; Rupp et al. 2015). These proteins are in turn proposed to enact a wound healing-like rearrangement of tumor ECM. Ligands promoting cancer cell migration are reportedly also secreted by stromal myofibroblasts. Specifically, myofibroblasts from colorectal cancers stimulated with PDGF were found to secrete stanniocalcin-1, a ligand shown to promote xenograft growth and metastasis for several colorectal cancer cell lines (Pena et al. 2013). Additionally, TGF-β1 was also found to stimulate the secretion of fibroblast growth factor 1 (FGF1) from cultured pericryptal myofibroblasts, which in turn increased the migratory abilities of several colorectal cancer cell lines (Henriksson et al. 2011). A role for myofibroblast-derived tenascin-C and HGF in cancer cell migration has also been proposed (De Wever et al. 2004). Clearly there is a considerable range of cross-stimulation reported between myofibroblasts and the colorectal tumors that they associate with, resulting in a muddied picture of the various triggered signaling pathways.

Regardless of the mechanisms involved, primary cancer cells tend to find their way into the bloodstream relatively quickly, and the rate-limiting step of metastasis is colonization at the secondary site (Chambers et al. 2002). It has long been known that metastases of specific primary tumors tend to occur in specific tissues (Paget 1889). For instance, colorectal cancer metastases most commonly arise in liver, as well as in the lymph nodes and lungs (Riihimäki et al. 2016). Such patterns have been explained by the dependence of incipient metastatic cells on a specific set of signals in the microenvironment for optimal colonization, and a wide range of secreted factors has

been implicated in this process in various tumor types (Santamaria-Martínez & Huelsken 2013). Recent studies have shown that the formation of such a metastatic niche may begin before any cancer cells even arrive there, as evidenced by the appearance of bone marrow-derived cells at premetastatic sites prior to metastasis, possibly activated by primary tumor-derived exosomes containing various signaling molecules (Sceneay et al. 2013; Lobb et al. 2017). Metastatic cells have also been proposed to co-opt native stem cell niches in target organs as sources of colonization-promoting factors, or to induce niches ad hoc by rearranging local stroma (Massagué & Obenauf 2016). As in the colonic crypt, such niches are often patterned by local myofibroblasts, and metastatic cells are known to activate these cells and stimulate their proliferation and invasion into the tumor (Olaso et al. 1997; Yeung et al. 2013). Besides secreting niche-patterning soluble factors, myofibroblasts activated by invading cancer cells also modify the ECM, which is critical to metastasis formation. In fact, liver fibrosis is known to greatly enhance the formation of hepatic metastases from breast, melanoma, and colorectal primary tumors (Cox & Erler 2014; Cox et al. 2013; Olaso et al. 1997; Kondo et al. 2016). The engagement of local myofibroblasts is thus a key event in the development of many metastases.

1.4 Genome-wide Expression Analysis in Cell Line and Tumor Panels

Over the last decade, big data has overtaken all fields of biology, bringing along great challenges in data handling and analysis (Marx 2013). In parallel, an expanding set of analytical tools has been applied to interact with such large datasets, from basic statistics and dimensionality reduction to more complex clustering, visualization, and machine learning. These approaches have allowed for extensive data mining and hypothesis generation, and have uncovered previously unnoticed details in biological data. This section will specifically focus on the analysis of genome-wide gene expression data from large cohorts of individuals, and on the conclusions that can be drawn from such studies.

1.4.1 Analysis and Clustering of Continuous Data

Gene expression is normally quantified as a continuous measure, with data from more modern methods counting discrete transcripts such as RNA-Seq generally still normalized to result in non-integer values. Analysis of sets of gene expression data thus depends on the statistics of continuous distributions (Slonim & Yanai 2009). Such distributions have been studied for a long time, and a number of standard distribution shapes defined by specific sets of parameters are known to arise from various processes. Most commonly, biological data is assumed to be best represented by normal distributions, in part because experimental noise is assumed to be normal. Accordingly, one of the most common statistical tests cited in biological studies is Student's t-test, which quantifies the likelihood that two samples of observations of

the same variable derive from distinct normal distributions. While the t-value is a measure of separation between two samples of the same variable, it is also common to analyze pairs of variables among observations within a single sample. The most common quantification of a relationship or distance between two continuous variables, often visualized on a scatter plot, is Pearson's correlation coefficient *r*.

While limited sets of continuous data can be readily visualized and interpreted using standard plots and histograms, multivariate datasets such as genome-wide expression data are too large for meaningful representation in such a way. In these cases, it is necessary either to break the data down into smaller and more approachable subsets or, if the completeness of the dataset is required for the analysis, to apply dimensionality reduction or clustering to reduce its complexity. The most common dimensionality reduction algorithm is principal component analysis (PCA), which iteratively calculates a new set of variables describing the data as linear combinations of the original variables, each chosen to maximally represent the remaining variance in the dataset at that iteration (Ringnér 2008). However, PCA is limited in two ways; on one hand, it is usually unclear what conclusions can be drawn about the original variables from the synthetic ones, and on the other, the fact that only linear combinations are allowed means that the separation between subsets is often not very clear. More recently, it has become common to use nonlinear dimensionality reduction algorithms such as t-Distributed Stochastic Neighbor Embedding (t-SNE), which exacerbate the lack of a clear relationship between the synthetic and original variables but result in much better defined subgroup distinctions (Van Der Maaten & Hinton 2008). Clustering algorithms work in a different way to achieve similar goals as dimensionality reduction – instead of transforming the data at hand into a more

approachable form, they look for certain structure within the original dataset. Every clustering algorithm is based on an *a priori* set of rules that defines the desired characteristics of the clusters to be found in the dataset, mathematically stated in the form of a scoring function. Next, iterative algorithms are run to progressively optimize the scoring function, and criteria are specified to determine when to stop these iterations. The result is a partition of the original data into some number of subgroups with certain characteristics. Clustering can be done by defining a certain pairwise distance measure between the observations, such as one related to the correlation coefficient between them, and then using that measure as a means of partitioning apart the observations most distant from each other through hierarchical clustering. Other approaches, such as k-means clustering or self-organizing maps, begin with a predetermined number of centroids, and then optimize the choice of centroids such that the distances from all observations to their assigned cluster-defining centroids are in some way optimal (D'haeseleer 2005). Finally, clustering can be model-based; in this case, a certain probability model or shape is given that the clusters are expected to conform to, and then the number and set of characteristics of such shapes that optimally accounts for the whole dataset is determined. One such method is Gaussian Mixture Modeling (GMM), which aims to describe the data at hand as derived from an underlying distribution consisting of some number of normal components (Mclachlan & Rathnayake 2014; Scrucca et al. 2016). The advantage of model-based clustering is that finer parameters of the data can be inferred, particularly scale parameters such as standard deviations and shape parameters such as skewness (Mcnicholas 2016).

1.4.2 Discretization of Continuous Data

A cornerstone of the modern evolutionary synthesis is the idea that variation of traits across a population arises as a result of Mendelian inheritance; in other words, a set of discrete parameters, namely a combination of genetic alleles, combines to give rise to an ostensibly continuous range of phenotypes (Fisher 1919; Mayr & Provine 1980). Evolution of a population occurs as the prevalence of alleles with greater advantages increases, while individuals carrying alleles with lower advantage have reduced fitness and ability to pass on their genetic information to the next generation. Oncogenesis is exactly such a process, with successive mutations and epigenetic changes accumulated over time, with the preferential outgrowth of the clone with the most advantageous combination of alleles (Nowell 1976; Merlo et al. 2006). However, evolution is not always a gradual process, and evidence from the fossil record suggests that species often exhibit rapid bursts of evolution followed by long periods of stasis; this is the theory of punctuated equilibrium (Gould & Eldredge 1993; Volkenstein 1987). Again, such evolutionary dynamics have been suggested in cancer development (Cross et al. 2016). There is thus a limited and discrete set of step-like changes that determine the identity of a cancer. Mapping continuous data into a finite set of clusters effectively turns it into discrete data, and in the case of cancer cell line or tumor-derived expression data, it is worthwhile to explore the possibility that these discrete subsets may relate directly to the key genetic changes responsible for that cancer's evolutionary trajectory. The genes most likely to be involved in cancer evolution would be expected to exhibit not just wide variation of expression across the panel of individual cancers but a dichotomous pattern,

separating the samples by evolutionary pathways or progression along these pathways.

Discrete distributions, like continuous ones, can be analyzed using certain classical statistical approaches. Analogously to the continuous case, it is possible to assess the likelihood that two samples of the same discrete variable derive from the same underlying distributions using either Fisher's Exact Test or a χ^2 test, and distance between two variables can be quantified as association coefficient *r* or Cramér's V, a generalized form of *r* for discrete variables with more than two levels. The most common dimensionality reduction technique for categorical data is correspondence analysis, which in a completely analogous fashion to PCA iteratively defines new variables as linear combinations of the original that maximally describe the remaining variability in the data. Distance-based clustering can be performed in the exact same way as before with an appropriate choice of distance measure, and even model-based clustering is possible, for example using Latent Class Models, which is a method somewhat similar to GMM except that the component distributions are multinomial rather than normal (Chen et al. 2012). While discretization of continuous gene expression data using methods such as GMM removes the finer details of the original dataset, it also counteracts the confounding factors of normally distributed technical noise, and hones in on the most striking differences between subsets of the samples. In the meantime, the parameters of the predicted distributions describing the original data can also be used to interpret with greater precision the heterogeneity within it. By using discretized gene expression data from cancers to cluster together the expression patterns of strongly variable and dichotomous genes, it may be possible to more

clearly infer the driver mutations that characterize the evolution of such cancers, and therefore to identify the most promising targets for therapy.

Chapter Two

Methods and Materials

2.1 Cell Culture Methods

2.1.1 Cell Lines

The 43 human colorectal cancer cell lines used in this thesis, along with their original sources, are summarized in **Table 2.1** (Bracht et al. 2010). The myofibroblast cell line CCD-18Co, originally derived from neonatal colorectal mucosa, was acquired from American Type Culture Collection (ATCC, Ref. CRL-1459). The skin fibroblast cell line was previously established at the Cancer and Immunogenetics Laboratory (CIL) of the Weatherall Institute of Molecular Medicine (WIMM), Oxford, UK (Hsia et al. 2016). All cell lines were obtained from in-house cryogenic storage for the experiments in this thesis.

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

Table 2.1: Cell lines used in this thesis, along with their original sources.

2.1.2 Cell Culture Conditions

Cells were maintained in 25 or 75 cm^2 vented tissue culture treated flasks (Corning, Refs. 430639 and 430641U). All cells were grown in 5-10 mL of Dulbecco's Modified Eagle Medium (DMEM) (Gibco, Ref. 41966-029), supplemented with 10% Fetal Bovine Serum (FBS) (Autogen Bioclear, Cat. No. 7.01CS) and 1% penicillin/streptomycin antibiotic solution (Lonza, Cat. No. DE17-602E). All cultures were incubated at 37°C in a humidified atmosphere containing 10% CO_2.

Cells were routinely grown to 70-80% confluence prior to passaging. Once confluent, cells were washed with 5-10 mL Absolute Phosphate Buffered Saline (PBSA) and incubated at 37°C for 5 minutes in 1-3 mL Trypsin-EDTA solution for detachment (Lonza, Cat. No. BE02-007E). 4-7 mL of fully supplemented medium were subsequently added to neutralize the trypsin, and the suspension was gently mixed to maximally disrupt cell clumps. Next, the cell suspensions were centrifuged at 1000 rpm (~400g) for 5 minutes, the supernatant was decanted, and the cell pellets were resuspended in 10 mL of fully supplemented medium. New cultures were established in sterile culture flasks at 1:5-1:10 dilution.

Experimental cultures were grown in 24-well and 96-well tissue culture treated plates (Corning, Refs. 3526 and 3595). Manipulations on suspended cells were performed in 15 mL and 50 mL Falcon tubes (Scientific Laboratory Supplies, Refs. SLS8103 and SLS8110), as well as in 1.5 mL microcentrifuge tubes (TubeOne Ref. S1615).

2.1.3 Cell Line Cryostorage, Preservation and Recovery

Cell lines were stored long-term in cryogenic vials (Corning, Ref. 430489), in liquid nitrogen. During cell line deposition, cells were detached and pelleted as during passaging, but instead of resuspension in culture medium, they were instead suspended in an appropriate volume of Fetal Bovine Serum (FBS) supplemented with 10% dimethyl sulfoxide (DMSO) (AppliChem, Ref. A3672,0250). The resultant suspension was subsequently aliquoted into cryogenic vials at 600 µL per vial, and the vials were placed in a Mr. Frosty Freezing Container (Thermo Scientific, Ref. 5100-0001). The Mr. Frosty Container was left at -80°C overnight to ensure gentle freezing of the cell suspensions prior to transfer to liquid nitrogen.

During cell line recovery, vials were first removed from liquid nitrogen storage and placed on dry ice. To thaw the cells, the frozen vials were placed in a 37°C water bath for 1 min, and immediately resuspended in 10 mL fully supplemented culture medium. To remove traces of DMSO, the resultant solutions were centrifuged at 1000 rpm (~400g) for 5 minutes, the supernatant was decanted, and the cell pellets were resuspended in 10 mL fully supplemented medium and transferred into a sterile culture flask for incubation.

2.1.4 Cell Counting

Cell counting was performed using the Cellometer Auto T4 device (Nexcelcom Bioscience, USA). 20 µL of cell suspension were transferred into a disposable counting chamber (Cellometer, Ref. SD100), which was subsequently inserted into

the cell counter. The software uses brightfield imaging to observe eight separate areas in the counting chamber, and detects cell shapes and sizes to calculate cell concentrations and estimate viability.

2.1.5 Conditioned Medium Harvesting and Usage

Conditioned media were harvested from the CCD-18Co myofibroblasts and the skin fibroblasts. The cells were initially grown as normal in 75 cm^2 flasks in fully supplemented DMEM. Once the cells reached 90% confluence, the culture medium was aspirated, and the cultures were washed thrice with 10 mL PBSA by repeat aspiration and addition of liquid, taking care not to disturb cell attachment to the flask. After washing, serum-free DMEM containing 1% penicillin/streptomycin was added to the flasks, and the cells were incubated as normal for 72 hours. The culture medium was then passed through a 0.2 μm diameter filter (Sartorius Ministart, Ref. 16534-K) to remove cell debris and stored at -20°C for at most two weeks. The conditioned medium was thawed in a 37°C water bath immediately before use and applied without dilution in subsequent experiments.

2.1.6 Transient siRNA Transfection

Small interfering RNAs (siRNAs) were transiently transfected into CCD-18Co myofibroblasts using the Lipofectamine RNAiMAX transfection reagent (Invitrogen, Ref. 13778150). One day before transfection, 4,000 or 20,000 cells were seeded in

96- or 24-well plates, respectively, in 100 µL or 500 µL DMEM supplemented with 10% FBS but lacking antibiotic. The next day, an appropriate volume of siRNA was diluted in serum-free Opti-MEM medium (Gibco, Ref. 11058-021) for a concentration of 120 nM. In an equal volume of Opti-MEM, Lipofectamine RNAiMAX was diluted at a 1:50 ratio. The two solutions were mixed and incubated at room temperature for 20 min, at which point 20 µL or 100 µL of the overall solution was added directly to the cells in the 96- or 24-well plates, respectively, for a final siRNA concentration of 10 nM per well. Culture medium was replaced with 200 µL or 1 mL fully supplemented DMEM containing antibiotic the next day. All siRNAs used in this thesis are summarized in **Table 2.2**.

Target	**Stock Concentration**	**Source**
Scrambled control	10 µM	Dharmacon
LRRC17	10 µM	Ambion by Life Technologies, Ref. s19974
EDNRB	10 µM	Ambion by Life Technologies, Ref. s4468
HGF	20 µM	Ambion by Life Technologies, Ref. s6530
NKX2-3	20 µM (three siRNAs, mixed in equal amounts)	OriGene, Ref. SR315976

Table 2.2: Summary of siRNAs used in this thesis, along with their sources and stock concentrations.

2.1.7 Matrigel Coculture

CCD-18Co myofibroblasts were cocultured with colorectal cancer cell lines in two geometric arrangements in Matrigel (Corning, Ref. 356231). All cocultures were made using Matrigel diluted to 50% in serum-free DMEM supplemented with 1% penicillin/streptomycin.

Colorectal cancer cells were prepared for coculture by detachment and pelleting in 15 mL Falcon tubes as during passaging. The cell pellets were washed three times in 10 mL PBSA by repeat resuspension, pelleting, and decanting, and finally resuspended in serum-free DMEM supplemented with 1% penicillin/streptomycin before counting. Appropriate numbers of cells were subsequently transferred into 1.5 mL microcentrifuge tubes, which were centrifuged at 2,200 rpm (~400g) for 15 min. Supernatant was removed from the cell pellets by pipetting, and the cells were then resuspended in an appropriate amount of 50% Matrigel prior to seeding.

In the layer coculture assay (**Figure 3.1**), 4,000 CCD-18Co myofibroblasts were first seeded in 96-well plates as before siRNA transfection. On the second day, the myofibroblasts were either transfected with siRNAs or treated with 20 µL unaltered Opti-MEM, and on the third day the culture medium was replaced with 200 µL fully supplemented DMEM. After another 2 days of growth, the culture medium was removed, the myofibroblasts were washed in 200 µL PBSA, and 40 µL of 50% Matrigel containing 1,000 colorectal cancer cells was added on top. After the gel was allowed to harden for one hour, it was covered with culture media containing various amounts of FBS and/or various drugs (see below). Colorectal cancer cell

monocultures in 50% Matrigel were performed in exactly the same manner, except that the myofibroblast steps were omitted and the gel was added directly to sterile 96-well plates prewashed with 200 μL PBSA to reduce surface tension and allow even spread of gel throughout the well. These cultures were allowed to grow for 18 days prior to fixation and immunostaining.

To establish cocultures for the myofibroblast recruitment assay (**Figure 4.1**), drops of 3 μL of 50% Matrigel containing 10,000 colorectal cancer cells prepared as before were first placed into dry sterile 96-well plates. After the gel hardened for one hour, it was covered with 200 μL of 1% serum-containing DMEM containing 2,000 CCD-18Co myofibroblasts, added carefully to ensure even dispersion of the myofibroblasts throughout each well. Medium was replaced in these cultures on the second, fifth, and eighth days of culture by direct pipetting, either with plain 1% serum-containing DMEM or with various drugs in addition (see below). After nine days, these cultures were fixed and immunostained.

2.1.8 Drug Screening

The various drugs utilized in this thesis are summarized in **Table 2.3**. The drugs were thawed on ice before each experiment and diluted to appropriate concentrations in desired culture medium prior to addition to cells.

Drug	Target	Concentration	Source
A8301	TGFBR1	1 μM	Tocris Bioscience, Cat. No. 2939
AMD3100	CXCR4	1 μM	Santa Cruz Biotechnology, Cat. No. sc-252367
BQ788	EDNRB	1 μM	Abcam, Ref. ab144504
Cetuximab	EGFR	10 μg/mL	Merck
Figitumumab (CP751851)	IGF1R	10 μg/mL	Pfizer
GSK690693	AKT	1 μM	Selleckchem, Cat. No. S1113
Imatinib	PDGFR	1 μM	Cayman Chemical, Item No. 13139
LDN193189	BMPR1A/ACVR1	500 nM	Sigma-Aldrich, Prod. No. SML0559
LY294002	PI3K	1 μM	Selleckchem, Cat. No. S1105
PD173074	FGFR1/2/3	100 nM	Cell Guidance Systems, Cat. No. SM27
PHA665752	c-Met	100 nM	Selleckchem, Cat. No. S1070

Table 2.3: Summary of drugs used in this thesis, along with their biochemical targets, treatment concentrations, and sources.

2.2 RNA Methods

2.2.1 RNA Extraction

Total RNA was extracted from cells as per manufacturer's instructions using the RNeasy Mini Kit (Qiagen, Ref. 74104). CCD-18Co cells were subjected to siRNA transfection in 24-well plates as described and subsequently grown for various amounts of time, with culture medium replacement every three days. Triplicate wells were grown for each condition. Prior to RNA extraction, the myofibroblasts were washed once with 500 µL PBSA. 350 µL of RLT lysis buffer was added to the first well for each condition; after 1 min of incubation at room temperature coupled with gentle pipetting, the buffer was transferred to the second, and subsequently the third of the triplicate wells for each condition, pooling together their RNA. 350 µL of 70% cell culture grade ethanol (VWR Chemicals, Ref. 437433T) diluted in Millipore filtered water was added to the RLT buffer, and the resulting 700 µL mixture was transferred to an RNeasy Mini spin column. The column was centrifuged at 10,000g for 15 sec and the flowthrough liquid was removed. The column was subsequently washed with 700 µL of RW1 buffer and then 500 µL of RPE buffer, with 15 sec centrifuging at 10,000g and flowthrough discarding each time. A second wash with 500 µL RPE buffer was followed by 2 min centrifugation at 10,000g, and the column was centrifuged again for another 1 min after the flowthrough was removed. The column was subsequently transferred to a 1.5 mL collection tube, and 30 µL RNAse-free water (Invitrogen, Ref. AM9937) was added to the filter. After 1 min incubation at room temperature, the column was centrifuged at 10,000g for 1 min, and the resulting RNA-containing solution was quantified and used in further experiments.

2.2.2 Reverse Transcription

Reverse transcription of total RNA to cDNA was performed using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Part No. 4368814). RNA samples were diluted to 10 ng/μL in nuclease-free water and mixed with 10X RT Buffer, 25X dNTP mix, 10X RT Random Primers, 20X MultiScribe Reverse Transcriptase, 20X RNAse inhibitor, and nuclease-free water to a final volume of 20 μL. The mixture was incubated in a thermal cycler as follows: 25°C for 10 min, 37°C for 120 min, and 85°C for 5 min. The resulting cDNA was quantified and used in further experiments.

2.2.3 Quantitative Real-Time Polymerase Chain Reaction (qRT-PCR)

qRT-PCR was performed in triplicate using 20X FAM dye-labeled TaqMan Gene Expression Assays (Applied Biosystems) (**Table 2.4**). cDNA samples from reverse transcription were diluted to 25 ng/μL, and 4 μL of sample was mixed with appropriate 20X TaqMan Gene Expression Assay, 2X TaqMan Fast Universal PCR Master Mix (Applied Biosystems, Ref. 4352042), and nuclease-free water for a final volume of 20 μL per well of MicroAmp Fast Optical 96-well Reaction Plate (Applied Biosystems, Ref. 4346906). The plates were sealed with MicroAmp Optical Adhesive Film (Applied Biosystems, Ref. 4311971), and qRT-PCR cycling and measurement were performed on a 7500 Fast Real-Time PCR System (Applied Biosystems, Ref. 4351107). Cycle times for the various assayed mRNAs were normalized to housekeeping protein Ubiquitin C (*UBC*).

Converted ΔCt (cΔCt) values were calculated for each assayed gene as previously reported (Hsia et al. 2016). In brief, standard comparative ΔΔCt calculations on qRT-PCR data assume that a range of samples are to be compared to a reference sample, and gene expression is then quantified as a fold change compared to this reference. In order to avoid the requirement of always choosing a reference sample, the largest ΔCt for every gene across a set of samples was determined, and all ΔCt values were then subtracted from the maximal one for each gene to give a cΔCt for each gene and each sample. This value could then be converted from a logarithmic to a linear scale by taking the $2^{c\Delta Ct}$ transformation without need for a reference sample. Error in the ΔCt reading was propagated using the standard variance method (Ku 1966).

Target	Reference	Source
UBC	Hs00824723_m1	Applied Biosystems
LRRC17	Hs00957873_m1	Applied Biosystems
EDNRB	Hs00240747_m1	Applied Biosystems
HGF	Hs00300159_m1	Applied Biosystems

Table 2.4: Summary of TaqMan Gene Expression Assays used in this thesis, along with their reference numbers.

2.2.4 Nucleic Acid Quantification

DNA and RNA quality and concentrations were quantified using the NanoDrop ND-1000 Spectrophotometer (Thermo Scientific). Blanking was performed with Millipore filtered water, and concentrations were reported in ng/μL.

2.2.5 Microarray Gene Expression Analysis

Microarray gene expression data for 78 cell lines were acquired as previously described (Wilding et al. 2010). Total RNA was extracted from cultures of each cell line using the RNeasy Mini Kit as described above. 20 μg of RNA for each sample were sent to the Molecular Biology Core Facility of the Paterson Institute for Cancer Research, where gene expression was performed using Affymetrix GeneChip Human Genome-U133 Plus 2.0 microarrays as per manufacturer's instructions. 500 ng of total RNA were reverse transcribed and converted to double stranded cDNA, and subsequently transcribed to make biotinylated cRNA. The cRNA was then fragmented, and 10 μg of the resultant mixture were applied to the microarray chips for hybridization. Resulting spot intensities for the 54,675 probesets on the chips were processed and normalized using the Robust Multi-array Average (RMA) algorithm, with GC correction and median polish to minimize noise, implemented in the Partek Genomics Suite software. Output expression measures were additionally processed to remove batch effects across the samples to give the final dataset used in downstream analyses.

2.3 Microscopy and Image Analysis

2.3.1 Fixation and Immunostaining

Upon completion of their 9-18 day culture times, the Matrigel cocultures were fixed and processed for immunostaining. Medium was carefully removed from the wells without disturbing the gels, and the cultures were washed once with 200 µL PBSA. Next, the wells were treated with 50 µL of 4% formol solution (VWR Chemicals, Ref. FOR0020AF59001) and fixed at 4°C for 20 min. The formol was removed, and the samples were then treated with 50 µL of Triton X-100 solution (Sigma-Aldrich, CAS No. 9002-93-1) diluted in PBSA to 0.2%, and permeabilized at 4°C for 5 min. Following that, the wells were washed four times with 200 µL of a 50 mM glycine (Sigma, Ref. G7126) solution in PBSA. Finally, the wells were covered with 50 µL of a 2% FBS in PBSA solution and left overnight at 4°C for blocking.

Following the blocking step, the samples were treated with primary antibodies diluted in PBSA. **Table 2.5** summarizes the primary antibodies used in this thesis, as well as the dilutions at which they were used. Co-staining was always performed with a pair of primary antibodies derived from two species, one from mouse and one from rabbit, or with a single primary antibody along with phalloidin-rhodamine to mark filamentous actin (F-actin). Samples were left overnight at 4°C with 50 µL of primary antibody solution. The next day, the samples were washed three times with 200 µL PBSA prior to the addition of the secondary antibody solution. Two secondary antibodies were used as necessary in all experiments – an antibody against mouse IgG grown in goat and conjugated to the Alexa 488 fluorophore (488-GAM) (Invitrogen, Ref. A11001), and an antibody against rabbit IgG grown in donkey and conjugated to

the Alexa 555 fluorophore (555-DAR) (Invitrogen, Ref. A31572). After overnight incubation at 4°C with the secondary antibodies, both used at a 1:200 dilution, the samples were again washed three times with 200 µL PBSA and treated with 50 µL of a 1:10,000 solution of 4',6-diamidino-2-phenylindole (DAPI) (Invitrogen, Ref. D1306) in PBSA immediately prior to imaging.

Primary Antibody	**Source**	**Dilution**	**Secondary**
EpCAM AUA1 mAb	in-house	1:100	488-GAM
Vimentin 3B4 mAb	Dako, Ref. M7020	1:100	488-GAM
Vimentin SP20 mAb	Spring Bioscience, Ref. M3202	1:200	555-DAR
Laminin Ab-1 pAb	Thermo Scientific, Ref. RB-082-A0	1:800	555-DAR
Laminin pAb	Abcam, Ref. ab11575	1:800	555-DAR
E-Cadherin 24E10 mAb	Cell Signaling Technology, Ref. 3195S	1:200	555-DAR
Reagent	**Source**	**Dilution**	
phalloidin-rhodamine	Sigma-Aldrich, Cat. No. P1951	1:1000	

Table 2.5: Reagents used for sample staining in this thesis, along with their sources, treatment dilutions, and corresponding secondary antibody where appropriate.

2.3.2 Epifluorescence Imaging

Epifluorescence imaging was performed using a Zeiss AxioObserver.Z1 microscope controlled by the ZEN Digital Imaging Software (Carl Zeiss AG, Germany). Filter sets enabled three-color imaging of multilabeled samples, allowing for the detection of two proteins of interest along with the DAPI nuclear stain.

2.3.3 Image Analysis

Images were prepared in the ZEN Software and exported for further processing using the Fiji package based on ImageJ. Scoring of the myofibroblast recruitment assay was performed subjectively on a five-point scale by three independent assessors. Cell culture numbers and areas were quantified in Fiji by binarizing the DAPI channel image, removing background spots by eroding and dilating, and then analyzing the spot numbers and areas using the built-in measurement tool.

2.4 Statistical Analysis

2.4.1 Continuous Gene Expression Analysis

Analysis of microarray gene expression for colorectal cancer cell lines was performed using Partek Genomics Suite software. Myofibroblast recruiting and non-recruiting colorectal cancer cell lines were directly compared by gene-by-gene Student's t-test, and output volcano plots were plotted as $\log_2$-transformed fold change against $-\log_{10}$-transformed uncorrected p-value. Multiple comparison p-value correction and determination of significance is not reported as none of the genes reached significance after correction due to low sample numbers.

2.4.2 Categorical Association Analysis

Microarray gene expression data for colorectal cancer cell lines was imported into R for Gaussian Mixture Modeling (GMM) and downstream association analysis. The public dataset of RNA-Seq data from 622 primary colorectal adenocarcinomas was obtained from the Genomic Data Commons (GDC) Data Portal of The Cancer Genome Atlas (TCGA). The TCGAbiolinks R package was used to query samples from the TCGA Colon Adenocarcinoma (COAD) and Rectum Adenocarcinoma (READ) databases. GMM was implemented using the mclust R package, and χ^2 tests were performed using the base R package. The interactive online interface for querying the associations calculated from categorical data derived from GMM was built using the shiny R package. All code is reported in the appendix.

Chapter Three

Myofibroblast Effects on Colorectal Cancer Proliferation, Morphology, and Differentiation

3.1 Introduction

The microenvironment is a fundamental determinant of tumor behavior. Myofibroblasts make up a prevalent cellular component of the stroma in both normal and cancerous tissues. As effectors of structural tissue rearrangements in wound healing and other processes, myofibroblasts interface with their neighboring cells by secreting a number of signaling molecules and extracellular matrix (ECM) components. The sources of myofibroblasts are thought to vary significantly, from the activation of local fibroblasts to the differentiation of circulating stem cells or epithelial-mesenchymal transition (EMT) of local epithelial cells (McAnulty 2007). A number of tissues also exhibit sizable myofibroblastic populations at steady state, hinting that this definition is likely to encompass a heterogenous family of cell types (Mifflin et al. 2011). In solid cancers, myofibroblast invasion is broadly reported as a predictor of more severe disease, a phenomenon that may be a result of signaling interactions between the myofibroblasts and the tumor cells (Liu et al. 2016).

The most apparent function of myofibroblasts in normal tissue occurs during wound healing. In the skin, myofibroblasts appear as components of granulation tissue following the initial inflammatory phase of healing, and are alleged to derive mostly from fibroblasts activated with transforming growth factor-β1 (TGF-β1) and platelet-derived growth factor (PDGF) (Darby et al. 2014; Singer & Clark 1999). In this context, myofibroblasts are reported to participate in ECM remodeling, and are also known to release a range of cytokines and chemokines, ostensibly to further promote the inflammatory response (Baum & Duffy 2011). Cancer-associated Fibroblasts (CAFs), prominent constituents of the stromal reaction to carcinomas, also exhibit the

myofibroblast phenotype of α-smooth muscle actin expression and secrete a range of immunomodulatory and other cytokines and chemokines (Hinz et al. 2012). Similarly to myofibroblasts in general, various sources of CAFs have been proposed, and significant heterogeneities are also present between and sometimes within individual tissues. Intriguingly, intestinal pericryptal myofibroblasts, which form a layer directly below the epithelium in normal intestine, have recently been implicated in the maintenance of the intestinal stem cell niche through the secretion of various signaling factors (Kabiri et al. 2014). This further highlights the heterogenous nature of this cell type and suggests that normal, locally resident myofibroblasts may already be present in the carcinoma microenvironment without the need of any activation of fibroblasts, endogenously secreting factors that promote early tumor development.

This chapter applies *in vitro* 3D culture techniques to probe the effects of myofibroblasts on colorectal cancer cell lines, namely their proliferation, morphology and differentiation. Cocultures of a panel of colorectal cancer cell lines with intestinal myofibroblast and skin fibroblast cell lines are established, and the phenotypes of the resulting structures are discussed. A panel of colorectal cancer cell lines is also grown with conditioned media from myofibroblast and skin fibroblast cell lines, and effects on proliferation and morphology are discussed. Gene expression analysis of the myo/fibroblasts is performed to determine a set of secreted factors possibly responsible for the effects of the conditioned media, and further drug and siRNA studies are done in an effort to block the effects of the myofibroblasts on the colorectal cancer cells.

3.2 Results

3.2.1 Cocultures on a Flat Layer of Myo/fibroblasts

In order to assess the possible influence of myofibroblasts and fibroblasts on colorectal cancer cells, a simple coculture assay was used. In brief, a confluent layer of cultured CCD-18Co myofibroblasts or in-house derived skin fibroblasts was grown in a 96-well culture plate, and 1000 dissociated colorectal cancer cells were subsequently added on top of that layer in 50% Matrigel (**Figure 3.1**). The presence of fetal bovine serum (FBS), a common culture medium supplement providing a range of nutrients and growth factors, was varied to compare a more permissive serum-containing environment that would allow maximal baseline cell survival to a restrictive serum-free medium, in which any ligands affecting the survival of the cancer cells would be expected to come from the myo/fibroblasts rather than the medium, and vice versa. The cocultures (or monocultures, without the layer of myo/fibroblasts) were grown for 18 days in serum-free or in 1% serum-containing culture medium prior to brightfield imaging to allow for optimal colony formation. A panel of ten colorectal cancer cell lines (**Table 3.1**) was used, and the results for six representative lines are shown in **Figure 3.2**.

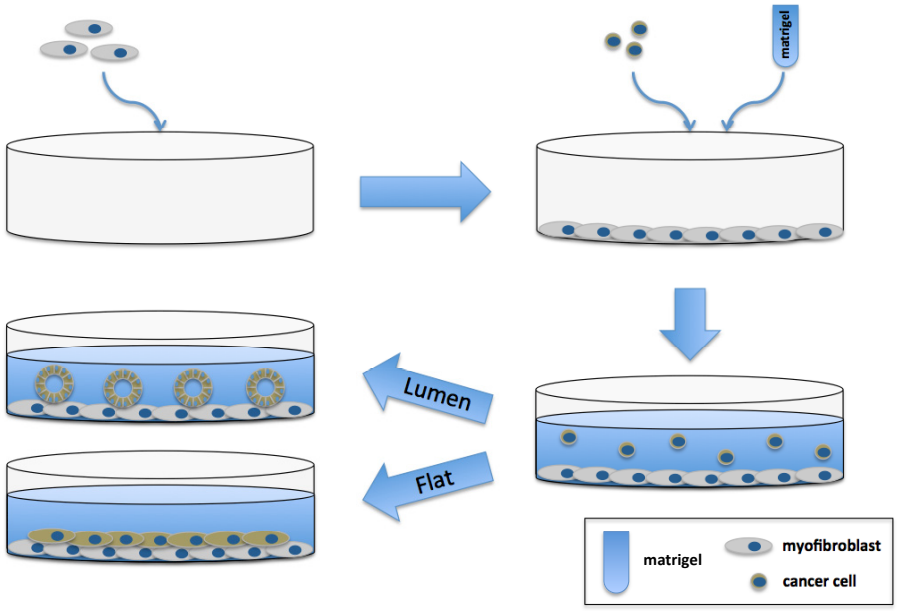


Figure 3.1: Outline of the layer coculture assay. As shown here, a layer of myofibroblasts (or skin fibroblasts) is first grown to confluence in the culture well. At this point, a population of colorectal cancer cells dispersed in 50% Matrigel is added atop the myo/fibroblast layer. Once the gel hardens, culture medium of desired serum content is added on top. Over the culture time, some colorectal cancer cell lines form structures containing some or all differentiated cell types of the colon in a distinct arrangement around a central lumen. Other cell lines do not differentiate, and instead form compact spheres in the gel or settle flat on top of the layer below.

Colony morphology: **Mismatch repair status:**	**Lumen-forming**	**Intermediate morphology**	**Not lumen-forming**	**Unknown**
Mismatch repair proficient	CACO2 SW1222	HT29	SW480 SW620	
Mismatch repair deficient	LS174T LS180		HCT116	RKO
Unknown		OUMS23		

Table 3.1: Cell lines used in the layer coculture assay, shown based on colony morphology in 3D culture and mismatch repair pathway functionality (Bracht et al. 2010; Ashley et al. 2013).

SF
alone
+MFs
+Skin Fibros
LS174T
LS180
HT29

SF
alone
+MFs
+Skin Fibros
HCT116
SW1222
SW620

1%
alone
+MFs
+Skin Fibros
LS174T
LS180
HT29

1%
alone
+MFs
+Skin Fibros
HCT116
SW1222
SW620

Figure 3.2 (previous four pages): Images for six colorectal cancer cell lines, labeled by name, of a panel of ten tested by the layer coculture assay (**Table 3.1**). Cells were grown either in serum-free (SF) or 1% serum-containing media, as indicated in the top corner of each page, alone or atop a layer of myofibroblasts (MFs) or skin fibroblasts. Scale bar: 500 μm

---

First off, it can be seen from the serum-free images that myofibroblasts, and to a lesser extent fibroblasts, induce substantial increases in colorectal cancer cell survival and/or growth in comparison to the cancer cells cultured alone. Notably, this effect is also clearly present in the 1% serum-containing cultures for all cell lines except SW1222. This suggests that both myofibroblasts and fibroblasts may be secreting or expressing on their membranes certain growth-promoting factors that are either absent from serum or present at insufficient concentrations. Another notable observation is the effect of coculture on colony morphology, shown at higher magnification for the LS180 cell line in **Figure 3.3**. Some colorectal cancer cell lines are known to form luminal structures in Matrigel, which consist of a differentiated layer of epithelial cells around a cavity mimicking the inside of the colonic crypt (Yeung et al. 2010). Note how myofibroblasts promote the formation of polyluminal clusters of LS180 cells, while skin fibroblast coculture results in large single lumens with reduced clustering, both in serum-free and 1% serum (**Figures 3.2 and 3.3**). Additionally, while LS180 grown alone in 1% serum form some polyluminal clusters extending into the gel, the presence of myofibroblasts seems to promote flatter cultures growing along the bottom layer both in serum-free and 1% medium, suggesting a possible importance of heterotypic cell-cell contact. Myofibroblasts also seem to promote more elongated cultures of the cancer cells, in LS180 (**Figure 3.3**, serum-free) and even more strikingly in SW1222 (**Figure 3.2**, 1% serum). These observations indicate a positive effect of colonic myofibroblasts on colorectal cancer cell proliferation, as well as a possible influence on lumen formation and differentiation.

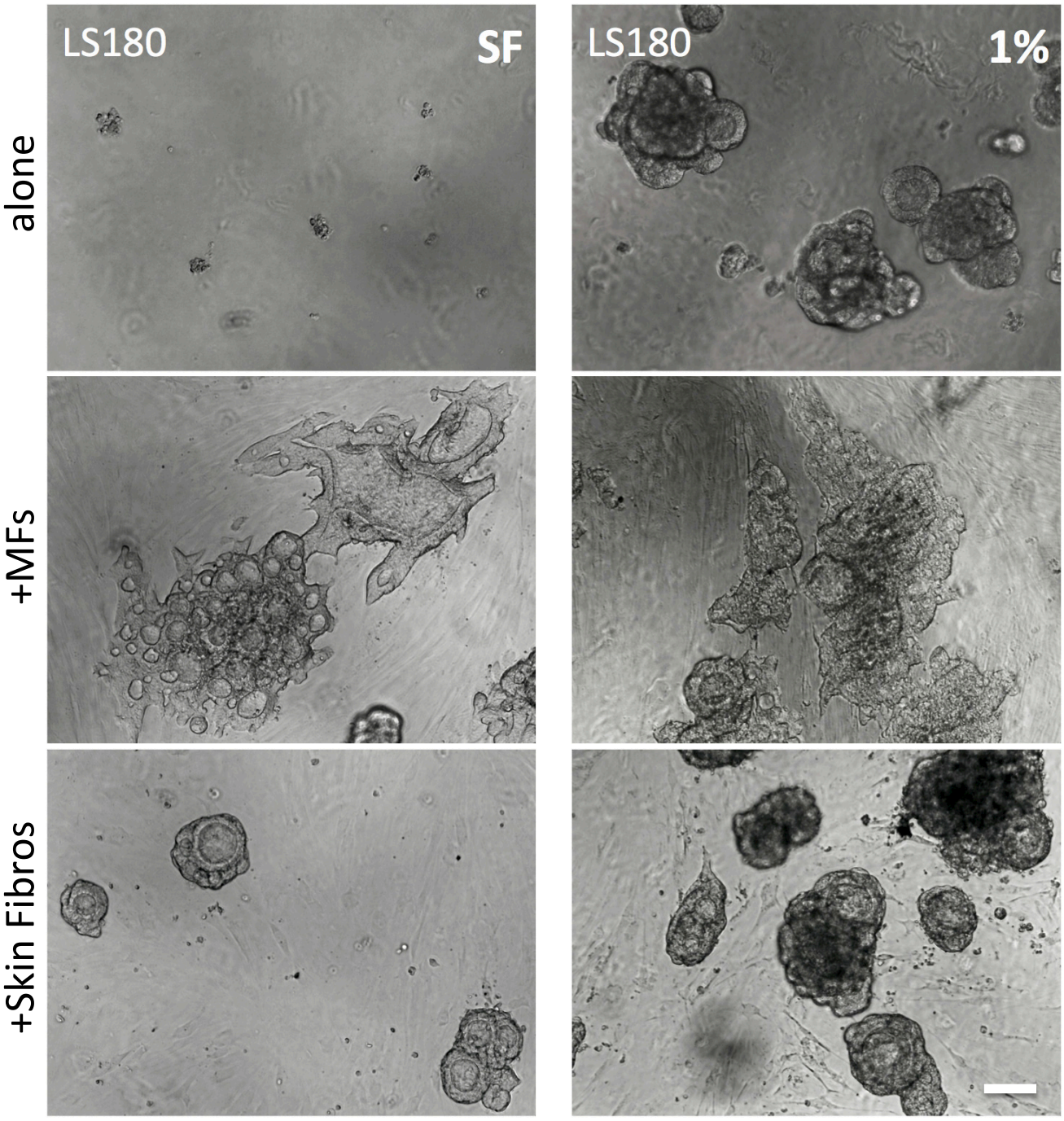
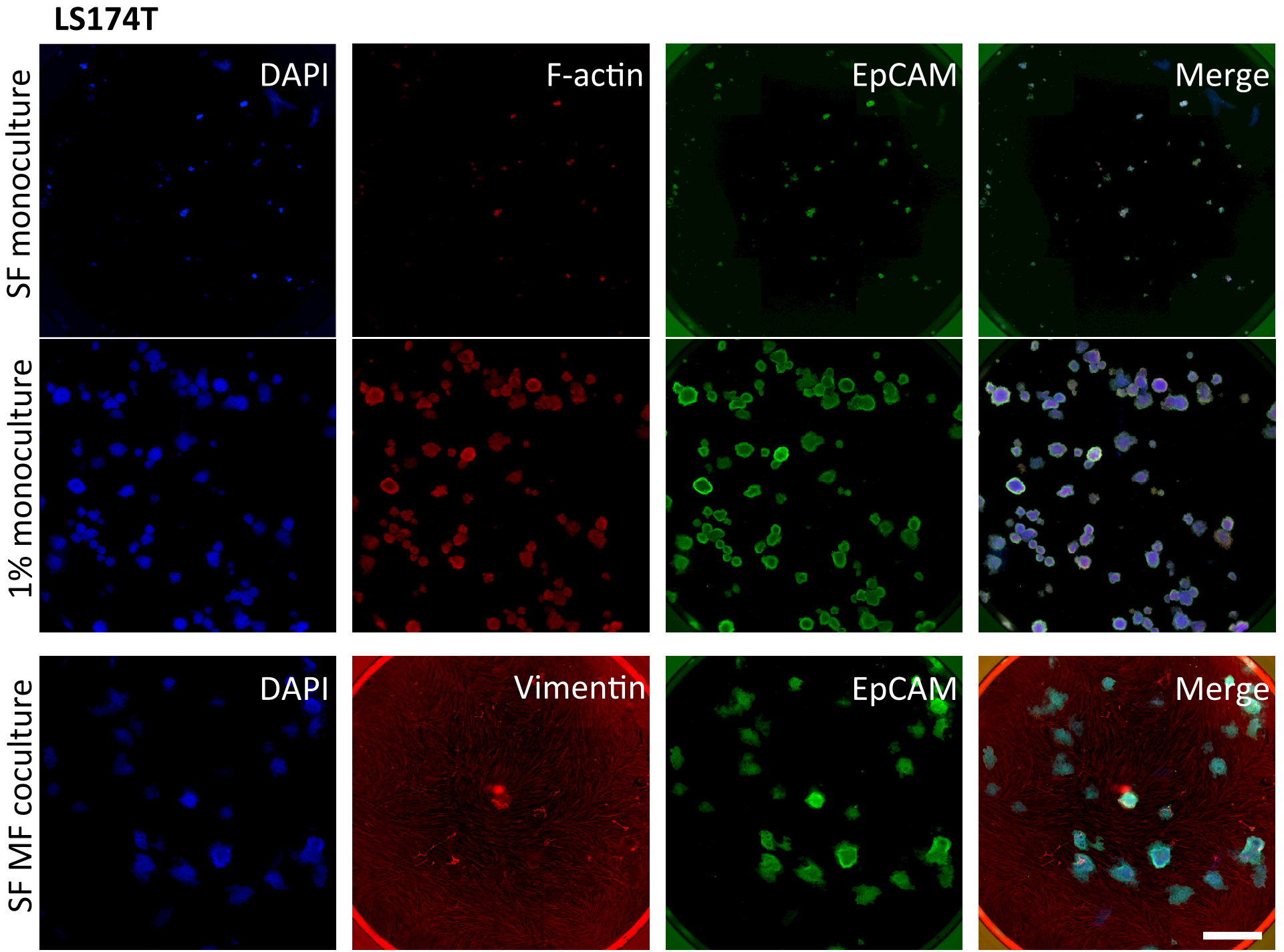
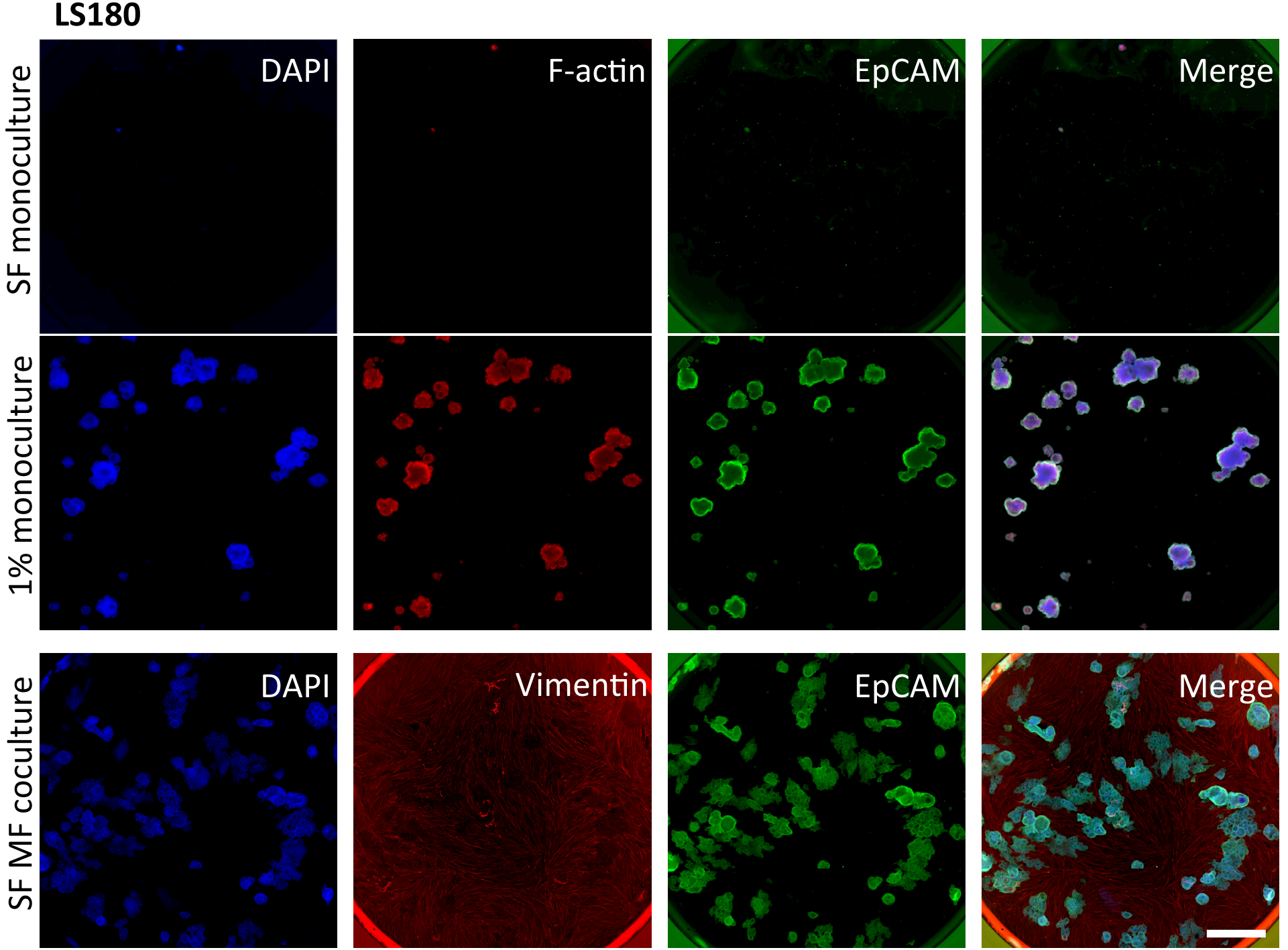


Figure 3.3: Zoomed brightfield images of LS180 colonies in Matrigel, grown alone or atop confluent layers of myofibroblasts (MFs) or skin fibroblasts, in serum-free (SF) or 1% serum-containing culture medium as indicated**.** Scale bar: 125 µm

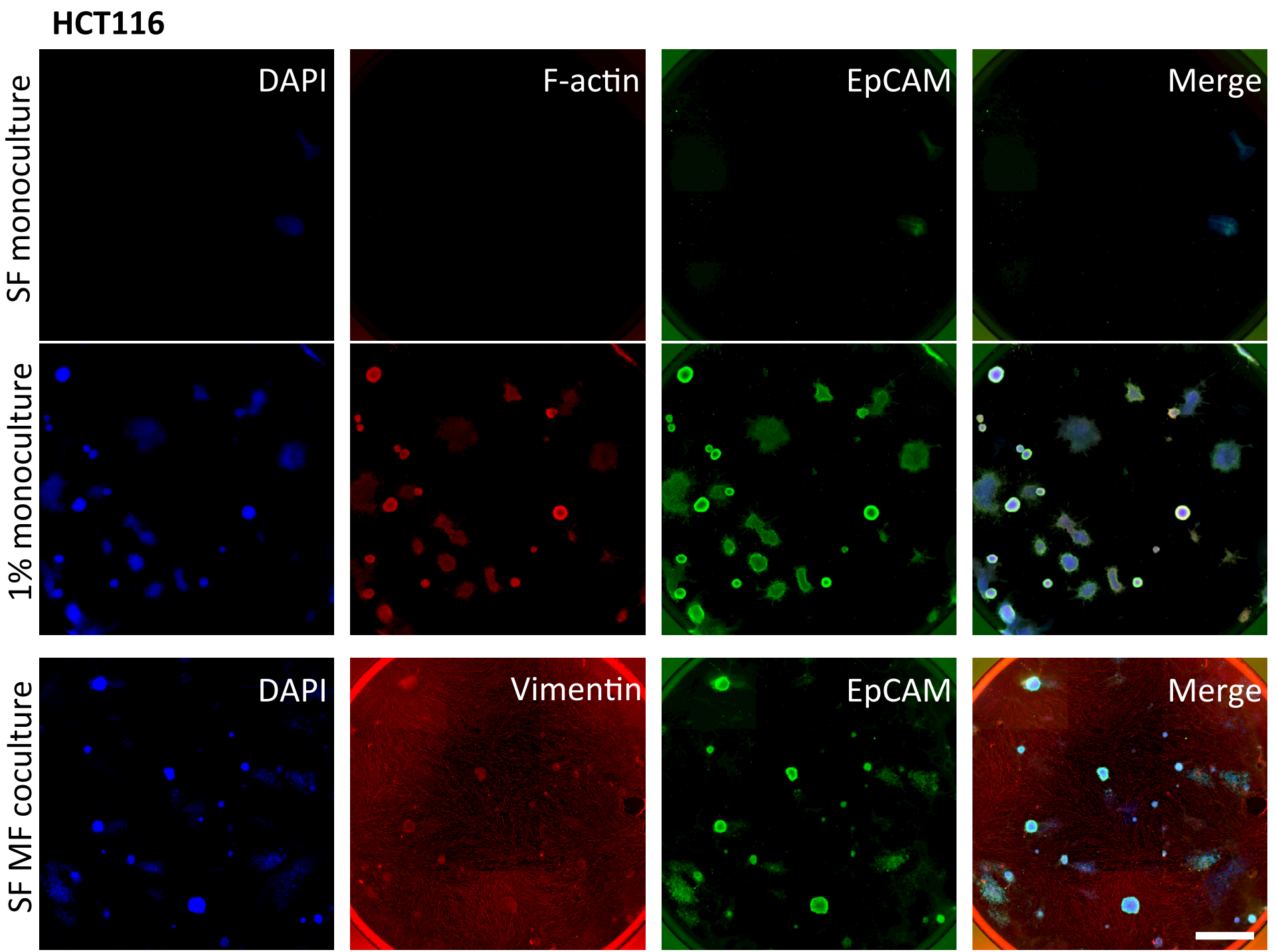
3.2.2 Immunostaining of Mono and Cocultures

To further examine the influence of myofibroblasts on colorectal cancer cells, five cell lines from the original panel of ten with striking coculture phenotypes (LS174T, LS180, HT29, HCT116, and SW1222) were again subjected to layer mono/coculture, this time comparing serum-free and 1% serum monocultures to serum-free coculture with myofibroblasts and staining for various protein markers (**Figure 3.4**). Specifically, the monocultures were stained for EpCAM, an epithelial cell surface protein, and F-actin, which is known to form distinct rings around luminal structures that form from differentiating cell lines and mimic the structure of the colonic crypt; cocultures were stained for vimentin instead of F-actin to mark the myofibroblast layer (Ashley et al. 2013). As expected, every cell line except SW1222 exhibited strikingly improved survival and proliferation in 1% serum compared to serum-free media in monoculture. Myofibroblasts also induced ample proliferation in the epithelial cells, even in the absence of serum. Some elongated morphology was again noted in the LS174T and LS180 colonies atop myofibroblasts, and the polyluminal clusters of LS180 again extended more vertically into the Matrigel in monoculture while forming flat along the bottom of the well in coculture. Most strikingly, while HT29 cells formed defined spherical colonies in monoculture, the cells were widely scattered in coculture with only a few dense foci, suggesting an effect of coculture on colony formation and cell differentiation. These images confirmed and refined the original observations, suggesting that myofibroblasts influence not only survival and proliferation but also differentiation and colony morphology in colorectal cancer cell lines.


LS174T
SF monoculture
1% monoculture
SF MF coculture
DAPI
F-actin
EpCAM
Merge
DAPI
Vimentin
EpCAM
Merge


LS180
SF monoculture
DAPI
F-actin
EpCAM
Merge
1% monoculture
SF MF coculture
DAPI
Vimentin
EpCAM
Merge

HT29
DAPI
F-actin
EpCAM
Merge
Vimentin
Merge
SF monoculture
1% monoculture
SF MF coculture


HCT116
SF monoculture
1% monoculture
SF MF coculture
DAPI
F-actin
EpCAM
Merge
DAPI
Vimentin
EpCAM
Merge

SW1222
DAPI
F-actin
EpCAM
Merge
DAPI
Vimentin
EpCAM
Merge
SF monoculture
1% monoculture
SF MF coculture

Figure 3.4 (previous five pages): Comparison of serum-free (SF) and 1% serum monoculture to myofibroblast (MF) coculture for five cell lines. All cell lines show beneficial effects of serum and myofibroblasts on survival and proliferation, and cocultures additionally exhibit altered colony morphologies compared to monoculture. Scale bar: 1 mm

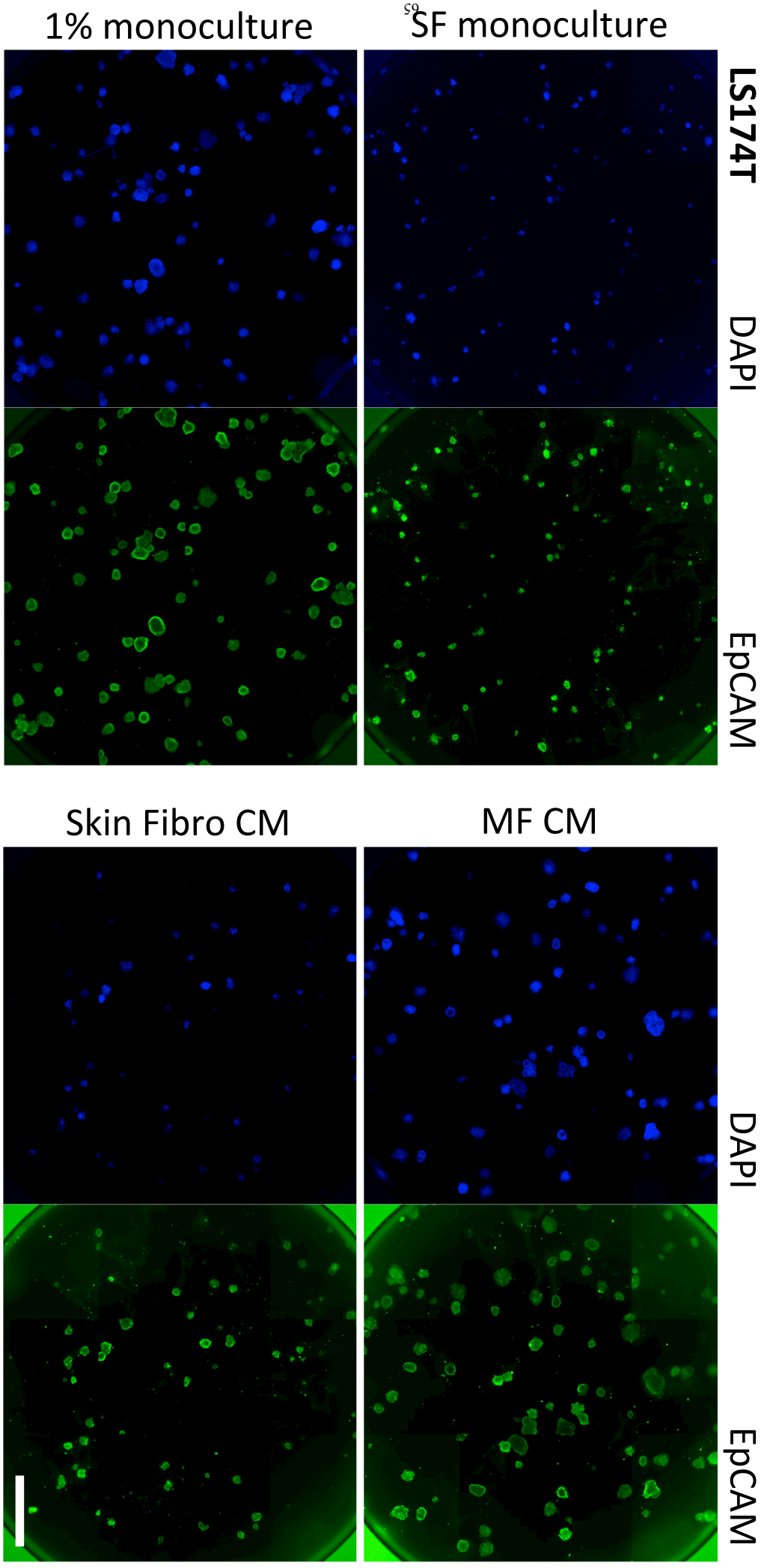
3.2.3 Colorectal Cancer Cell Lines in Myo/fibroblast Conditioned Media

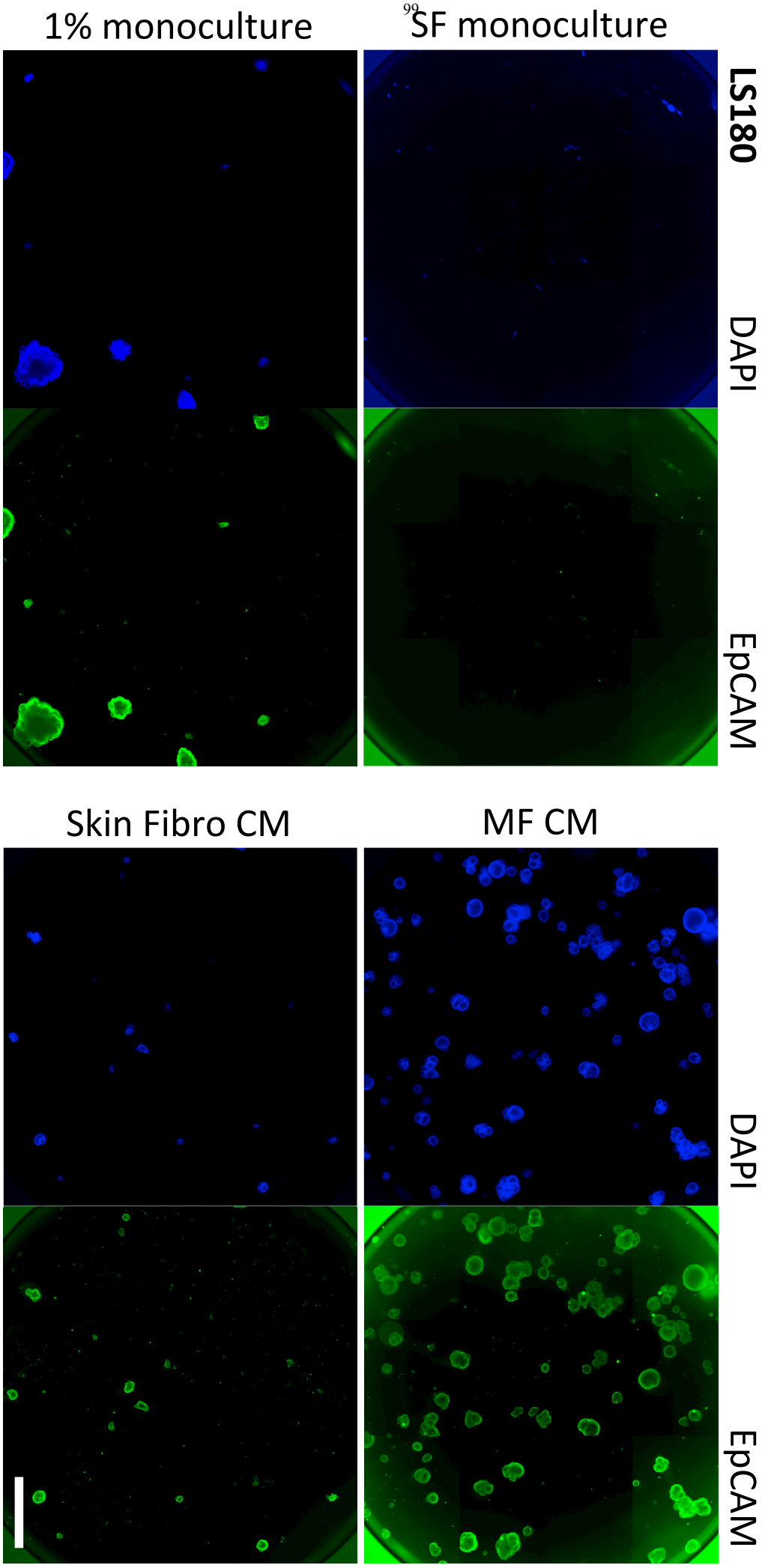
One significant caveat of the layer coculture assay is its inability to distinguish if the observed effects of myofibroblasts on cancer cells are due to secreted or membrane-bound signals. In order to isolate the influence of secreted factors, colorectal cancer cell monocultures in Matrigel were treated with conditioned media (CM) from myofibroblast and fibroblast cell lines. Conditioned media were derived by growing myo/fibroblasts to full confluence in large culture flasks in 10% serum-containing media, and then washing them and replacing with serum-free media for three days. These 72-hour conditioned serum-free media were harvested, filtered to remove debris, and used directly in subsequent experiments in lieu of fresh media. Conditioned media were prepared from the same lines as those used in the layer cocultures – cultured CCD-18Co colonic myofibroblasts and one skin fibroblast line prepared in-house. The same ten colorectal cancer cell lines as previously were grown in Matrigel for 18 days in the various conditioned media, as well as serum-free and 1% serum DMEM, and the results for five lines that exhibited good proliferation in at least one condition are shown in **Figure 3.5**.

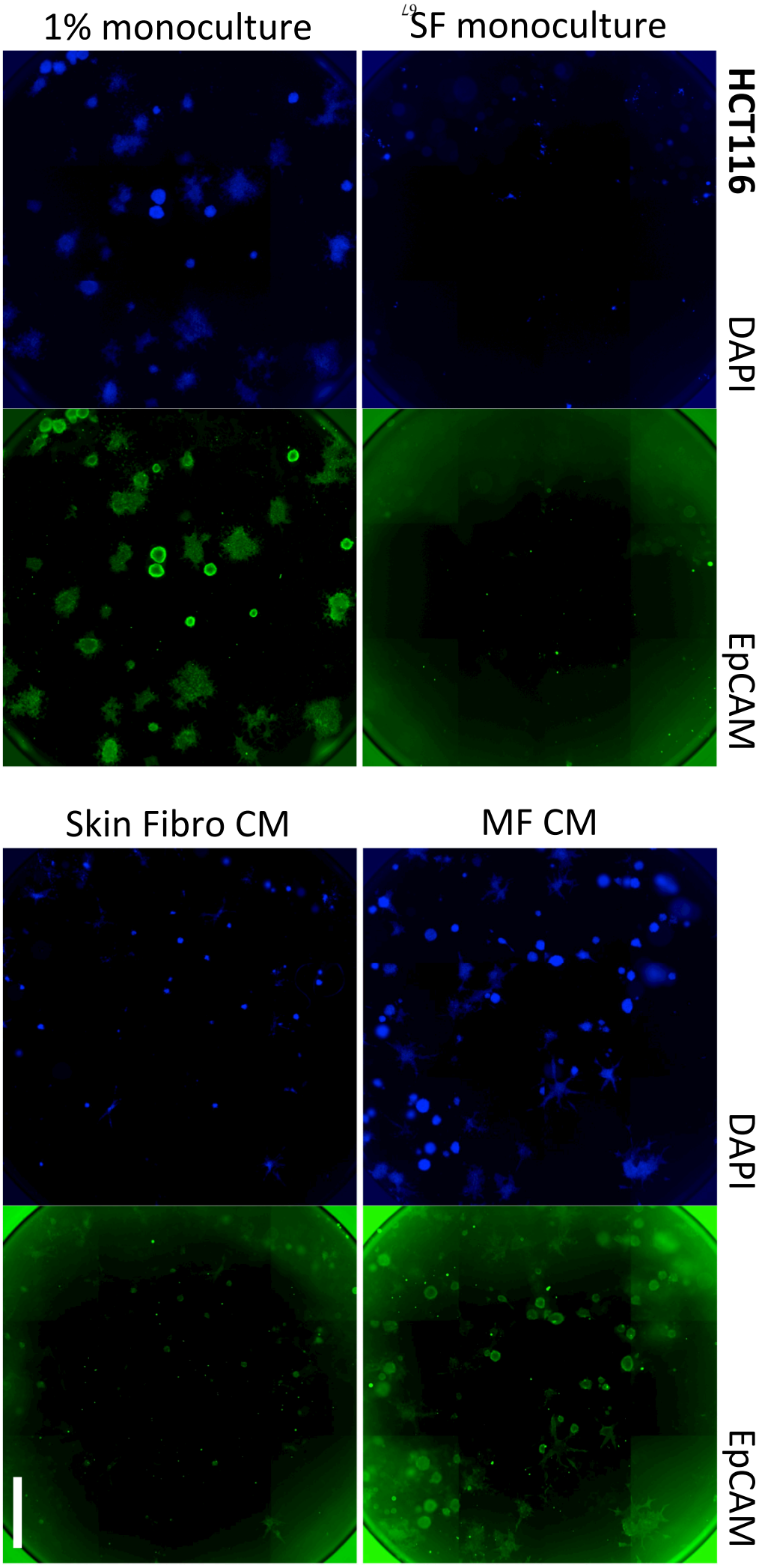
First of all, most cell lines again proliferated very poorly in serum-free unconditioned media, with only LS174T and OUMS23 exhibiting modest cell numbers. On the other hand, every cell line shown survived and proliferated at least to some extent in 1%

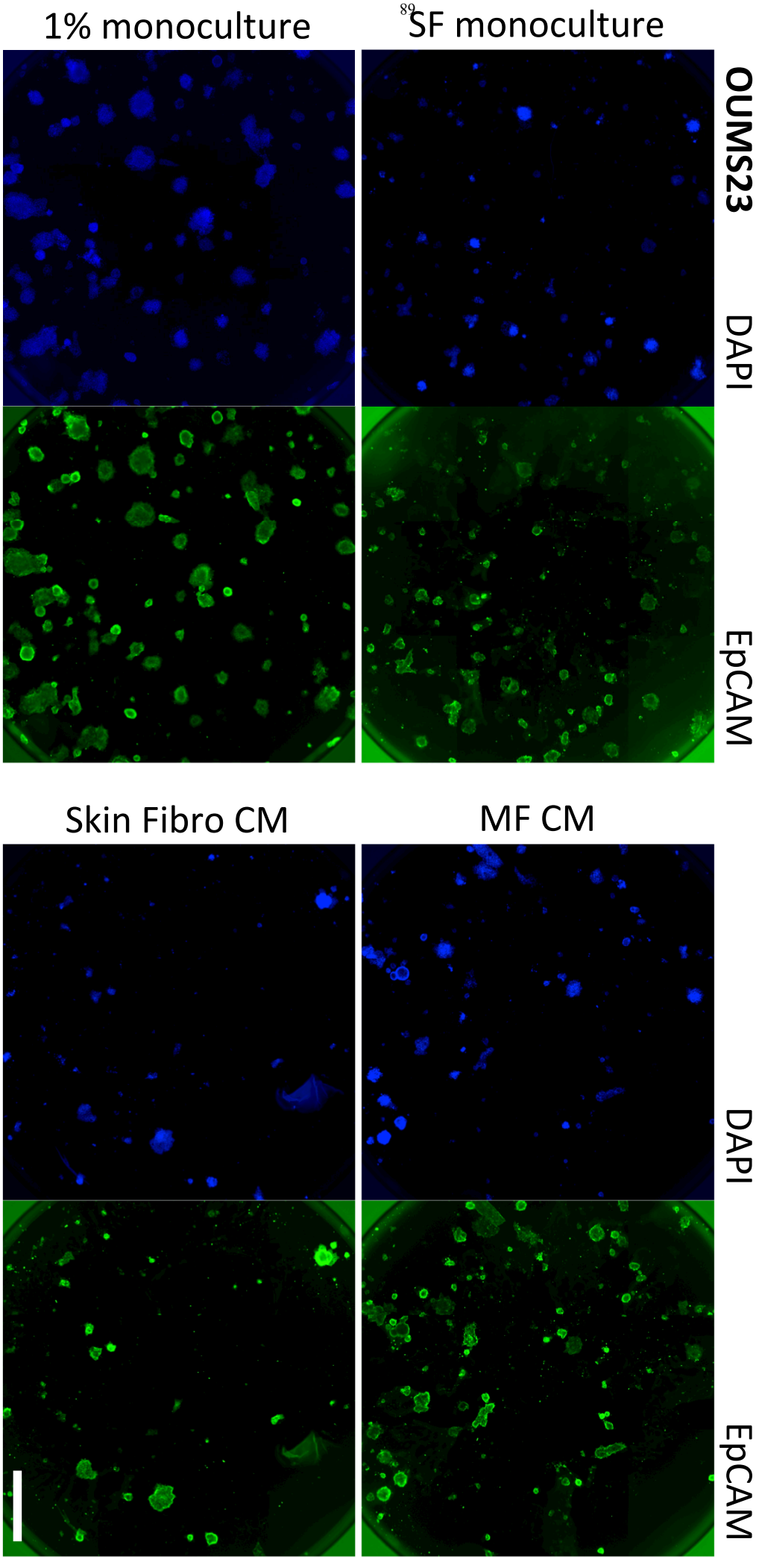
serum, and to a similar degree in myofibroblast CM. There was strikingly more proliferation in myofibroblast CM compared to skin fibroblast CM, with the latter performing only slightly better than serum-free medium (**Figure 3.6**). No obvious colony elongation akin to that seen in cocultures was observed. It was concluded that some effects of myofibroblasts on colorectal cancer cells might be due to secreted factors that replace and/or augment the effects of serum.

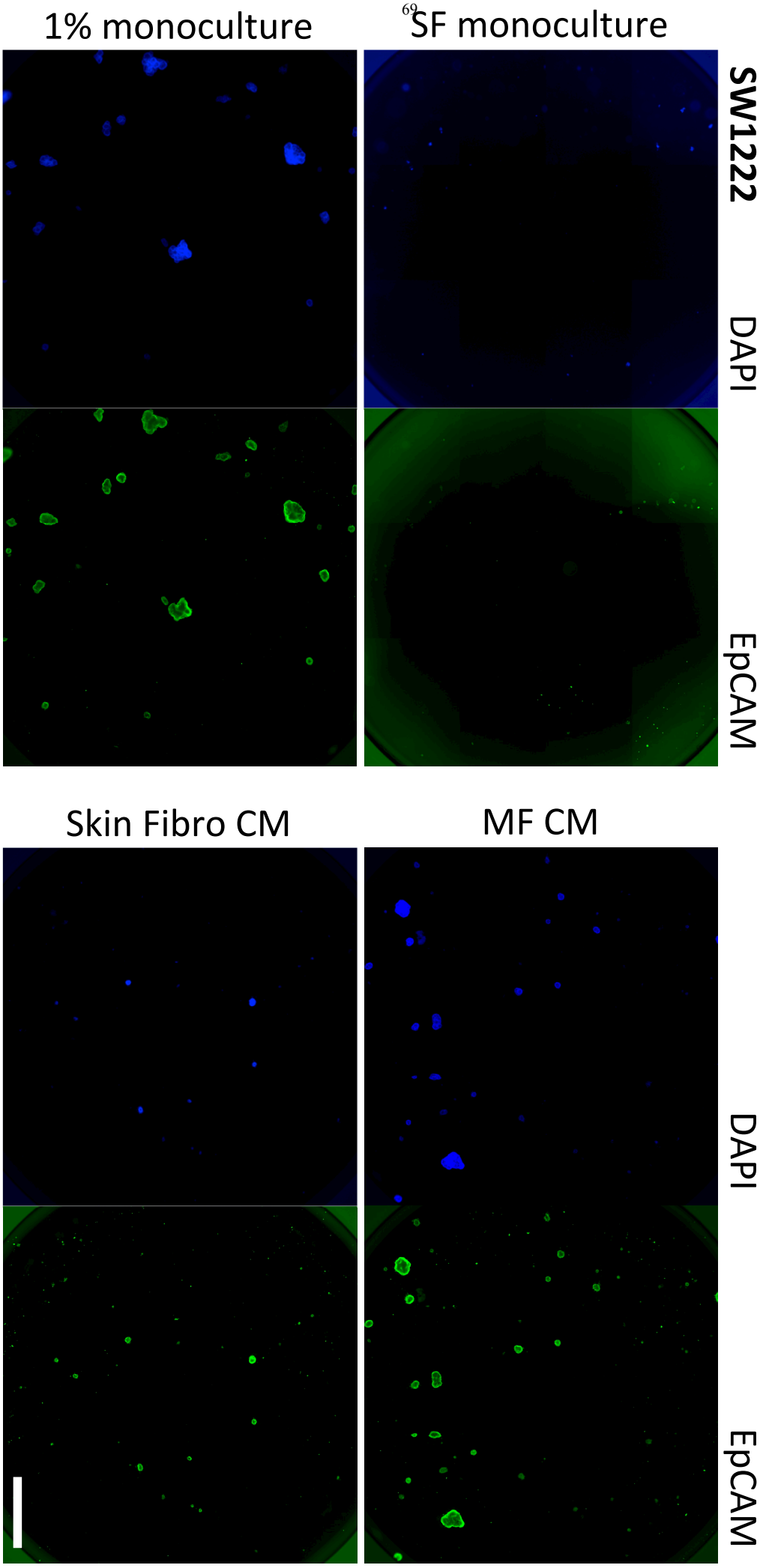
Figure 3.5 (following five pages): Effects of myofibroblast and skin fibroblast conditioned media on colorectal cancer cells grown in monoculture in Matrigel. In every case, a beneficial effect of myofibroblast CM on cancer cell survival and proliferation is evident and comparable to that of 1% serum. Scale bar: 1 mm


1% monoculture
SF monoculture
LS174T
DAPI
EpCAM
Skin Fibro CM
MF CM
DAPI
EpCAM


1% monoculture
99% SF monoculture
LS180
DAPI
EpCAM
Skin Fibro CM
MF CM
DAPI
EpCAM


1% monoculture
SF monoculture
HCT116
DAPI
EpCAM
Skin Fibro CM
MF CM
DAPI
EpCAM


1% monoculture
SF monoculture
OUMS23
DAPI
EpCAM
Skin Fibro CM
MF CM
DAPI
EpCAM


1% monoculture
69SF monoculture
SW1222
DAPI
EpCAM
Skin Fibro CM
MF CM
DAPI
EpCAM

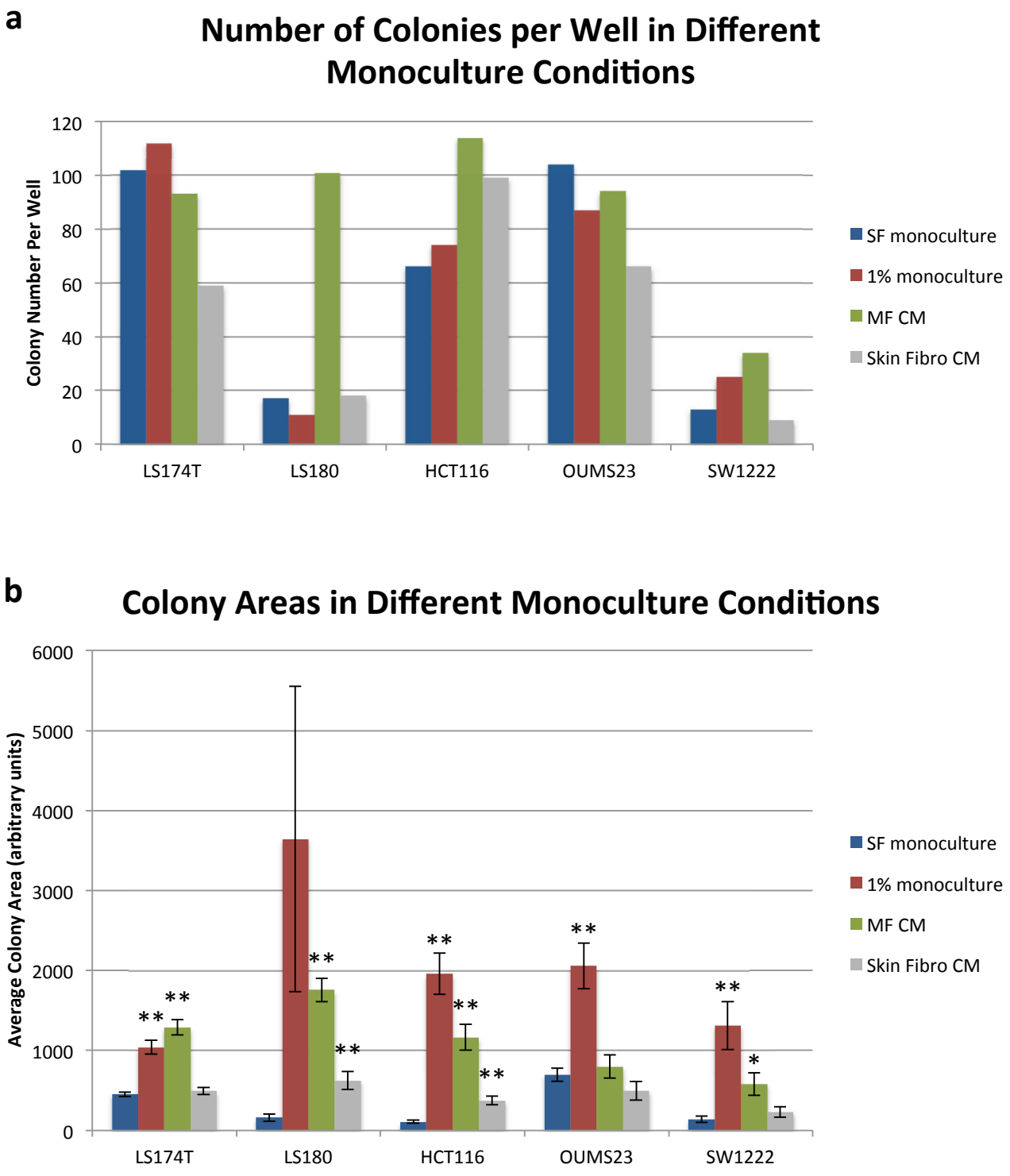


Figure 3.6: Quantification of the monoculture experiment in **Figure 3.5**. (**a**) Number of colonies per well for each cell line and each treatment. Note especially the effect of myofibroblast CM on colony formation of LS180, as well as to a milder extent HCT116 and SW1222, which in some cases outperforms 1% serum-containing medium. (**b**) Average colony area for each cell line and each treatment, reporting on cell proliferation. Note that 1% serum results in substantially increased cell proliferation for all cell lines, as does myofibroblast CM for most of them, while skin fibro CM has almost no improvement over serum-free (SF) medium.
*p<0.01 **p<0.001 (Two-tailed Student's t-test), compared to corresponding SF monoculture control

Microarray gene expression data for the CCD-18Co myofibroblasts and skin fibroblasts, grown in monoculture on plastic in 10% serum-containing DMEM, were available from previous studies in the lab. This dataset was pared down to focus only on factors secreted from the myo/fibroblasts using the secreted/membrane protein annotations from the Human Protein Atlas; expression differences between myofibroblasts and skin fibroblasts were calculated for each probeset, and the results were ordered by decreasing difference (Uhlen et al. 2015). As shown in **Table 3.2**, the top genes expressed by myofibroblasts and not skin fibroblasts include endothelin receptor types A and B (*EDNRA*, *EDNRB*), leucine rich repeat containing 17 (*LRRC17*), and NK2 transcription factor related, locus 3 (*NKX2-3*), all previously reported as markers of myofibroblasts over activated fibroblasts (Hsia et al. 2016). Additionally, the myofibroblasts express insulin-like growth factor 2 (*IGF2*), hepatocyte growth factor (*HGF*), and bone morphogenetic protein 2 (*BMP2*), ligands known to influence colorectal cancer growth and migration (Zanella et al. 2015; Jiang et al. 1993; Zhang et al. 2014). Notably, endothelin receptor ligand endothelin 1 (*EDN1*), as well as the IGF2 receptor *IGF1R*, the HGF receptor *MET*, and the BMP2 receptor *BMPR1A*, are expressed in subsets if not in all colorectal cancer cell lines in a panel of over 100 that are cultured in our lab (data not shown), strongly suggesting that these signaling axes may be responsible for the interactions between myofibroblasts and colorectal cancer cells.

Affymetrix Probeset ID	Gene Symbol	Gene Title	Secreted Annotation	CCD-18Co Expression	Skin Fibro Expression	Difference
202409_at	IGF2 /// INS-IGF2	insulin-like growth factor 2 (somatomedin A) /// INS-IGF2 readthrough transcript	Yes	11.4712	3.90287	7.56833
219478_at	WFDC1	WAP four-disulfide core domain 1	Yes	11.0276	3.47531	7.55229
206701_x_at	EDNRB	endothelin receptor type B	Yes	10.4901	3.06224	7.42786
226322_at	TMTC1	transmembrane and tetratricopeptide repeat containing 1	Yes	9.19849	1.96498	7.23351
207420_at	COLEC10	collectin sub-family member 10 (C-type lectin)	Yes	9.73396	2.66435	7.06961
220532_s_at	TMEM176B	transmembrane protein 176B	Yes	10.4326	3.85294	6.57966
205381_at	LRRC17	leucine rich repeat containing 17	Yes	10.1648	3.58847	6.57633
202410_x_at	IGF2 /// INS-IGF2	insulin-like growth factor 2 (somatomedin A) /// INS-IGF2 readthrough transcript	Yes	10.64	4.26423	6.37577
229485_x_at	SHISA3	shisa homolog 3 (Xenopus laevis)	Yes	8.24713	1.96498	6.28215
204271_s_at	EDNRB	endothelin receptor type B	Yes	11.3421	5.1593	6.1828
210432_s_at	SCN3A	sodium channel, voltage-gated, type III, alpha subunit	Yes	8.20602	2.0825	6.12352
244506_at	TMTC1	transmembrane and tetratricopeptide repeat containing 1	Yes	9.76088	3.67715	6.08373
218345_at	TMEM176A	transmembrane protein 176A	Yes	10.5524	4.55722	5.99518
218717_s_at	LEPREL1	leprecan-like 1	Yes	11.445	5.48287	5.96213
209961_s_at	HGF	hepatocyte growth factor (hepapoietin A; scatter factor)	Yes	10.8915	5.03585	5.85565
233882_s_at	SEMA6D	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6D	Yes	10.5132	4.7016	5.8116
224397_s_at	TMTC1	transmembrane and tetratricopeptide repeat containing 1	Yes	9.46795	3.74211	5.72584
208600_s_at	GPR39	G protein-coupled receptor 39	Yes	10.4841	4.85985	5.62425
212909_at	LYPD1	LY6/PLAUR domain containing 1	Yes	10.9706	5.41698	5.55362
209960_at	HGF	hepatocyte growth factor (hepapoietin A; scatter factor)	Yes	8.40144	2.97885	5.42259
204818_at	HSD17B2	hydroxysteroid (17-beta) dehydrogenase 2	Yes	7.80883	2.44124	5.36759
209631_s_at	GPR37	G protein-coupled receptor 37 (endothelin receptor type B-like)	Yes	7.73069	2.43456	5.29613
210881_s_at	IGF2 /// INS-IGF2	insulin-like growth factor 2 (somatomedin A) /// INS-IGF2 readthrough transcript	Yes	10.5707	5.36348	5.20722
235666_at	ITGA8	integrin, alpha 8	Yes	9.38527	4.21597	5.1693
1553808_a_at	NKX2-3	NK2 transcription factor related, locus 3 (Drosophila)	Yes	9.23822	4.14091	5.09731
214596_at	CHRM3	cholinergic receptor, muscarinic 3	Yes	8.57376	3.50127	5.07249
205266_at	LIF	leukemia inhibitory factor (cholinergic differentiation factor)	Yes	9.96891	4.90744	5.06147
214265_at	ITGA8	integrin, alpha 8	Yes	10.1537	5.10611	5.04759
223449_at	SEMA6A	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A	Yes	9.27197	4.28547	4.9865
204273_at	EDNRB	endothelin receptor type B	Yes	7.30814	2.34292	4.96522
230788_at	GCNT2	glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (I blood group)	Yes	7.98212	3.05132	4.9308
209569_x_at	D4S234E	DNA segment on chromosome 4 (unique) 234 expressed sequence	Yes	9.15535	4.22809	4.92726
225660_at	SEMA6A	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A	Yes	8.51592	3.6299	4.88602
1553180_at	ADAMTS19	ADAM metallopeptidase with thrombospondin type 1 motif, 19	Yes	7.59909	2.71385	4.88524
209087_x_at	MCAM	melanoma cell adhesion molecule	Yes	10.2684	5.48106	4.78734
226492_at	SEMA6D	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6D	Yes	8.32838	3.57748	4.7509
204464_s_at	EDNRA	endothelin receptor type A	Yes	8.34549	3.59556	4.74993
210869_s_at	MCAM	melanoma cell adhesion molecule	Yes	8.83226	4.09344	4.73882
207414_s_at	PCSK6	proprotein convertase subtilisin/kexin type 6	Yes	9.16753	4.42958	4.73795
210998_s_at	HGF	hepatocyte growth factor (hepapoietin A; scatter factor)	Yes	9.96736	5.25018	4.71718
217590_s_at	TRPA1	transient receptor potential cation channel, subfamily A, member 1	Yes	8.99105	4.27731	4.71374
205290_s_at	BMP2	bone morphogenetic protein 2	Yes	9.87476	5.22143	4.65333
231192_at	LPAR3	lysophosphatidic acid receptor 3	Yes	6.57577	1.93107	4.6447
203980_at	FABP4	fatty acid binding protein 4, adipocyte	Yes	9.34231	4.71531	4.627
211340_s_at	MCAM	melanoma cell adhesion molecule	Yes	9.62	4.9986	4.6214
202037_s_at	SFRP1	secreted frizzled-related protein 1	Yes	9.76703	5.17921	4.58782
226931_at	TMTC1	transmembrane and tetratricopeptide repeat containing 1	Yes	9.0238	4.50835	4.51545
203680_at	PRKAR2B	protein kinase, cAMP-dependent, regulatory, type II, beta	Yes	7.71035	3.24272	4.46763
209583_s_at	CD200	CD200 molecule	Yes	6.3055	1.85625	4.44925
231969_at	STOX2	storkhead box 2	Yes	7.08571	2.733	4.35271

Table 3.2: Top fifty secreted protein coding genes upregulated in myofibroblasts compared to skin fibroblasts, ordered by expression difference. Log_2-transformed expression values are shown, meaning that the difference column corresponds to $\log_2$ of the fold difference in the expression of the particular gene between myofibroblasts and skin fibroblasts. Genes of especial interest, with well-known signaling functions and previous implications in myofibroblast-colorectal cancer cell communication, are highlighted in red.

3.2.4 Drug and siRNA Testing of the Candidate genes

In order to test the effects of the candidate myofibroblast-derived secretions identified in the previous section on colorectal cancer cells, a drug screen was performed using the layer coculture assay. Five drugs with various targets were chosen, as shown in **Table 3.3**, and tested on the smaller subset of five cell lines used previously (LS174T, LS180, HT29, HCT116, and SW1222). In addition to the signaling axes mentioned above, drugs were also included to target the CXCL12-CXCR4 axis previously reported to influence colorectal cancer cell proliferation, migration and invasion (Brand et al. 2005). Serum-free DMEM containing these drugs at the stated concentrations was added directly atop the hardened Matrigel layer, and the cocultures were grown for 18 days prior to staining for EpCAM and vimentin. **Figure 3.7** shows full-well images for all drugs, along with the serum-free control.

Drug	Target	Concentration
AMD3100	CXCR4	1 μM
BQ788	EDNRB	1 μM
Figitumumab (CP751851)	IGF1R	10 μg/mL
LDN193189	BMPR1A/ACVR1	500 nM
PHA665752	c-Met	100 nM

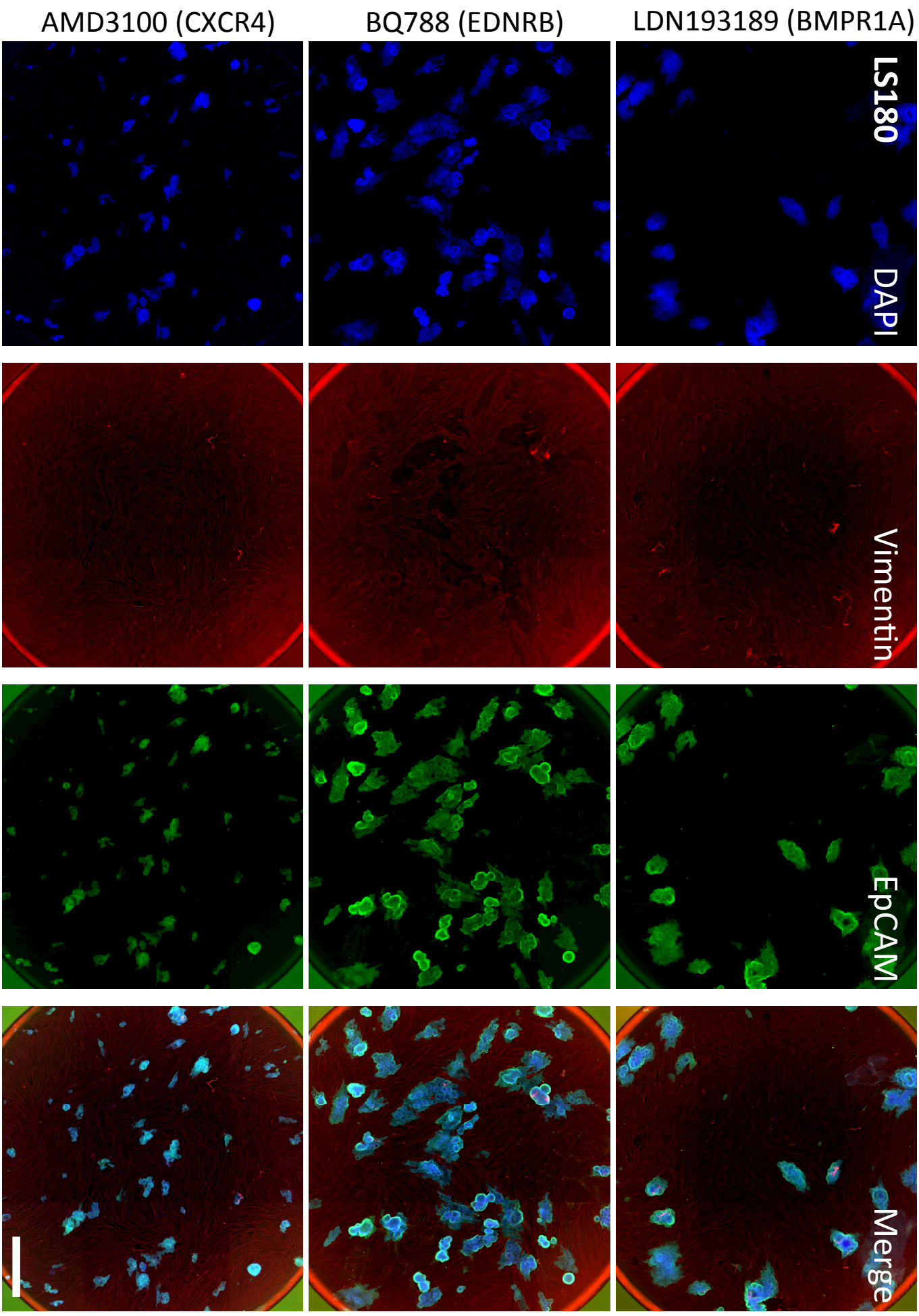
Table 3.3: Drug panel used to test the involvement of various signaling pathways in the effects of myofibroblasts on colorectal cancer cell lines in Matrigel. All drugs were used in the layer coculture assay at the given concentration in serum-free DMEM media.

Figure 3.7 (following ten pages): Responses of all five tested colorectal cancer cell lines, grown in layer coculture with myofibroblasts, to all five drugs in **Table 3.3**. Drug targets are shown in parentheses for reference. Note especially the antiproliferative effects of LDN193189 and AMD3100, and the morphological changes in HT29 in response to PHA665752. Scale bar: 1 mm

LS174T
DAPI
Vimentin
EpCAM
Merge
SF MF coculture
Figitumumab (IGF1R)
PHA665752 (c-Met)

LS174T
DAPI
Vimentin
EpCAM
Merge
LDN193189 (BMPR1A)
BQ788 (EDNRB)
AMD3100 (CXCR4)

LS180
DAPI
Vimentin
EpCAM
Merge
SF MF coculture
Figitumumab (IGF1R)
PHA665752 (c-Met)


AMD3100 (CXCR4)
BQ788 (EDNRB)
LDN193189 (BMPR1A)
LS180
DAPI
Vimentin
EpCAM
Merge

HT29
DAPI
Vimentin
EpCAM
Merge
PHA665752 (c-Met)
Figitumumab (IGF1R)
SF MF coculture

HT29
DAPI
Vimentin
EpCAM
Merge
LDN193189 (BMPR1A)
BQ788 (EDNRB)
AMD3100 (CXCR4)

HCT116
DAPI
Vimentin
EpCAM
Merge
SF MF coculture
Figitumumab (IGF1R)
PHA665752 (c-Met)

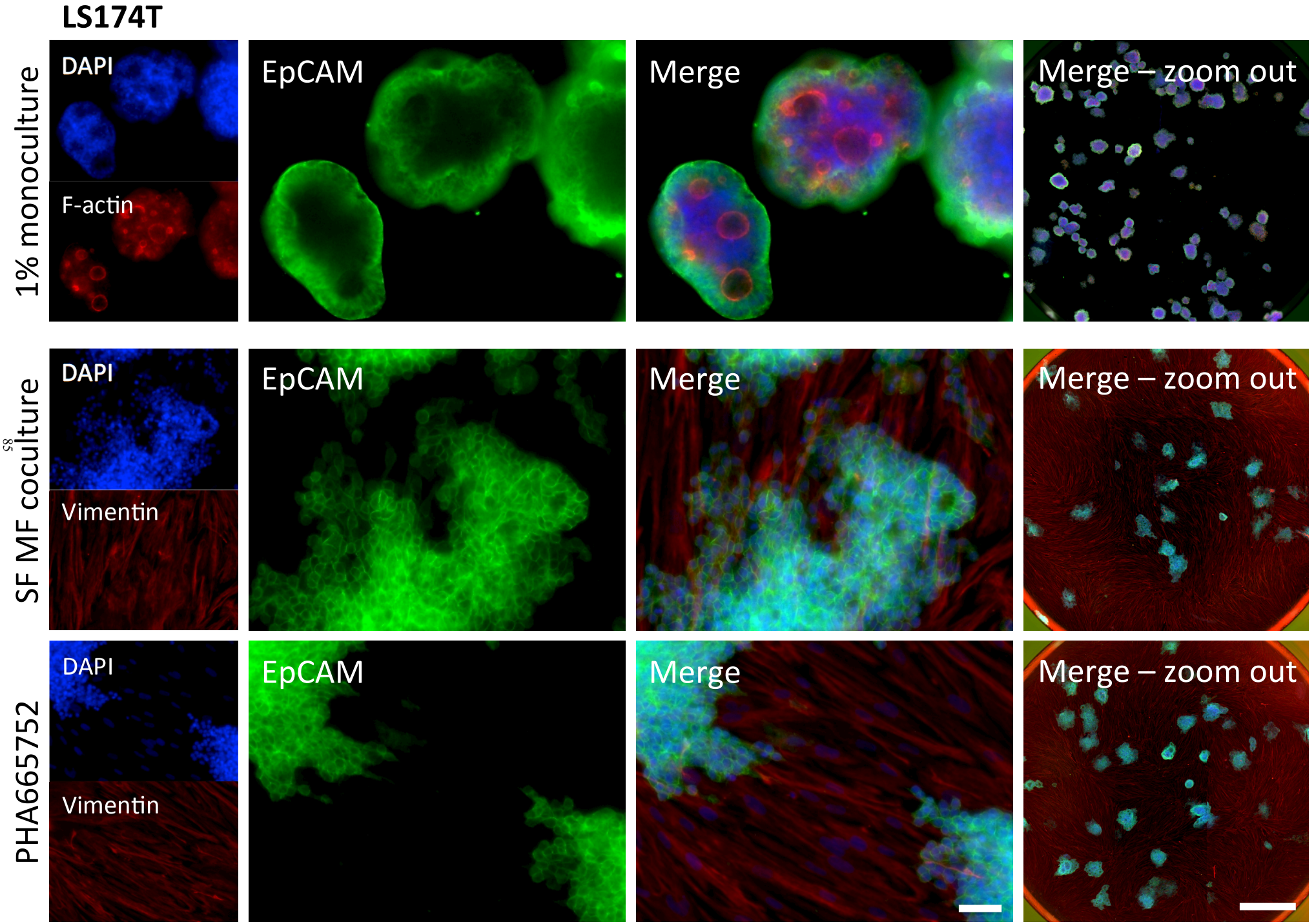
HCT116
DAPI
Vimentin
EpCAM
Merge
LDN193189 (BMPR1A)
BQ788 (EDNRB)
AMD3100 (CXCR4)

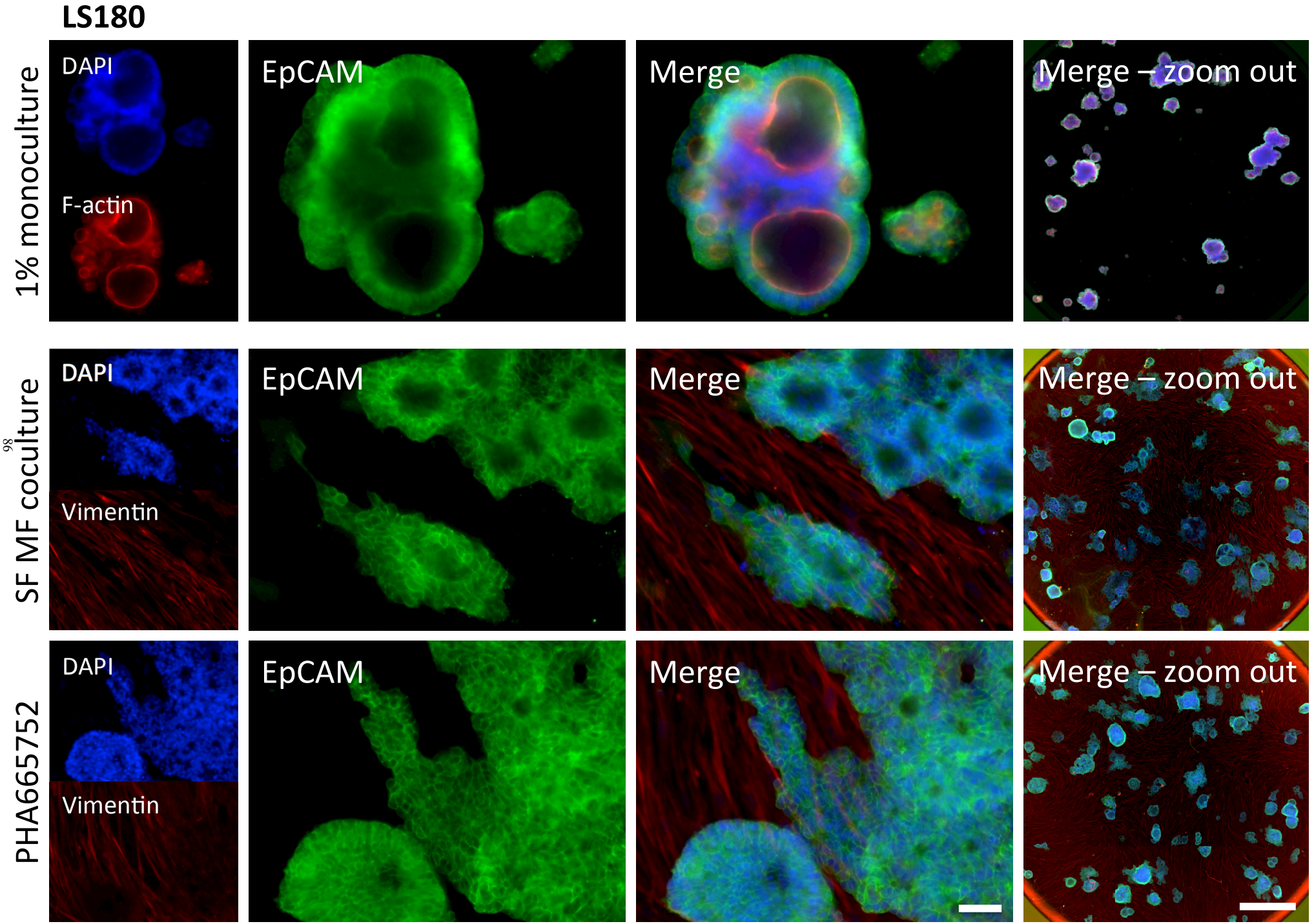
SW1222
DAPI
Vimentin
EpCAM
Merge
SF MF coculture
Figitumumab (IGF1R)
PHA665752 (c-Met)

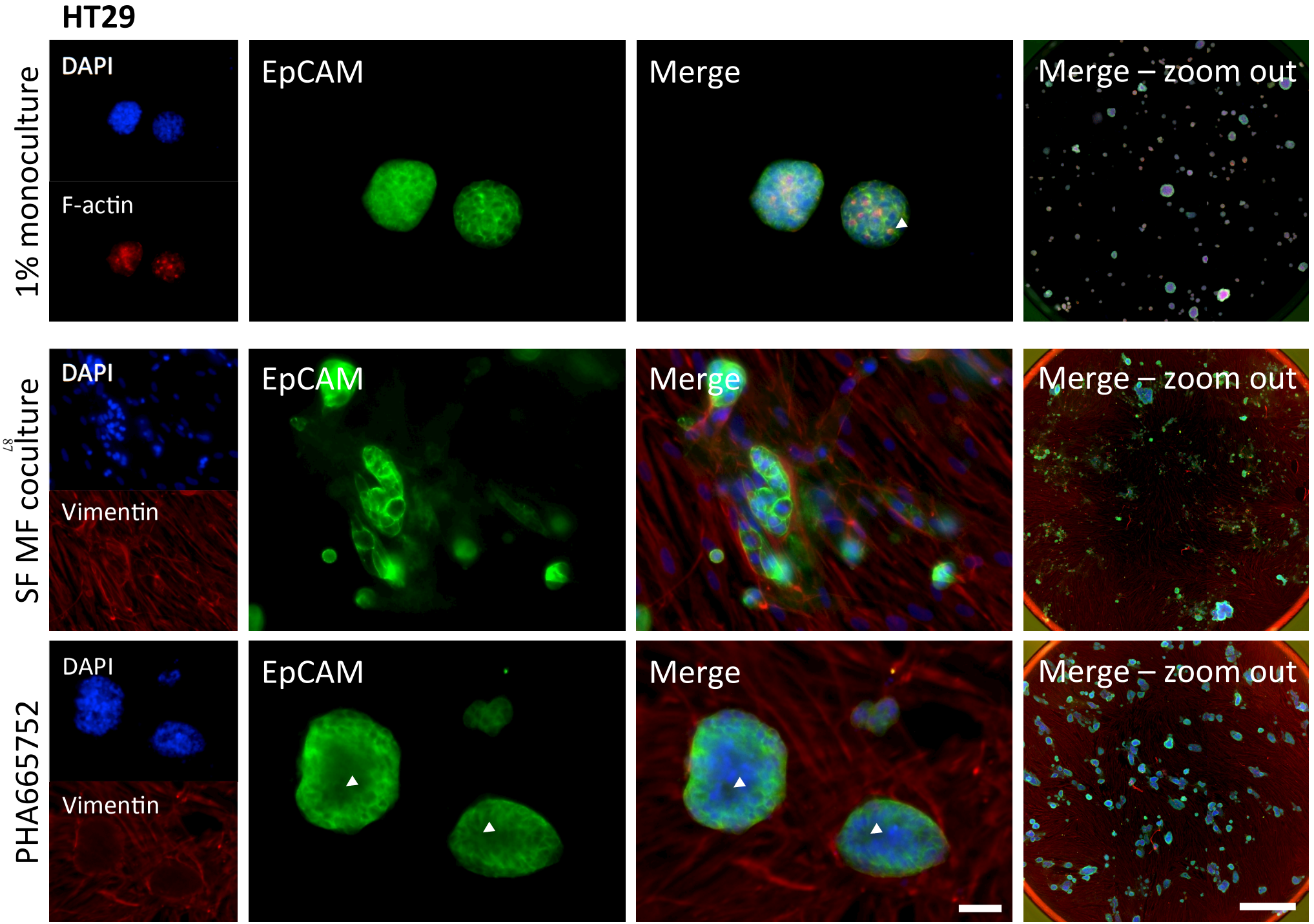
SW1222
DAPI
Vimentin
EpCAM
Merge
LDN193189 (BMPR1A)
BQ788 (EDNRB)
AMD3100 (CXCR4)

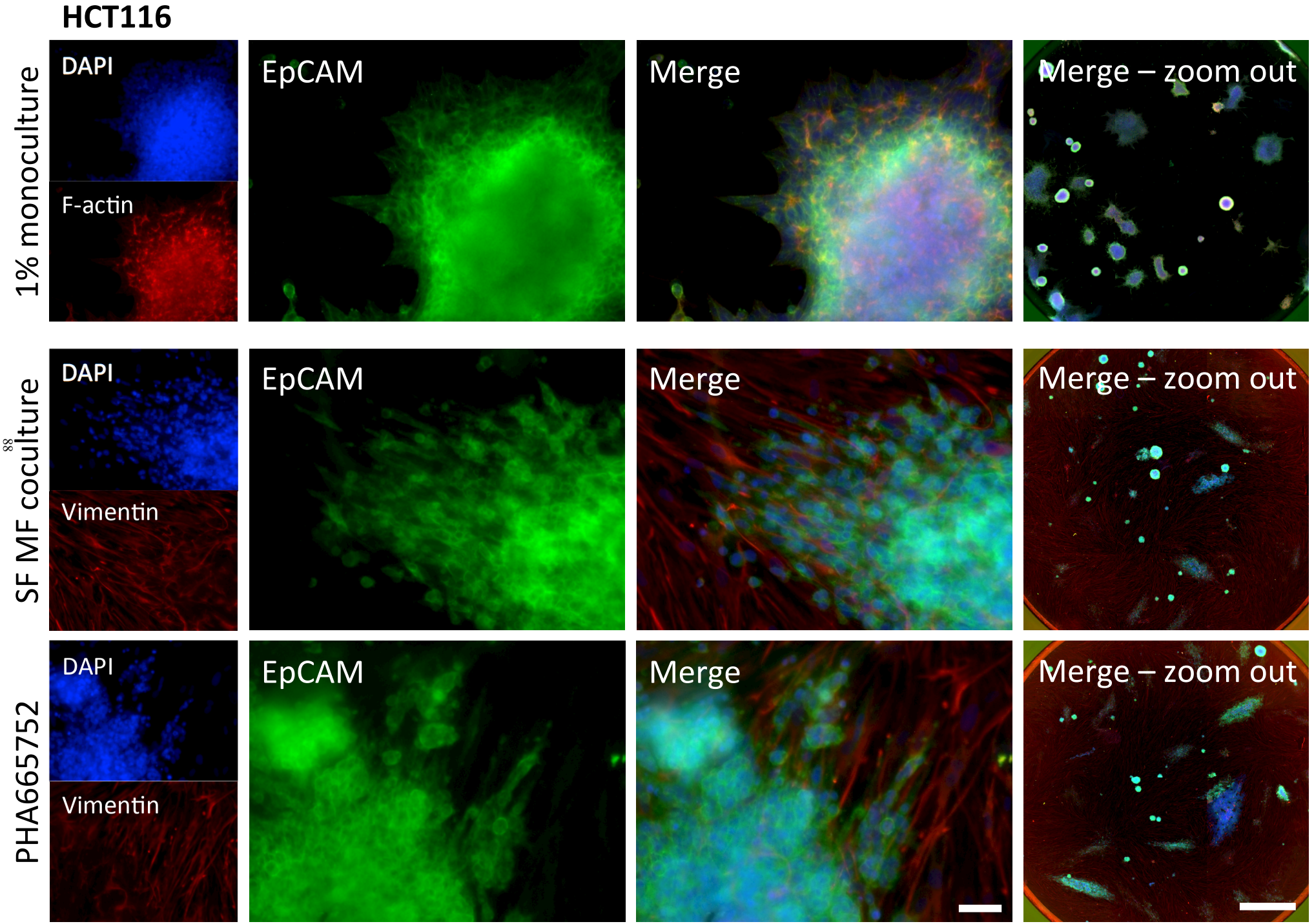
Both LDN193189 and AMD3100 (antagonists for BMPR1A and CXCR4, respectively) exhibited noticeable negative effects on colorectal cancer cell proliferation across all cell lines, possibly reflecting an importance of these receptors to cancer cell viability but equally explainable by nonspecific cytotoxicity. IGF1R blocking antibody Figitumumab had no visible effects on any of the cell lines, and while EDNRB antagonist BQ788 seemed to result in somewhat more rounded colony morphology for LS174T, this phenotype was not clearly interpretable. Most striking was the effect of c-Met antagonist PHA665752 – while no noticeable effect was seen for four of the five cell lines, HT29 exhibited a clear shift from a scattered morphology to tight, localized colonies with minimal dispersal and branching. **Figure 3.8** compares the morphologies of the colorectal cancer cell lines grown in coculture with and without the presence of PHA665752, as well as in 1% serum monoculture. As indicated by white arrowheads, HT29 cocultures treated with the drug seem to exhibit lumen-like structures akin to those visible in the monocultures, although F-actin staining was not performed on cocultures, and it is thus impossible to conclude this unequivocally. Evidently, while most cell lines exhibit no clear responses to the drug, HT29 colonies in coculture form tight structures akin to those in monoculture upon treatment.

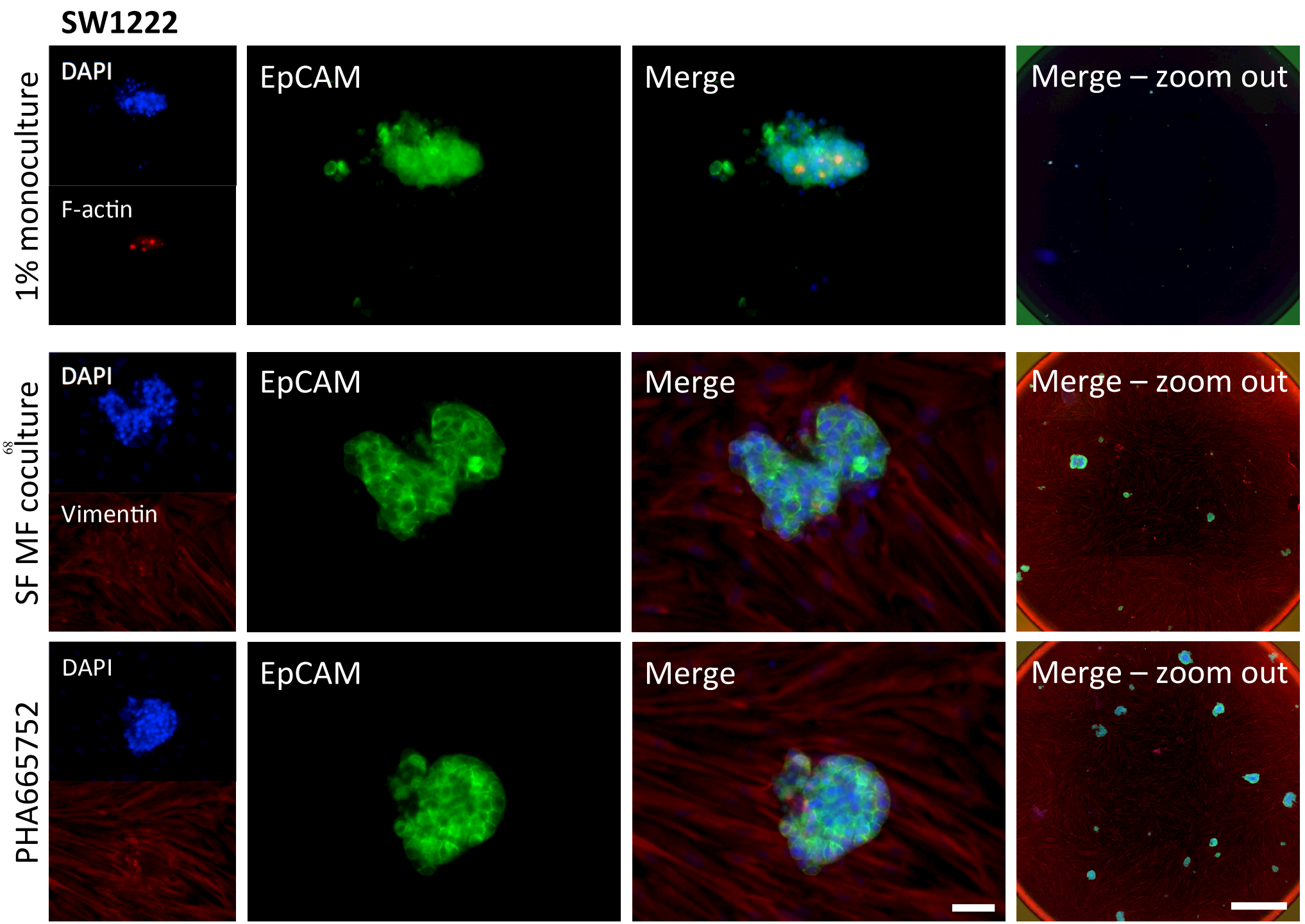
Figure 3.8 (following five pages): Zoomed images of five colorectal cancer cell lines in layer coculture with myofibroblasts treated with PHA665752, compared to untreated coculture and monoculture in 1% serum. White arrowheads indicate lumen-like cavities in cocultured HT29 cells in the presence of the drug, akin to those marked by F-actin in monocultures. Corresponding zoomed out images from **Figure 3.7** shown again for reference. Scale bar: 50 μm (zoomed in), 1 mm (zoomed out)


LS174T
1% monoculture
SF MF coculture
PHA665752
DAPI
F-actin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out


LS180
1% monoculture
SF MF coculture
PHA665752
DAPI
F-actin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out


HT29
1% monoculture
SF MF coculture
PHA665752
DAPI
F-actin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out


HCT116
1% monoculture
SF MF coculture
PHA665752
DAPI
F-actin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out


SW1222
1% monoculture
SF MF coculture
PHA665752
DAPI
F-actin
EpCAM
Merge
Merge – zoom out
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out
DAPI
EpCAM
Merge
Merge – zoom out

A second verification was also performed using siRNA knockdown. In this case, knockdown of *HGF*, *LRRC17*, and *EDNRB* was first tested on myofibroblasts in monoculture using qRT-PCR; *NKX2-3* knockdown was confirmed for at least 48 hours in previous work in the lab (Hsia & Bodmer 2015). Strong knockdown of *LRRC17* and *HGF* (≥85%) was confirmed for up to 10-13 days following transfection, while *EDNRB* expression returned to baseline after about one week (**Figure 3.9**). Next, a modified layer coculture assay was performed on the same five cell lines used in the drug screen; in this case, myofibroblasts were first seeded into 96-well plates at intermediate density, transfected with siRNAs the next day, and subsequently allowed to recover in 10% serum-containing medium for three days prior to the addition of colorectal cancer cells in Matrigel and coculture for the usual 18 days in serum-free medium.

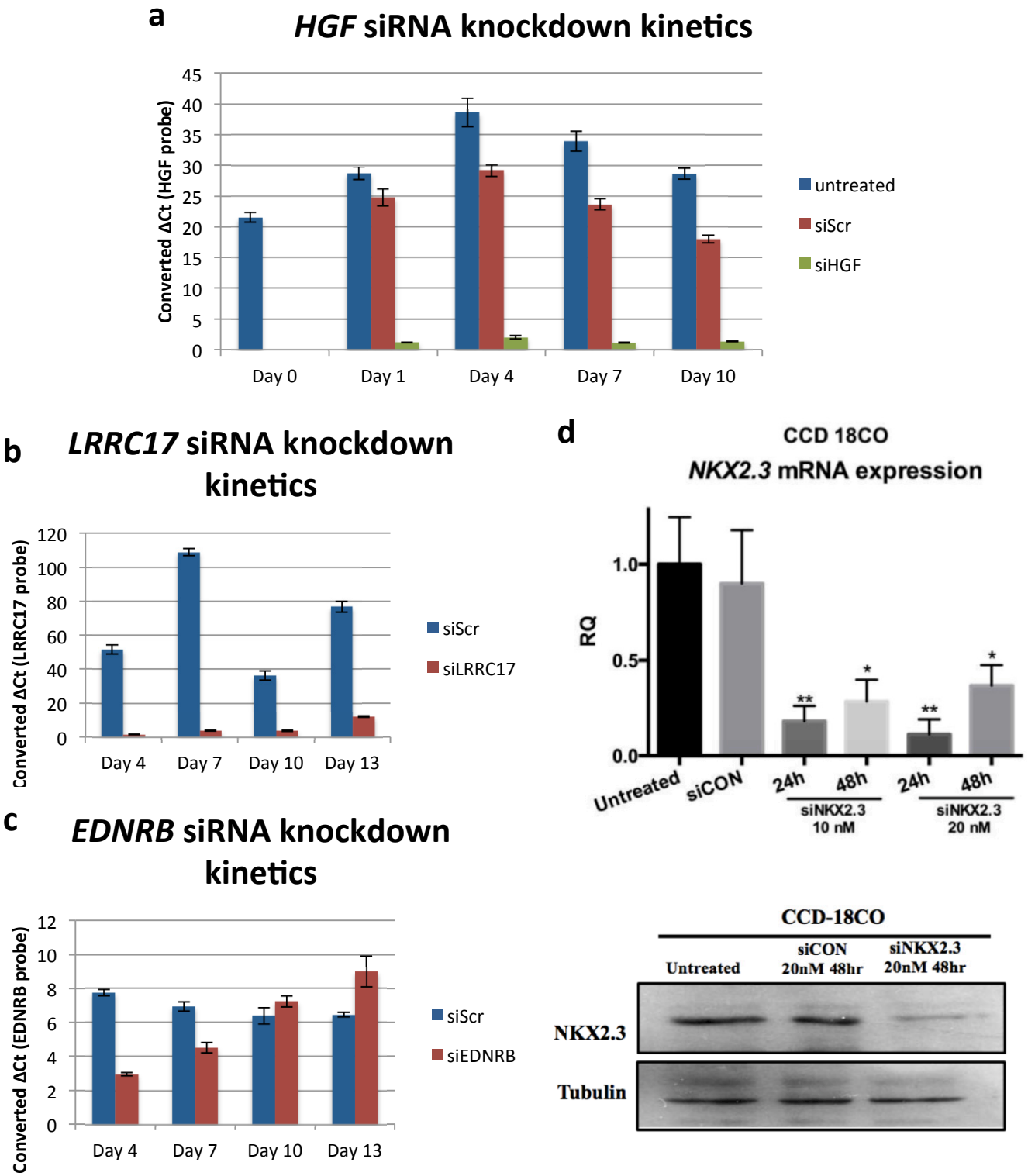


Figure 3.9: Confirmation of siRNA knockdown of *HGF*, *LRRC17*, *EDNRB*, and *NKX2-3* using RT-qPCR and Western Blot. Note that compared to scrambled siRNA control, both *HGF* (**a**) and *LRRC17* (**b**) remain significantly knocked down up to 10-13 days post-transfection, while *EDNRB* (**c**) recovers to control levels by day 10. *NKX2-3* results (**d**) taken from previous lab publication (Hsia & Bodmer 2015).

While most siRNAs exhibited no particularly striking effects on any of the cell lines, knockdown of *HGF* in myofibroblasts exactly mimicked PHA665752 treatment of HT29 cells in coculture (**Figure 3.10**). As shown in greater detail in **Figure 3.11**, HT29 in coculture with myofibroblasts formed small, dispersed colonies in control conditions, while giving rise to tight, possibly lumen-forming structures when grown with myofibroblasts transfected with siRNA targeting *HGF*. White arrowheads again indicate lumen-like cavities, akin to those in monoculture and PHA665752-coculture colonies in **Figure 3.8**. This result clearly confirms that while some colorectal cancer cell lines form well-differentiated lumen-forming colonies in mono or coculture regardless of myofibroblast coculture, other lines like HT29 respond strongly to myofibroblast-derived HGF. Specifically, it seems that HGF inhibits differentiation and lumen formation in such cells, as blocking HGF signaling either by chemical antagonist or by siRNA knockdown of the ligand results in the formation of compact, likely lumen-forming colonies akin to those seen in serum-stimulated monoculture.

Figure 3.10 (following ten pages): Effects of siRNA knockdown in myofibroblasts on five colorectal cancer cell lines grown in layer coculture. Note the resemblance of HT29 colonies grown with myofibroblasts transfected with *HGF* siRNA to PHA665752-treated HT29 cocultures in Figure 3.7. Scale bar: 1 mm

LS174T
DAPI
Vimentin
EpCAM
Merge
SF MF coculture
siScr
siLRRC17

Merge
EpCAM
Vimentin
DAPI
LS174T
siEDNRB
siHGF
siNKX2-3

siLRRC17
siScr
SF MF coculture
LS180
DAPI
Vimentin
EpCAM
Merge

siNKX2-3
siHGF
siEDNRB
LS180
DAPI
Vimentin
EpCAM
Merge

HT29
DAPI
Vimentin
EpCAM
Merge
SF MF coculture
siScr
siLRRC17

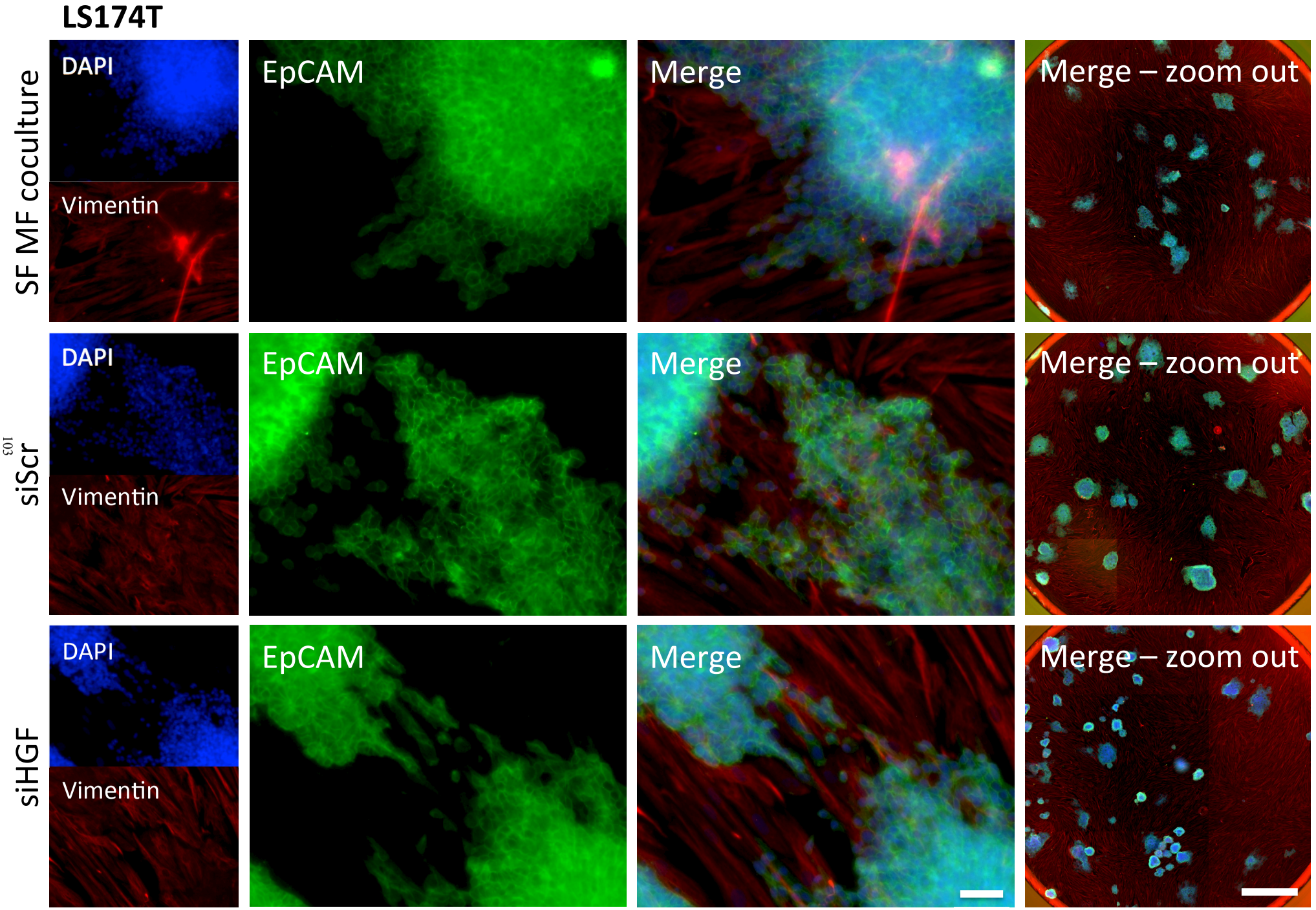
HT29
DAPI
Vimentin
EpCAM
Merge
siEDNRB
siHGF
siNKX2-3

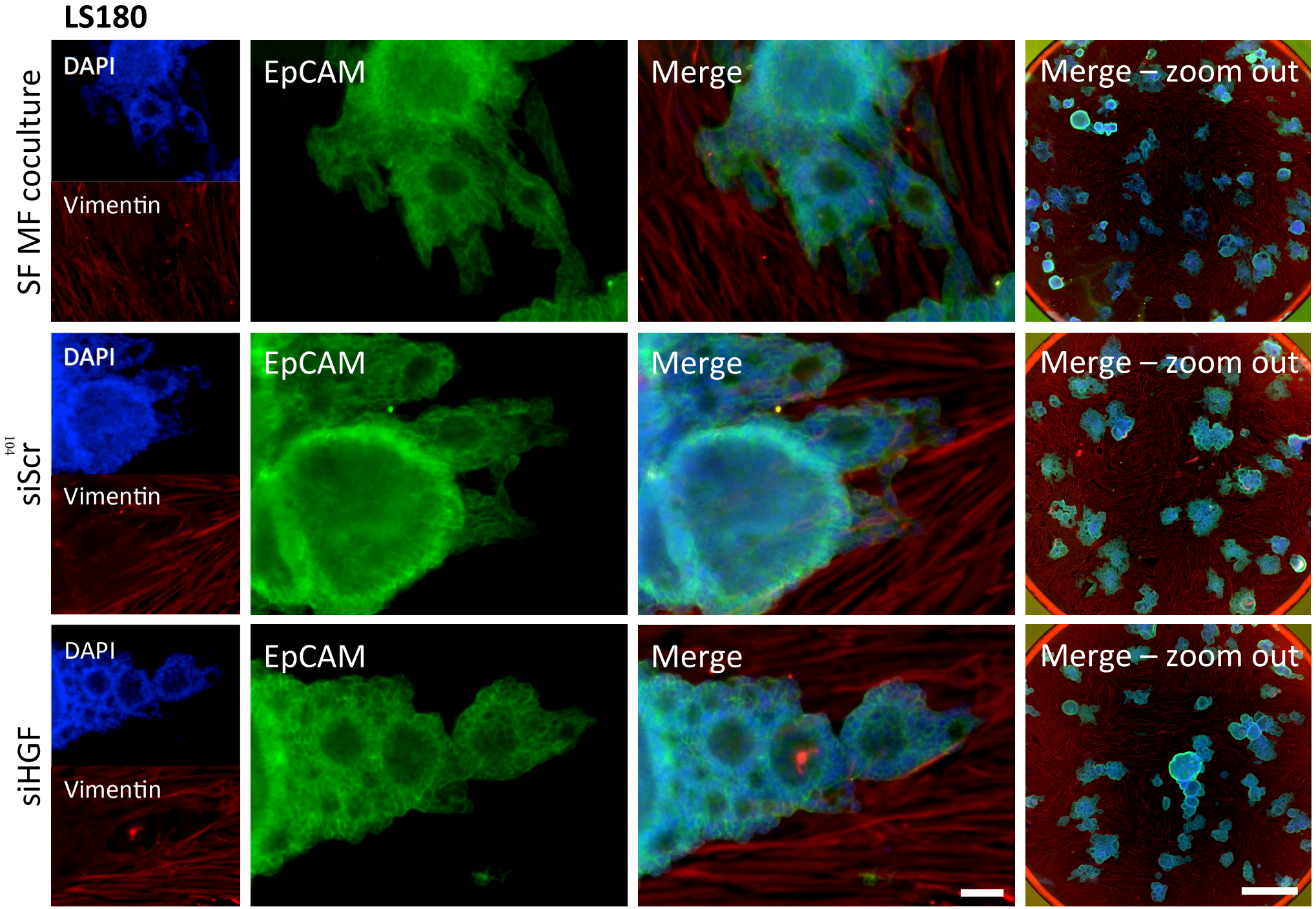
HCT116
DAPI
Vimentin
EpCAM
Merge
SF MF coculture
siScr
siLRRC17

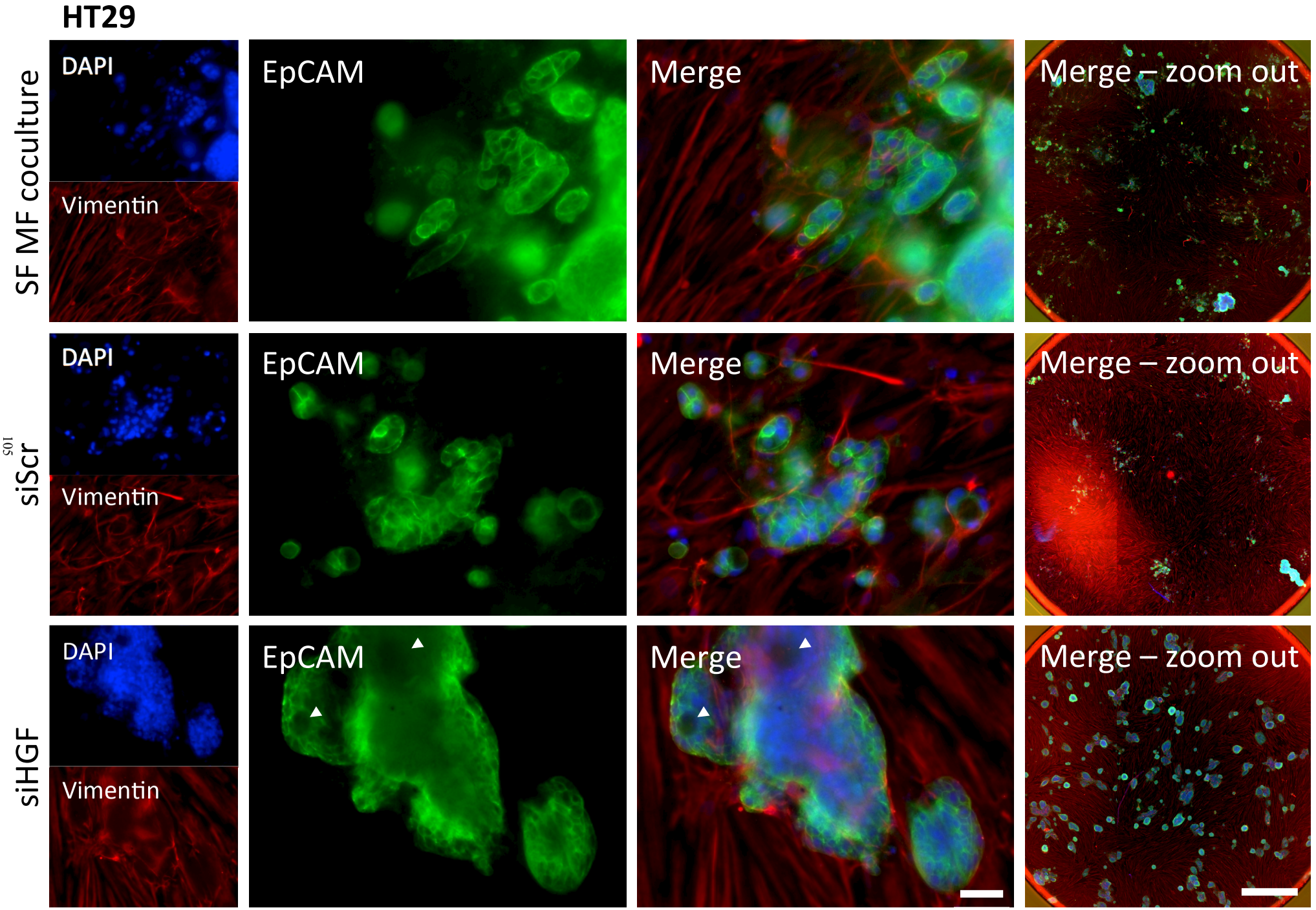
siEDNRB
siHGF
siNKX2-3
HCT116
DAPI
Vimentin
EpCAM
Merge

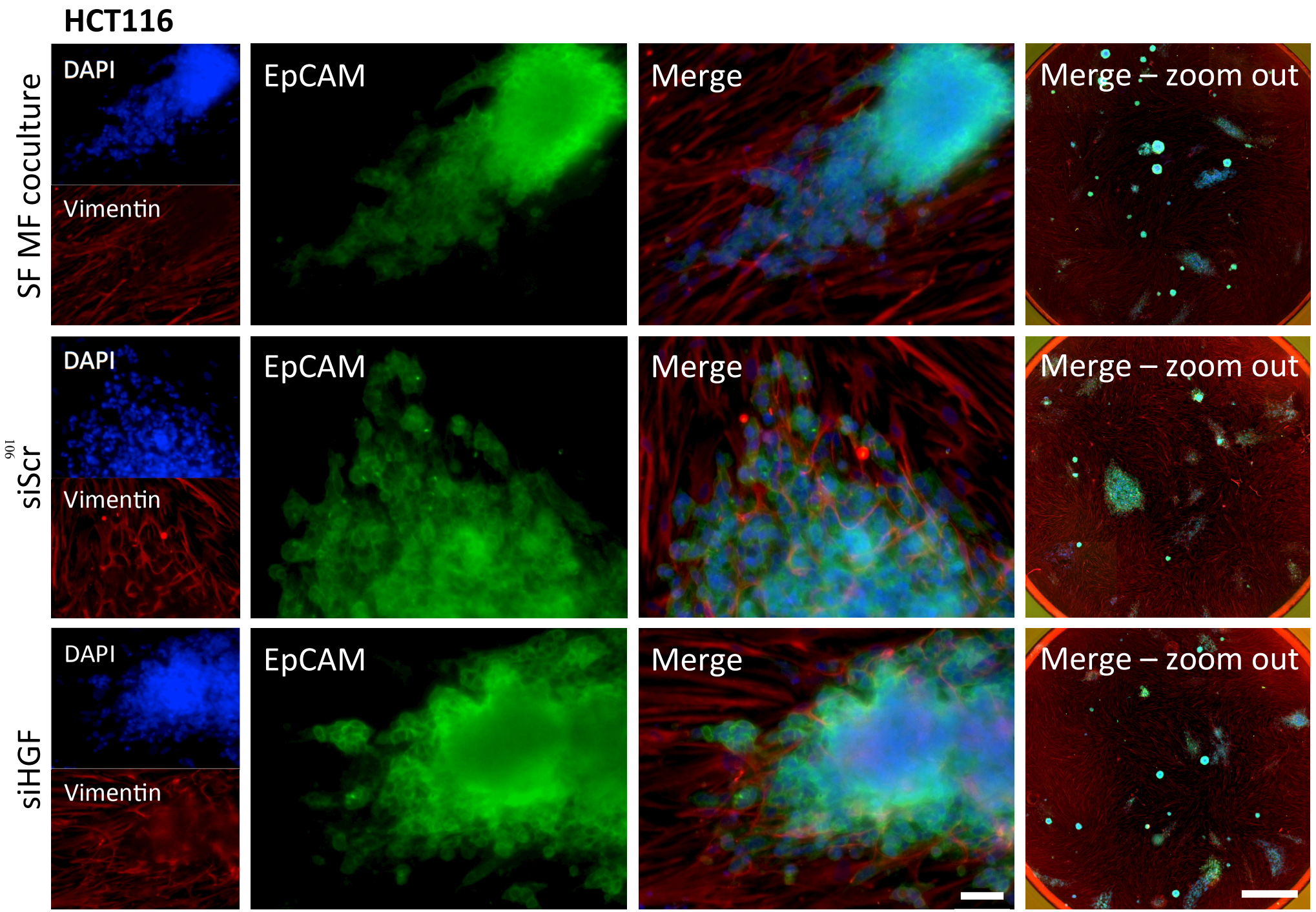
SW1222
DAPI
Vimentin
EpCAM
Merge
SF MF coculture
siScr
siLRRC17

SW1222
DAPI
Vimentin
EpCAM
Merge
siEDNRB
siHGF
siNKX2-3


LS174T
SF MF coculture
siScr
siHGF
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out


LS180
SF MF coculture
siScr
siHGF
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out


HT29
SF MF coculture
siScr
siHGF
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out


HCT116
SF MF coculture
siScr
siHGF
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out

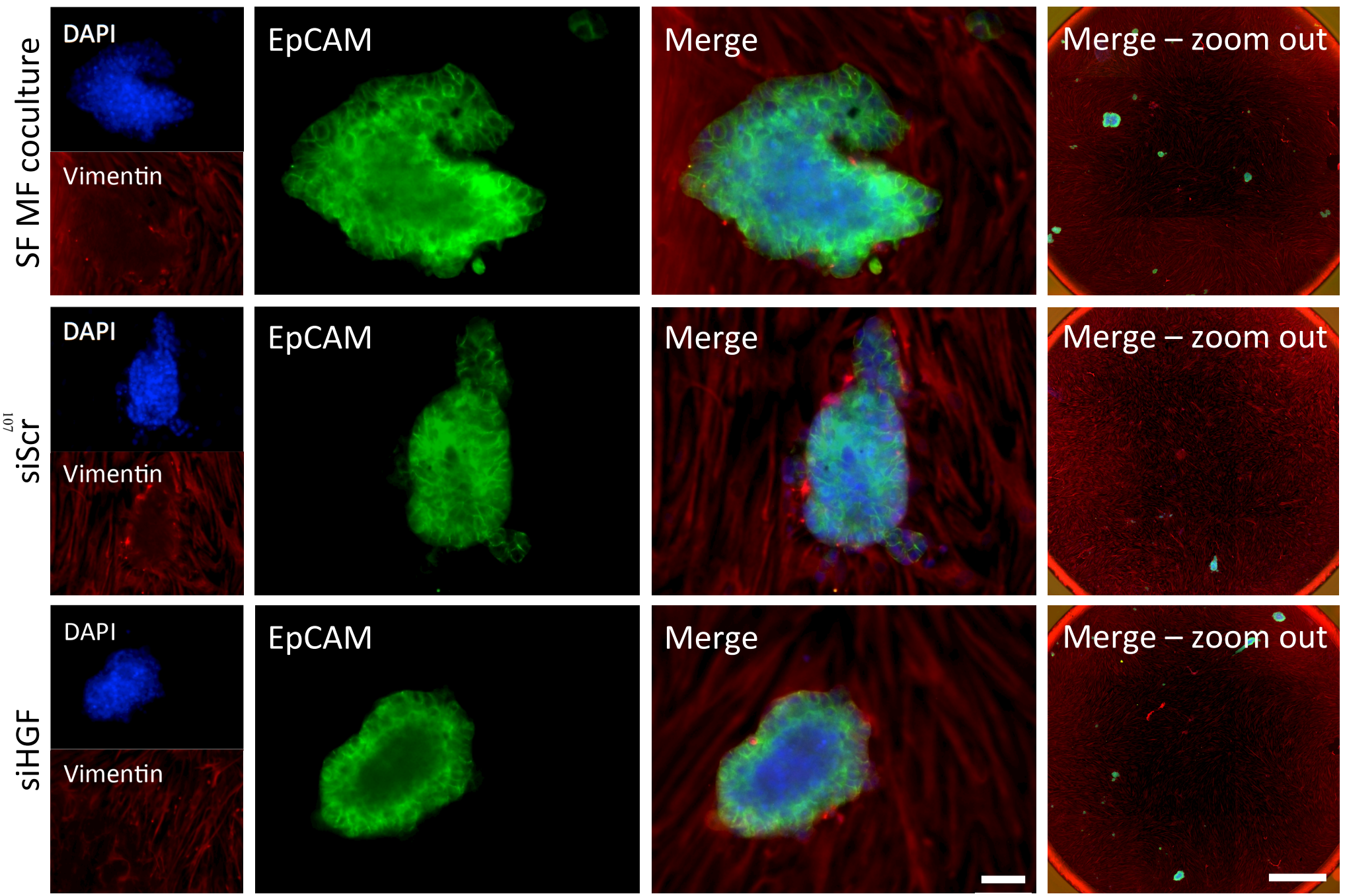

SW1222
SF MF coculture
siScr
siHGF
DAPI
Vimentin
EpCAM
Merge
Merge – zoom out

Figure 3.11 (previous five pages): Zoomed images of five colorectal cancer cell lines in layer coculture with myofibroblasts transfected with scrambled control siRNA or *HGF* siRNA, compared to untreated coculture. White arrowheads indicate lumen-like cavities in HT29 colonies grown atop myofibroblasts with *HGF* knockdown (compare to monoculture and PHA665752 coculture in **Figure 3.7**).
Scale bar: 50 μm (zoomed in), 1 mm (zoomed out)

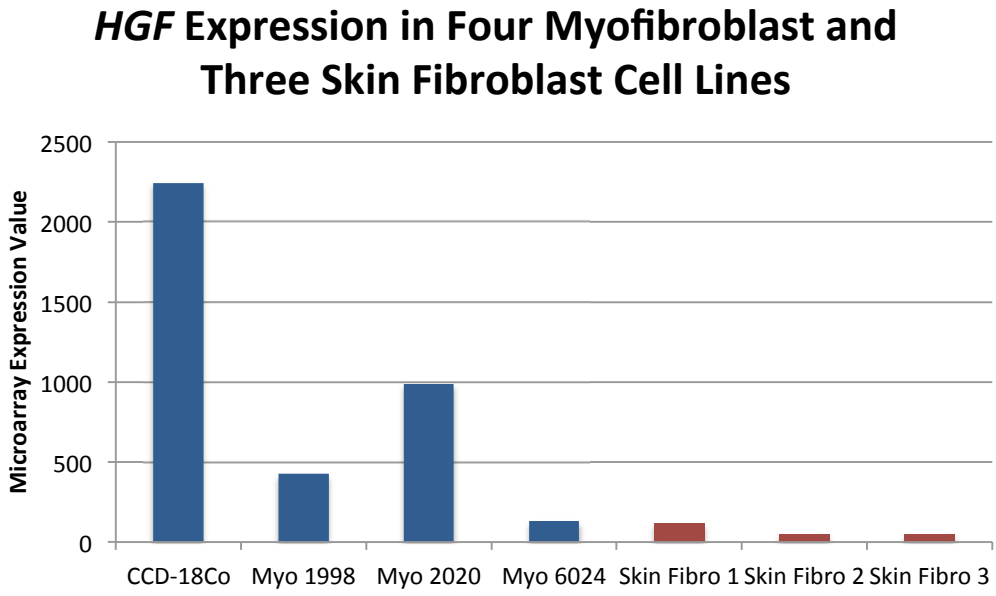


Figure 3.12: *HGF* expression across four colonic myofibroblast (blue) and three skin fibroblast (red) cell lines, from an in-house microarray gene expression dataset. Note that skin fibroblasts are consistently low for *HGF*, while myofibroblasts exhibit heterogeneous *HGF* expression, possibly reflecting their activation state.

3.3 Discussion

This chapter began by showing that colonic myofibroblasts substantially promote the survival and growth of colorectal cancer cells, an effect mediated at least in part by secreted ligands and somewhat replicated by the addition of serum. In line with these observations, chemical inhibition of two receptor proteins sensitive to mesenchymal signals reported to affect colorectal cancer proliferation, BMPR1A and CXCR4, led to a substantial decrease of cancer cell growth and survival in coculture (**Figure 3.7**). While these effects may be due to nonspecific toxicity, and must be confirmed with dose response experiments as well as genetic knockouts of the ligands in question, they are promising. BMP2, one of the ligands of BMPR1A, shows substantially increased expression in myofibroblasts compared to skin fibroblasts and may therefore be involved in the much improved ability of conditioned medium from the former to support cancer cell growth compared to the latter. Conflicting reports exist regarding the influence of BMP2 on colorectal cancer cells, with some indicating that the ligand inhibits proliferation while others suggest that it promotes survival in serum-free conditions; the findings of this chapter support the latter interpretation (Vishnubalaji et al. 2016; Kim et al. 2015). CXCL12, the ligand of CXCR4, has also been implicated in promoting colorectal cancer cell survival and growth, and tumors overexpressing CXCR4 have worse prognosis and higher incidence of metastasis (Song et al. 2014).

Another effect of myofibroblasts on colorectal cancer cells, most evident for the HT29 cell line, involves the control of differentiation. HGF, originally called scatter factor, was first identified as a fibroblast-secreted factor capable of inducing

dissociation and migration of a range of epithelial cell types, as well as promoting anchorage-independent growth (Stoker et al. 1987; Uehara & Kitamura 1992; Jiang et al. 1993). Further studies additionally implicated HGF in angiogenesis, as it induced growth and motility in endothelial cells (Bussolino et al. 1992; Grant et al. 1993). HGF has also been found to influence differentiation in various tissues, acting as a pro-differentiation factor in some contexts and inhibiting differentiation in others (Leshem et al. 2000; Shen et al. 1997; Karp et al. 1994). In colorectal cancers, HGF has long been implicated as a mitogen capable of protecting tumors from cytotoxic therapies, and has also been found to correlate with increased invasiveness and metastasis formation (Otte et al. 2000; Liska et al. 2011; Liu et al. 2012; Kermorgant et al. 2001). In fact, recent studies have found that the HGF receptor c-Met is upregulated at the invasive front of colorectal carcinomas, in structures known as tumor buds (Bradley et al. 2016; Satoh et al. 2014). Tumor budding is proposed to result as a consequence of partial EMT; WNT signaling has also been implicated in this process, as nuclear β-catenin is known to accumulate in cells at the invasive front of colorectal tumors (Mitrovic et al. 2012; Grigore et al. 2016; El-Gendi & Al-Gendi 2011). Several studies have shown that HGF stimulation can lead to nuclear accumulation of β-catenin in liver, breast, and colon cancer cells, and in the latter case it has been suggested that this type of WNT activation defines colorectal cancer stem cells (Huang et al. 2012; Monga et al. 2002; Vermeulen et al. 2010). Thus, there is some controversy over whether HGF signaling stimulates EMT in some cell types, or whether it actually inhibits differentiation and maintains the cells in a mesenchymal-like stem cell state. Cancer-associated fibroblasts have been implicated as sources of HGF in the microenvironments of breast, liver, lung, esophageal, oral, and colorectal

cancers (Tyan et al. 2011; Jia et al. 2013; Mahale et al. 2016; Grugan et al. 2010; Costea et al. 2013; Koliaraki et al. 2012).

Previous studies have asserted a tight link between HGF signaling and stemness in colorectal cancer stem cells, even going so far as to suggest that HGF can revert differentiated cancer cells back to a stem-like phenotype (Vermeulen et al. 2010). However, as the findings in this chapter show, there seems to be a substantial heterogeneity in the responses of colorectal cancer cell lines to HGF. While HT29, known to be a partially differentiating cell line with limited lumen formation in Matrigel, showed an obvious reduction in the formation of differentiating colonies in the presence of HGF, other well-differentiating cell types like LS174T and LS180 did not respond to variations in this ligand in any apparent fashion, continuing to form differentiated lumens as before. Further study screening the effects of coculture, and the specific role of HGF, on a larger population of cell lines could help uncover any underlying characteristics that may differentiate HGF-sensitive lines from insensitive ones; one intriguing possibility is a correlation of WNT signaling alterations with HGF sensitivity. It is also worthwhile to note that removal of HGF did not appreciably inhibit proliferation in any of the cell lines (in fact, it may have been promoted in HT29), suggesting that other myofibroblast-derived mitogenic and/or survival factors are responsible for the vastly improved growth of the cancer cells in serum-free coculture compared to monoculture. Additionally, as shown in **Figure 3.12**, the CCD-18Co colonic myofibroblast cell line used in this chapter expresses high levels of *HGF*, while other colonic myofibroblast lines exhibit mixed levels of this mRNA and all skin fibroblasts in the in-house microarray gene expression dataset are essentially negative for it. This suggests that HGF secretion is variable across

colonic myofibroblasts, and may be induced when they are activated by certain cancer cell-derived ligands, a further level of complexity to this interaction. While a role for HGF, specifically derived from local myofibroblasts, is evident in colorectal cancer, the data at hand suggest that its influence may be more nuanced than previously thought.

Chapter Four

Myofibroblast Recruitment by Colorectal Cancer Cells

4.1 Introduction

In order to perform their roles as Cancer-associated Fibroblasts and influence tumor cell behaviors as described in the previous chapter, myofibroblasts must be in an appropriate physical proximity to the recipients of their signaling. The cellular composition and arrangement of the tumor microenvironment depends on both the innate organization of the local stroma as well as the subsequent alterations induced by the incipient tumor. Both primary and metastatic cancers are known to enact significant changes in their surroundings through the expression of membrane-bound and secreted signaling molecules, and it is suspected that similar mechanisms may be responsible for the recruitment of myofibroblasts toward and into tumors.

Myofibroblasts are known to exhibit chemotactic responses in various contexts and to various ligands. The primary factor reported to induce myofibroblast activation is Transforming Growth Factor-β1 (TGF-β1) (Desmoulière et al. 1993). TGF-β1 is strongly expressed in a variety of carcinomas, especially at later stages, and can induce α-smooth muscle actin expression as well as contractility in fibroblasts (Derynck et al. 2001; Meyer-ter-Vehn et al. 2006). On the other hand, recent work from our lab showed that colorectal myofibroblasts represent a cell type distinct from activated skin fibroblasts, and differ substantially in their response to TGF-β1, along with a range of other characteristics (Hsia et al. 2016). Members of the platelet-derived growth factor (PDGF) family have also been implicated in fibroblast and myofibroblast activation, and signaling via this pathway has been found to promote tumor growth, invasion and metastasis in a number of cancers (Heldin 2013). Local myofibroblast recruitment is a well-documented phenomenon in wound healing

(Desmoulière & Gabbiani 1988). In addition to TGF-β1 and PDGF, secreted ligands such as tenascin-C and IL-1β, as well as extracellular matrix components such as fibronectin, have been suggested as possible effectors of this process (Ko et al. 2008; Tamaoki et al. 2005; Kawamoto et al. 1997). In-depth *in vitro* studies of myofibroblast activation and recruitment in colorectal cancer have been hampered by a limited availability of assays that recapitulate the critical aspects of the tumor microenvironment in a robust and accessible fashion, which has further complicated the question.

In this chapter, an *in vitro* assay modeling the physical arrangement of colorectal cancer cells and stromal myofibroblasts is applied to test myofibroblast recruitment in a panel of colorectal cancer cell lines. Gene expression correlates of the recruitment phenotype are determined and discussed, and a panel of drugs is tested on a subset of cell lines with the goal of blocking myofibroblast recruitment.

4.2 Results

4.2.1 Myofibroblast Recruitment Assay Design and Optimization

4.2.1.1 Assay Design

Colorectal tumors and pericryptal myofibroblasts coexist in a compact volume containing cells and extracellular matrix. It was reasoned that an appropriate *in vitro* model of the interactions between these cells must recreate this biophysical arrangement, and a Matrigel-based assay was devised as follows (**Figure 4.1**):

1. Cultured colorectal cancer cells were placed onto a plastic surface in a drop of 50% Matrigel
2. Once the gel hardened, it was covered with cell culture growth medium containing dissociated cultured CCD-18Co myofibroblasts, which subsequently attached to the substrate to form a flat population on top and around the drop
3. Movement of the myofibroblasts toward the cancer cells was assayed by immunofluorescence staining of fixed samples after several days of culture

Staining was performed using antibodies for vimentin, which labeled the myofibroblasts (along with some vimentin-expressing cancer cell lines), and for laminin, which delineated the extent of the Matrigel drop. Resulting images were scored on a five-point scale based on the degree of myofibroblast invasion into the drop.

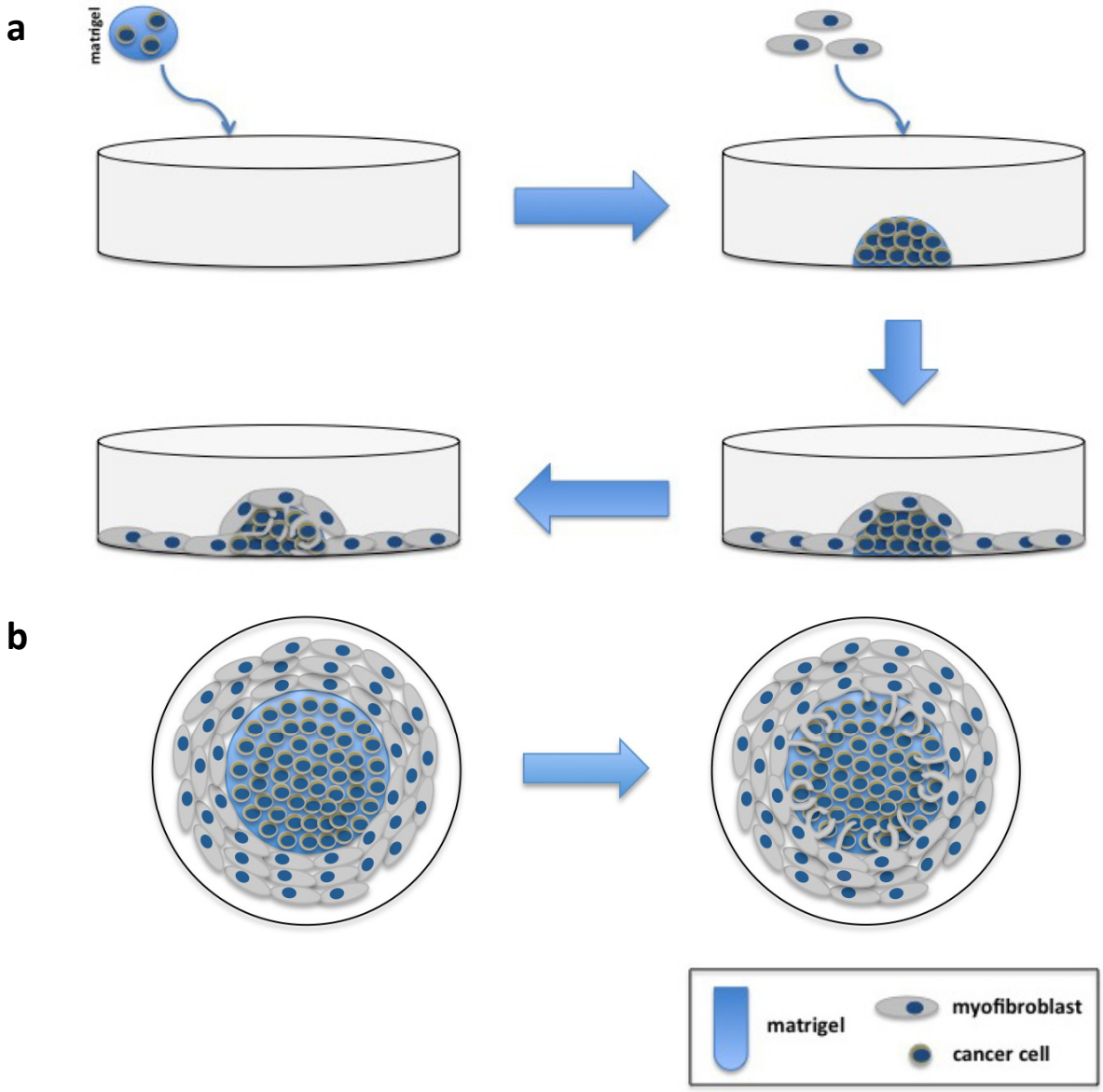


Figure 4.1: Outline of the myofibroblast recruitment assay. (**a**) A drop of cancer cells in 50% Matrigel is first placed into the culture well. Next, once the drop has hardened, myofibroblasts are added on top of the drop in culture medium. Finally, the coculture is allowed to grow for several days, over which time the myofibroblasts have a chance to invade into the drop of cancer cells. (**b**) Top view of the culture well. Once the myofibroblasts have invaded into the Matrigel as shown on the right, staining and imaging are performed to give the experiment readout.

4.2.1.2 Assay Optimization

Preliminary experiments were performed to determine assay conditions that optimally balanced cell viability and proliferation, allowing sufficient culture time for myofibroblast recruitment phenotypes to be exhibited while limiting cancer cell overgrowth. Fetal bovine serum (FBS) is the most common growth factor supplement in standard cell culture applications, and its concentration in culture media is a primary stimulant of cell survival and growth. As such, seeded cell numbers and FBS concentrations were the variables on which optimization was performed. Specifically, drops containing either 5,000 or 10,000 cancer cells were placed into wells of 96-well culture plates, and were subsequently covered either in 2,000 myofibroblasts in 1% serum containing DMEM culture medium or 5,000 myofibroblasts in serum-free medium. The cultures were grown for nine days, with media changed every three days, to allow maximal time for myofibroblast proliferation and invasion while ensuring that the cancer cells did not overgrow and obscure the process. This test was performed on a limited panel of ten colorectal cancer cell lines, two of which are shown in **Figure 4.2**; seeded cancer cell numbers had no appreciable effect on the cultures, so only the wells with 10,000 cancer cells are shown.

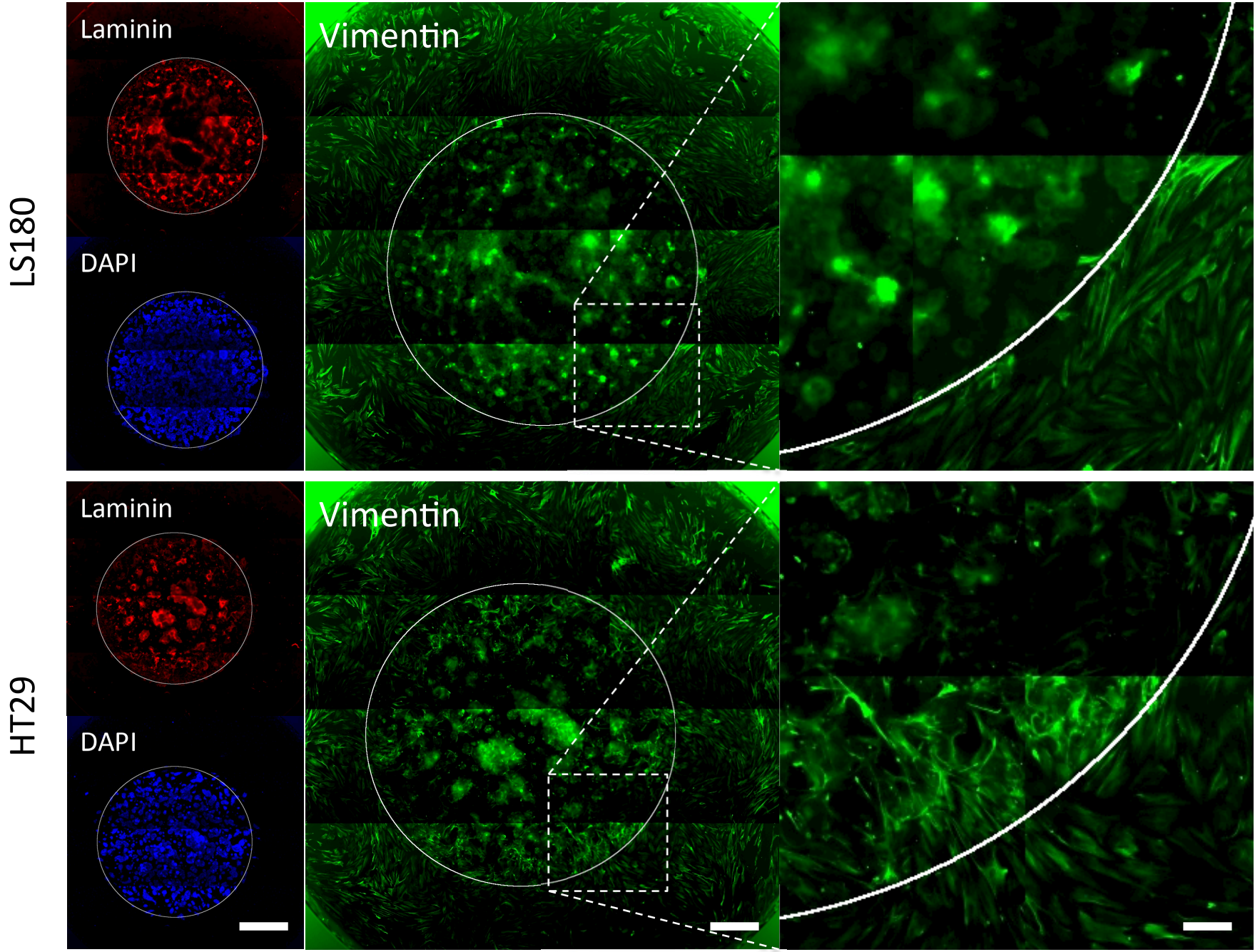
Stronger vimentin staining and somewhat denser growth suggests that myofibroblasts (MFs) were healthier and proliferated to a greater extent in 1% serum-containing media, especially when taking into account the initial cell numbers. Cancer cells also proliferated more in 1% serum, although appreciable proliferation was seen in serum-free conditions as well. Importantly, cancer cells in drops proliferated better in the presence of myofibroblasts than on their own under serum-free conditions (data not

shown), consistent with findings in Chapter 3. Myofibroblasts exhibited a slight level of invasion into empty drops, especially in 1% serum; HT29-containing drops exhibited myofibroblast invasion under both serum-free and 1% serum conditions. Notably, myofibroblasts were only found on top of drops containing HT29, but not empty drops or those containing LS180, and a large fraction of invading myofibroblasts were confined to the bottom surface of the well.

Optimal parameters were determined to be 10,000 colorectal cancer cells per drop grown with 2,000 myofibroblasts for nine days in 1% serum-containing media. This set of conditions was used for subsequent experiments.

Figure 4.2 (following three pages): Myofibroblast recruitment assay – optimization. Images for two cancer cell lines of the panel of ten tested in this experiment are shown. 10,000 cancer cells were seeded in the drop, followed by the addition of 5,000 myofibroblasts (MFs) in serum-free medium or 2,000 MFs in 1% serum-containing medium. Co-staining for laminin and vimentin was performed, and the laminin pattern was used to define the boundary of the Matrigel drop (white outline). Scale bar: 1 mm (DAPI and Laminin), 500 μm (Vimentin big), 125 μm (Vimentin zoom)

Empty drop control
Laminin
DAPI
Vimentin
Serum-free (5000 MFs)
1% serum (2000 MFs)


Serum-free (5000 MFs)
LS180
HT29
Laminin
DAPI
Vimentin
Laminin
DAPI
Vimentin

1% serum (2000 MFs)
Vimentin
Laminin
DAPI
LS180
HT29

4.2.2 Myofibroblast Recruitment Assay Across Cell Lines

The myofibroblast recruitment assay was subsequently applied to a panel of 43 colorectal cancer cell lines selected to exhibit a wide spectrum of phenotypes in terms of *in vitro* lumen formation/differentiation capacity, as described in Chapter 3, and mismatch repair proficiency (**Table 4.1**). Recruitment scoring was performed by three assessors on a de-identified set of images, and consensus scores were determined by averaging all assessors' scores and rounding to the nearest integer. Cell lines marked unscorable by at least one assessor, usually due to cancer cell overproliferation outside the drop, were subjectively reassessed and either given the average score of the remaining assessors or marked unscorable by consensus. Images from a representative set of seven cell lines are shown in **Figure 4.3**, and the complete experiment is summarized in **Table 4.2**.

The cancer cell lines were scored on a five-point scale based on the extent of myofibroblast invasion that they induced. A score of 3 was given to cell lines that showed no appreciable difference in their myofibroblast invasion pattern compared to the empty drop control, a score of 1 corresponded to a complete absence of myofibroblast invasion, and a score of 5 reported myofibroblast invasion much higher than the control. Scores of 2 and 4 were given to intermediate levels. **Figure 4.3** shows the range of invasion patterns seen across the cell lines, including examples of cell lines with each of the five scores. Additionally, a subset of cancer cell lines expressed vimentin, meaning that they were stained along with the myofibroblasts. As shown in **Figure 4.3**, the invasion scores for these cell lines were determined in part by identifying the varying subcellular vimentin distributions between cancer cells and

myofibroblasts. As in the case of HDC8, some drops of cancer cells were washed off during the staining process; no appreciable effect was seen on the myofibroblasts invading along the bottom of the well. The cancer cell drops exhibit a range of diameters; this was dependent on the surface characteristics of individual culture plates, as every drop in every well in a single plate exhibited roughly the same diameter.

Colony morphology: / **Mismatch repair status:**	**Lumen-forming**	**Intermediate morphology**	**Not lumen-forming**	**Unknown**
Mismatch repair proficient	C10 C32 CACO2 CCO7 HT55 RW7213 SW1222	CX1 COLO206 HDC8 HT29 NCIH716 SKCO1 SW1417 VACO4A	CAR1 HDC57 SW1116 SW480 SW620 T84 VACO429	CC20 COLO205 COLO320HSR OXCO1
Mismatch repair deficient	CCK81 LS174T LS180 OXCO2 SNUC2B	GP2D HDC9 LOVO SW48	DLD1 HCA7 HCT116 LS411	RKO
Unknown		OUMS23	PMFK014	CL11

Table 4.1: Cell lines used in the myofibroblast recruitment assay, shown based on colony morphology in 3D culture and mismatch repair pathway functionality (Bracht et al. 2010; Ashley et al. 2013).

Figure 4.3 (following four pages): Myofibroblast recruitment assay – cancer cell line screen. Images for seven cancer cell lines of the panel of 43 (**Table 4.1**) tested in this experiment are shown. A spectrum of myofibroblast invasion phenotypes is evident across the cell lines. As before, a white line delineates the boundary of the cancer cell drop. Scale bar: 1 mm (DAPI and Laminin), 500 μm (Vimentin big), 125 μm (Vimentin zoom)

LS180 – score 1

Empty drop control

HDC8 – score 2

COLO205 – score 3

SW1116 – score 5

HT29 – score 4

Vimentin-positive Examples
Vimentin
Laminin
DAPI
SW620 – score 1
OXCO1 – score 4

Cell Line	Assessor 1	Assessor 2	Assessor 3	Consensus
C10	4	4	4	4
C32	4	4	4	4
CACO2	-	1	1	-
CAR1	5	5	5	5
CC20	-	1	-	-
CCK81	1	1	1	1
CCO7	5	5	5	5
CL11	3	4	4	4
COLO205	3	3	4	3
COLO206	5	5	5	5
COLO320HSR	3	3	-	3
CX1	5	5	5	5
DLD1	-	1	1	-
GP2D	1	1	2	1
HCA7	-	1	1	-
HCT116	-	1	1	-
HDC57	2	3	4	3
HDC8	2	2	2	2
HDC9	5	5	5	5
HT29	5	3	4	4
HT55	1	2	2	2
LOVO	1	3	-	2
LS174T	1	1	1	1
LS180	1	1	2	1
LS411	4	2	3	3
NCIH716	3	4	4	4
OUMS23	-	1	1	-
OXCO1	5	4	4	4
OXCO2	5	4	5	5
PMFK014	1	1	2	1
RKO	1	1	1	1
RW7213	5	5	5	5
SKCO1	3	3	4	3
SNUC2B	3	5	4	4
SW1116	5	5	5	5
SW1222	1	2	2	2
SW1417	5	5	5	5
SW48	5	5	4	5
SW480	1	2	2	2
SW620	1	1	-	1
T84	1	4	2	2
VACO429	5	5	4	5
VACO4A	1	2	2	2

Table 4.2 (previous page): Summary of the myofibroblast recruitment screen across the cancer cell line panel. Scores from each assessor, as well as the consensus scores, are shown. A (-) denotes that the cell line was deemed unscorable by the given assessor and/or by consensus.

Next, microarray gene expression comparisons were performed between the resulting myofibroblast recruiter and non-recruiter cancer cell line categories. Two such analyses, a strict one (comparing only cell lines scored 1 to cell lines scored 5) and a broad one (comparing cell lines scored 1 or 2 to cell lines scored 4 or 5), specifically filtered to include only secreted and membrane protein-coding genes as in Chapter 3, are shown in **Figure 4.4**. Several informative associations were determined from the volcano plots. First off, a number of differentiation markers show noticeable upregulation in the non-recruiter cell lines, albeit with weak p-values. Specifically, *CEACAM5*, *GPA33*, *HEPH*, and *CFTR* are well-known markers of differentiated colon epithelial cells, while *LGR5* is reported to mark stem cells in the small intestine (Zheng et al. 2011; Wong et al. 2006; Brookes et al. 2006; Than et al. 2016; Barker et al. 2007). Additionally, *BMP7* is thought to be secreted by differentiated intestinal epithelium and alter the behaviors of pericryptal myofibroblasts below (Kosinski et al. 2007). Meanwhile, *DKK1*, a transcriptional target of epithelial-mesenchymal transition (EMT)-inducing transcription factor ZEB1, is noticeably upregulated in recruiter cell lines (de Barrios et al. 2017). Lysozyme (*LYZ*), a Paneth cell marker normally restricted to small intestine but known to appear in colons damaged by diet or inflammatory bowel disease, is also present in recruiters over non-recruiters (Wang et al. 2011; Fahlgren et al. 2003; Klockars et al. 1977). Other correlates of the recruitment phenotype with less clear relevance include membrane-bound signaling ligand ephrin-A5 (*EFNA5*), reportedly downregulated in colorectal cancers but with no evidence of a functional role, as well as actin regulator gelsolin (*GSN*) and *GPR37*,

which encodes a receptor for the neurotrophin prosaposin (Surawska et al. 2004; Sun et al. 1999; Meyer et al. 2013). Overall, it seems that non-recruiter cell lines generally exhibit markers of conventional enterocyte differentiation, while recruiter cell lines may be enriched with cell lines that form Paneth cells or have undergone EMT.

Even more intriguing, however, is that the recruiter cell lines show increased expression of several members of signaling pathways known to be involved in myofibroblast-colorectal cancer interactions (**Figure 4.4**, orange boxes). For instance, as mentioned in the previous chapter, the CXCL12-CXCR4 signaling axis reportedly stimulates cancer cell migration and invasion, while IGF ligands, of which IGFBP7 is an inhibitor, are thought to promote colorectal cancer cell survival and proliferation (Brand et al. 2005; Zanella et al. 2015). EGFR is a well-known growth factor receptor known to modulate colorectal cancer proliferation and targeted by successful clinical treatments (Krasinskas & M. 2011). Possibly most interestingly, myofibroblast-recruiting cell lines express increased amounts of endothelin-1 (*EDN1*) and platelet-derived growth factor-C (*PDGFC*). EDN1 is a ligand of endothelin receptor B (EDNRB), which is one of the strongest markers of colonic myofibroblasts over skin fibroblasts as described in Chapter 3 (Hsia et al. 2016). Previous reports suggest that EDNRB is a primary mediator of colonic myofibroblast migration in response to EDN1 (Kernochan et al. 2002). PDGF-C, on the other hand, is a member of the PDGF signaling family widely reported to participate in myofibroblast activation and the induction of migration, and has previously been implicated in metastatic niche formation (Kastanis et al. 2011; Bandapalli et al. 2012; Yamauchi et al. 2014). These genetic correlates were taken as candidate signals possibly responsible for the invasion phenotype and analyzed further.

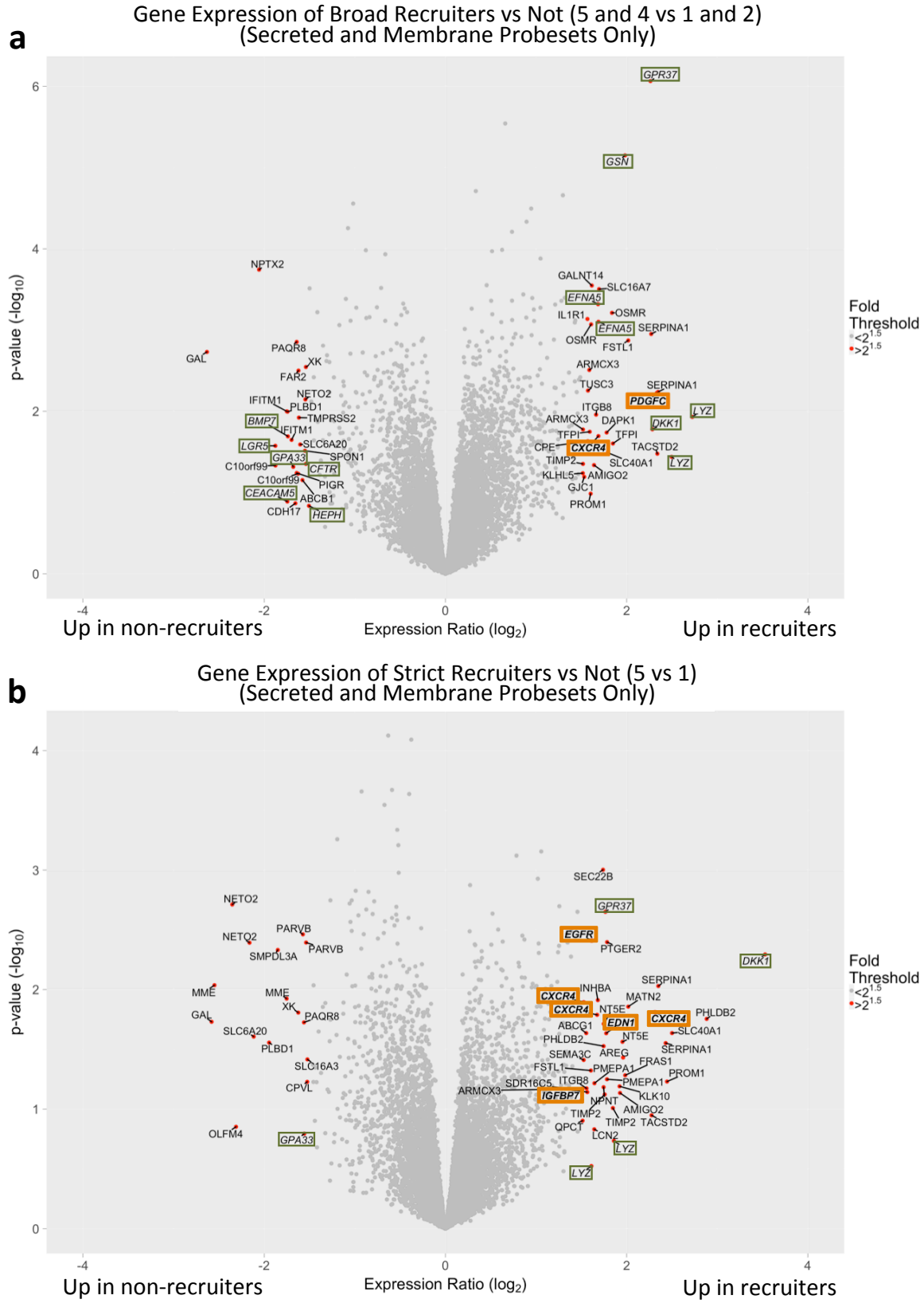


Figure 4.4: Gene expression comparison between cell lines showing myofibroblast invasion (recruiters) and not (non-recruiters) in the myofibroblast recruitment screen. (**a**) Broad comparison of cell lines scored 4 or 5 against those scored 1 or 2. (**b**) Strict comparison of lines scored 5 against those scored 1. Probesets with a fold difference greater than $2^{1.5}\approx2.83$ are labeled. Genes with strong literature support for involvement in myofibroblast activation pathways are boxed in orange. Other genes of interest, either strongly correlated with MF recruitment or known from literature to be involved in cell-cell signaling or colorectal differentiation, are boxed in green. P-values from Student's t-test uncorrected for multiple comparisons are shown.
Top: n=18 (scores 5 and 4) vs n=14 (scores 1 and 2); bottom: n=11 (score 5) vs n=7 (score 1)

4.2.3 Myofibroblast Recruitment Drug Screen

Following the determination of the candidate genes in **Figure 4.4**, a drug screen was devised to test the effects of blocking a panel of signaling pathways on the myofibroblast recruitment phenotype. As listed in **Table 4.3**, in addition to the five drugs used in the previous chapter (**Table 3.3**), several others were added, targeting the EGFR and FGFR pathways, previously reported to influence colorectal cancer cell proliferation, migration and invasion, as well as the TGF-β and PDGF pathways reportedly responsible for myofibroblast activation, and intracellular signaling components PI3K and AKT (Krasinskas & M. 2011; Matsuda et al. 2012; Singer & Clark 1999). The standard myofibroblast recruitment assay was performed for eight cell lines scored as strong recruiters in the cell line screen (CAR1, CCO7, CX1, HDC9, OXCO1, SW1116, SW1417, SW48), along with negative control line LS180 and empty drop control. Culture media containing the various drugs were added to the cells on the day after cultures were seeded, and subsequently replaced with fresh drug-containing media on days 5 and 8 prior to sample fixation on day 9. Results of the screen are summarized in **Figure 4.5**.

Drug	Target	Concentration
A8301	TGFBR1	1 μM
AMD3100	CXCR4	1 μM
BQ788	EDNRB	1 μM
Cetuximab	EGFR	10 μg/mL
Figitumumab (CP751851)	IGF1R	10 μg/mL
GSK690693	AKT	1 μM
Imatinib	PDGFR	1 μM
LDN193189	BMPR1A/ACVR1	500 nM
LY294002	PI3K	1 μM
PD173074	FGFR1/2/3	100 nM
PHA665752	c-Met	100 nM

Table 4.3: Drug panel used to test the involvement of various signaling pathways in determining the myofibroblast recruiter phenotype of cancer cell lines. All drugs were used in the myofibroblast recruitment assay at the given concentration in 1% serum DMEM.

Figure 4.5 (following twelve pages): Myofibroblast recruitment assay – drug screen. Images for five of the nine screened cell lines, along with empty drop control, are shown. Every screened cell line had a visible downregulation of myofibroblast recruitment in response to imatinib treatment, compared to BQ788 or 1% serum control. Drug targets are shown in parentheses for reference. Once again, a white line delineates the extent of the cancer cell drop. Scale bar: 1 mm (DAPI and Laminin), 500 μm (Vimentin big), 125 μm (Vimentin zoom)

Empty drop control
Laminin
DAPI
Vimentin
1% serum control
BQ788 (EDNRB antagonist)

LS180
Laminin
DAPI
Vimentin
1% serum control
BQ788 (EDNRB antagonist)

CAR1
Laminin
DAPI
Vimentin
1% serum control
BQ788 (EDNRB antagonist)

CCO7
Laminin
DAPI
Vimentin
1% serum control
BQ788 (EDNRB antagonist)

SW1417
Laminin
DAPI
Vimentin
1% serum control
BQ788 (EDNRB antagonist)

SW48
Laminin
DAPI
Vimentin
1% serum control
BQ788 (EDNRB antagonist)

Empty drop control
1% serum control
Imatinib (PDGFR antagonist)
Laminin
DAPI
Vimentin

LS180
Laminin
DAPI
Vimentin
1% serum control
Laminin
DAPI
Vimentin
Imatinib (PDGFR antagonist)

CAR1
Laminin
DAPI
Vimentin
1% serum control
Imatinib (PDGFR antagonist)

CCO7
Laminin
DAPI
Vimentin
1% serum control
Laminin
DAPI
Vimentin
Imatinib (PDGFR antagonist)

SW1417
Laminin
DAPI
Vimentin
1% serum control
Laminin
DAPI
Vimentin
Imatinib (PDGFR antagonist)

SW48
Laminin
DAPI
Vimentin
1% serum control
Imatinib (PDGFR antagonist)

Most of the eleven drugs had no discernible effects on myofibroblast invasion or, in the case of PI3K inhibitor LY294002, resulted in widespread cell death that made it impossible to interpret the result. However, **Figure 4.5** shows a representative summary of the effects of BQ788 (EDNRB antagonist) and imatinib (PDGF receptor antagonist) on a subset of screened cell lines. While BQ788 resulted in no significant reduction in myofibroblast recruitment by any cell line, imatinib, strikingly, led to an almost complete abrogation of the phenotype in all screened cell lines. Note that in this experiment, myofibroblasts grew at a lower density than previously, and under these conditions, increased vimentin staining in the wells containing cancer cells compared to the empty drop controls suggests a growth stimulatory effect of cancer cells on myofibroblasts. However, no apparent differences are present in either the overall DAPI or vimentin staining extents when comparing paired control and imatinib-treated wells with the same cell line combinations. Blocking PDGF receptor signaling thus blocks myofibroblast recruitment by colorectal cancer cell lines without appreciably affecting cell viability.

According to the in-house microarray expression dataset, *PDGFA* and *PDGFC* are the only two PDGF ligands expressed to an appreciable level in any of the available cell lines**.** As shown in **Figure 4.6a**, although *PDGFC* was found to be a correlate of myofibroblast recruitment in the cell line screen (**Figure 4.4**), it is not expressed in all recruiter cell lines subjected to the drug screen. However, the recruiter lines that do not express *PDGFC* do express *PDGFA* – in fact, the expression of these ligands tends to be at least somewhat mutually exclusive in the cell lines with known recruitment scoring, and the expression of at least one of them significantly associates with the recruiter phenotype (**Figure 4.6b**). Importantly, CCD-18Co cells, as well as

other colonic myofibroblast cell lines established in-house, express strong levels of PDGF receptor genes *PDGFRA* and *PDGFRB* (**Figure 4.6c**). Some of the non-recruiting cell lines do express one of the ligands, indicating that an additional factor may also be needed for the phenotype to be exhibited, that these cell lines proliferate too quickly and crowd out any incoming myofibroblasts, or that these cell lines have a gene expression pattern different in the drop coculture conditions compared to those from which the microarray-bound RNA sample was taken. Some cell lines also exhibit strong recruitment phenotype with low expression of either *PDGFA* or *PDGFC*; these lines may recruit myofibroblasts using a different mechanism, or may again have a gene expression pattern different in the conditions at hand compared to those for which the expression data was obtained. These observations suggest that PDGFA and PDGFC secreted from colorectal cancer cells may be involved, possibly in a redundant fashion, in the recruitment of local myofibroblasts into the tumor body, a process with possible implications to primary tumor development as well as to the formation of a hospitable metastatic niche.

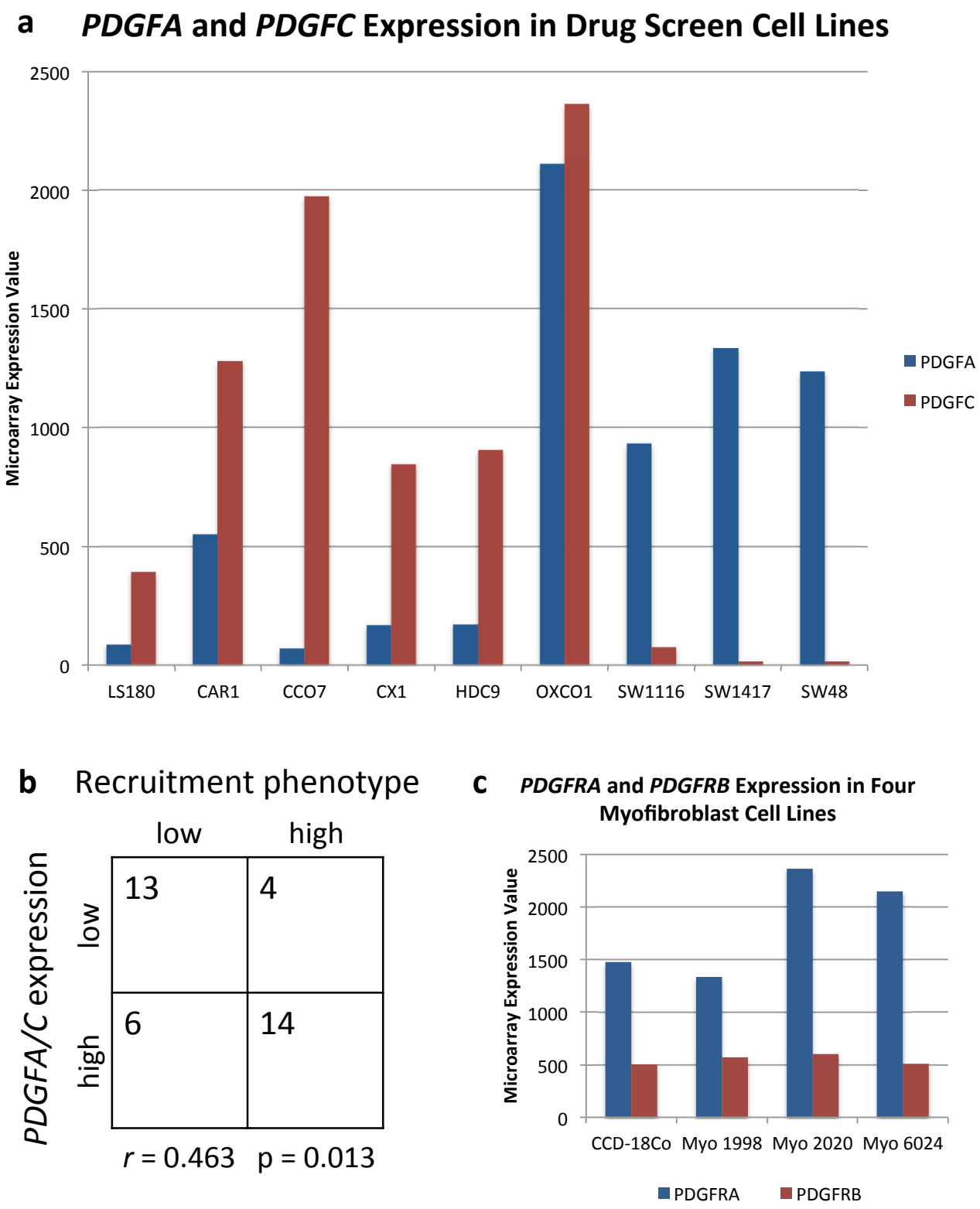


Figure 4.6: *PDGFA* and *PDGFC* microarray expression in colorectal cell lines. (**a**) *PDGFA* and *PDGFC* exhibit mostly mutually exclusive expression in the myofibroblast-recruiting cell lines used in the drug screen, suggesting that these ligands may substitute for one another. (**b**) *PDGFA* and *PDGFC* expression patterns across all 37 scored cell lines were categorized using a Gaussian Mixture Modeling-based approach (see Chapters 5 and 6). High expression of either *PDGFA* or *PDGFC* (low corresponds to cell lines with poor expression of either gene) according to this algorithm significantly ($p<0.05$) associates with the recruitment phenotype (recruitment score 1-3: low; score 4-5: high), although several outlier cell lines do remain. r is the coefficient of association for the 2x2 table; p-value calculated by χ^2 test with Yates's correction for continuity. (**c**) Expression of PDGF receptor genes *PDGFRA* and *PDGFRB* across four colonic myofibroblast cell lines, showing strong expression in all lines.

4.3 Discussion

The goal of this chapter was to develop and test an *in vitro* assay of myofibroblast migration into colonies of colorectal cancer cells, with the specific goal of mimicking the close proximity and ECM environment in which these cells interact *in vivo*. Using this assay, 43 colorectal cancer cell lines were tested for the ability to recruit cultured CCD-18Co colonic myofibroblasts. It was found that a subset of cell lines is capable of doing this, and subsequent genetic analyses and drug tests were performed to pinpoint what signals these cells uniquely produce to enable this phenotype. Although limited by the fact that only one myofibroblast cell line was used, these experiments suggest a key role of the PDGF signaling pathway in the activation and recruitment of local myofibroblasts by colorectal cancer cells.

As previously described, PDGF ligands act synergistically with TGF-β during wound healing of the skin, secreted by platelets that invade a new wound and subsequently stimulating the invasion of stromal and immune cells (Hosgood 1993). In fact, PDGF signaling is such a potent inducer of these processes that single doses of picomolar amounts of active PDGF-BB dimer have been reported to significantly accelerate wound closure and *de novo* ECM deposition, signifying an increased activation of local fibroblasts and myofibroblasts (Pierce et al. 1991). A role for PDGF in oncogenesis was first noted when a >90% homology was discovered between *PDGFB* and $p28^{v\text{-}sis}$, the transforming protein of the Simian Sarcoma Virus (Waterfield et al. 1983; Doolittle et al. 1983; Bejcek et al. 1992). Later studies found that mutations in PDGF ligands and receptors are present in a range of cancers, often resulting in increased PDGF signaling within the tumor cells themselves by

upregulation of receptor expression and increased sensitivity to environmental PDGF or by activating mutations in the receptor kinase domains (Heldin 2013). Breast and colorectal cancer cells undergoing epithelial-mesenchymal transition have been reported to engage a PDGF autocrine loop by expressing high levels of both ligand and receptor (Jechlinger et al. 2006; Steller et al. 2013). Such loops have additionally been noted in gliomas and glioblastomas, the rare skin tumor dermatofibrosarcoma protuberans and other sarcomas, hepatocellular carcinoma, and prostate cancer (Uhrbom et al. 2000; Shimizu et al. 1999; Gotzmann et al. 2006; Ustach et al. 2010). PDGF signaling has also been implicated in tumor angiogenesis, specifically by recruiting pericytes and possibly inducing their transition to fibroblast-like phenotypes, as well as by inducing erythropoietin expression in the stroma (Manzat Saplacan et al. 2017; Cao et al. 2002; Potente et al. 2011; McCarty et al. 2007; Hosaka et al. 2016; Xue et al. 2011). Specific roles for PDGF ligands in the activation and recruitment of cancer-associated fibroblasts have been described in melanoma, cervical cancer, and colorectal cancer (Anderberg et al. 2008; Pietras et al. 2008; Pena et al. 2013; Sundberg et al. 1997).

The findings in this chapter suggest that PDGF ligands secreted from colorectal cancer cells directly activate the invasion of local myofibroblasts into the tumor bulk. Although further confirmation is necessary, these data also suggest that other factors aside from PDGF secreted by the cancer cells promote myofibroblast proliferation, while PDGF itself is responsible for inducing this migratory phenotype. This is an important distinction, as the effects of growth factors on cell survival, proliferation, and migration are often listed in the same breath. None of the screened cancer cell lines express either of the PDGF receptors according to the microarray dataset (data

not shown), and this in combination with the lack of appreciable imatinib effects on their proliferation suggests that no autocrine stimulation by PDGF is present in these lines.

Although an involvement of PDGF signaling in myofibroblast recruitment is nothing new in itself, this coculture assay using a large panel of cell lines with fully characterized baseline gene expression patterns allows for a finer definition of the interaction. The cancer cell lines express only *PDGFA* and/or *PDGFC*, meaning that they are only expected to produce the PDGF-AA and/or PDGF-CC functional dimeric ligands (Reigstad et al. 2005). Moreover, while these ligands have been described to have distinct receptor specificities and functional effects, they seem to act in a redundant fashion in this case. Further study should be performed to confirm and characterize the impacts of PDGF-AA and PDGF-CC on myofibroblast behavior and gene expression, both in the drop coculture setup and in monoculture. Additionally, some form of *in vivo* confirmation would strengthen the conclusion that PDGF is truly the primary factor responsible for myofibroblast recruitment by colorectal cancer cells. Recruitment of local myofibroblasts by colorectal cancer lymph node metastases is a well-characterized phenomenon, and it is worth checking if PDGF-AA or PDGF-CC is the ligand responsible (Yeung et al. 2013). Overall, this chapter suggests that signaling through the PDGF pathway, specifically via ligands PDGF-AA and PDGF-CC, is an important facet of the interaction between colorectal cancer cells and surrounding myofibroblasts, and confirms that this pathway is a worthwhile clinical target for novel colorectal cancer treatments targeting tumor-stromal interactions.

Chapter Five

Categorization of Gene Expression Using Gaussian Mixture Modeling

5.1 Introduction

Global gene expression analysis has become a cornerstone of modern biological research, and is poised to enter the clinic as a powerful diagnostic tool. A concomitant explosion in analytical tools has allowed for these data to be mined and interpreted in increasingly thorough and intricate ways. Most recently, in parallel with the artificial intelligence revolution occurring in computing, a range of pattern recognition and machine learning tools have been applied to various kinds of high-throughput biological data. These approaches allow for the identification of complex features in data without human supervision, and provide novel interpretations and hypotheses.

The most widely used and easily implemented approach for pattern recognition in biological data is clustering, the unsupervised separation of data into categories. One such approach is model-based clustering, which involves the *a priori* stipulation that the data at hand is to be fitted with a certain probability model. In practice, the most common form of model-based clustering is Gaussian Mixture Modeling (GMM), which assumes that every component of the overall distribution is normal and thus limits the estimation to location (mean) and scale (standard deviation) parameters (Melnykov & Maitra 2010). This limitation ends up minimally problematic for most experimental data due to significant underlying technical noise, which makes any fine estimates of shape unreliable.

GMM has recently been applied to a number of questions involving biological data. For instance, mixed Gaussian models have recently been used to analyze heart rate variability, the performance of prostheses, and gene expression networks (Costa et al.

2012; Huang et al. 2005; Dai et al. 2009; Ko et al. 2007). Some analysis has also been performed on publically available gene expression data in cancers, uncovering novel gene interrelationships (Lee et al. 2012). The overarching benefit of GMM in these applications is twofold – it allows for continuous data to be subdivided into reasonable subgroups in an unsupervised manner, and it provides a clear reflection of the shape of each subgroup's distribution.

In this chapter, an approach using GMM is introduced as a means of analyzing gene expression data across samples. Specifically, GMM is used to detect and characterize bimodal expression patterns of individual genes. The various steps of the analysis are described and discussed. Validation tests on simulated data are shown to confirm the performance of the analyses. Examples are also shown of the methods applied to a microarray expression dataset from 78 independent colorectal cancer cell lines, as well as a publically available RNA-Seq expression set of a panel of primary colorectal tumors from The Cancer Genome Atlas (TCGA), and output datasets are obtained for further analysis.

5.2 Analytical Approach

The primary objectives of this analytical approach are twofold:

1. Develop an unsupervised method for determining whether a gene's expression pattern across a set of cell or tissue samples is unimodal or exhibits a clear bimodality, and make an appropriate cutoff separating the samples into the two component subgroups in the latter case (**Figure 5.1**)
2. For the genes found to exhibit a bimodal expression pattern, analyze the distributions of the component subgroups to infer cellular heterogeneities within the samples

This analytical pipeline takes as input a set of standard normalized and batch-removed gene expression datasets, obtained either from microarrays or by RNA-Seq. Performance of subsequent calculations was tested on both $\log_2$-transformed and raw expression data. Full details of the code are provided in the appendix.

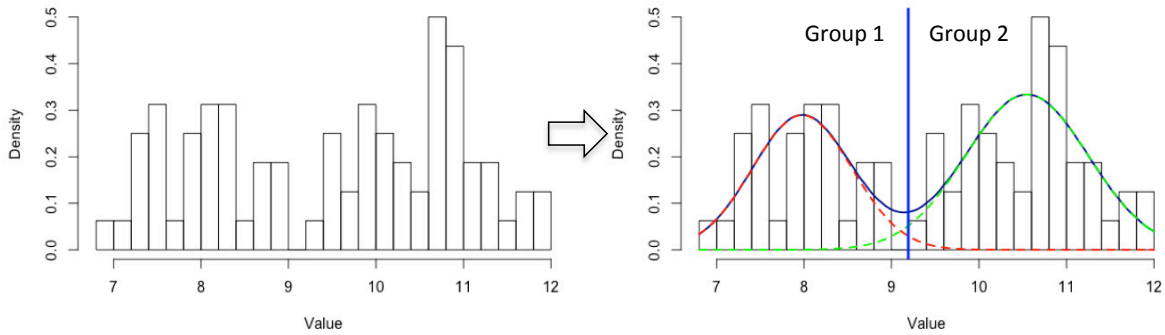


Figure 5.1: Example of desired GMM output on simulated data mimicking $\log_2$-transformed microarray gene expression data. Note the detection of bimodality, the fitting of two normal distributions to the underlying subgroups, and the vertical line indicating the cutoff value separating the two subgroups.

5.2.1 Gaussian Mixture Modeling

GMM was performed in R using the mclust package. Mclust is a popular and powerful package centered on fitting normal models to univariate and multivariate data using the expectation-maximization (EM) algorithm. In the particular case of the expression datasets, every gene's expression pattern is a univariate distribution, so the data is simple enough that the EM algorithm would be expected to reliably converge to a sensible fitting. When two normal distributions were fitted, their variances were allowed to vary. Model selection, in the particular case the decision whether a certain gene's expression pattern is better fitted by one normal distribution or two, is performed using the Bayesian Information Criterion (BIC), a log-likelihood function with a penalty term to reduce overfitting. Among the fits of one or two models to each gene's expression pattern, the one with the lower BIC was chosen.

5.2.2 Tail Trimming

One common caveat when fitting data with normal distributions of varying variances is that the best-scoring model may result in a narrow distribution fitted in the middle of a broad one, meaning that the observations on either end of the overall distribution are categorized as part of one subgroup (**Figure 5.2a**). This is undesirable in this case because one of the primary goals of the analysis is to make a single cutoff separating the samples into two subgroups. To handle such situations, a tail-trimming step was included. Every gene expression pattern scored as bimodal and exhibiting the problematic characteristic, namely that samples on either end of the overall

distribution were part of the same category, was subjected to a repeat round of GMM. This time, the samples from the shorter of the two tails categorized together were excluded, and the remaining samples were fitted with one or two normal distributions as before. If the output of the second GMM iteration was bimodal and lacking the tail problem, the newly calculated cutoff was used to separate the original distribution; else, if the second GMM gave a unimodal result, or a bimodal result that still had a tail problem, the gene's whole expression pattern was deemed unimodal. An example outcome of this step is shown in **Figure 5.2b**. The outputs of the aforementioned steps are two new datasets – namely, one containing expression data for every gene across the sample panel, scored as 1s or 2s depending on which category that sample fell into for each particular gene, and a second containing the means and variances of the one or two normal distributions fitted to each gene's continuous gene expression pattern. Applications of these datasets are discussed in Chapter 6.

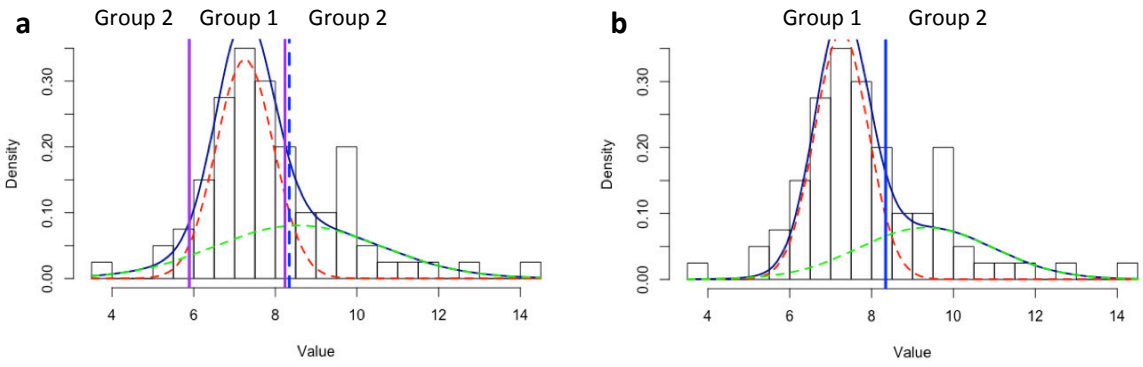


Figure 5.2: Example of tail trimming on simulated data mimicking $\log_2$-transformed microarray gene expression data. (**a**) This is the original GMM output; note that one subgroup is modeled by a much wider normal distribution than the other, resulting in two cutoffs and tails categorized in the same group. (**b**) Here, the left tail of five observations was omitted in the GMM calculation; this time, only one cutoff was predicted, and this was used to separate the full original distribution. Purple vertical lines indicate the two cutoffs of the first GMM iteration; blue vertical lines indicate the single cutoff following tail trimming.

5.3 Analytical Process

This analytical approach was applied to several in-house and public datasets. The in-house dataset consisted of gene expression data for 78 unique, non-duplicate colorectal cancer cell lines obtained using Affymetrix GeneChip HG-U133 Plus 2.0 microarray, normalized using the RMA approach and batch-removed in the Partek Genomics Suite software (**Table 5.1**). The public TCGA RNA-Seq dataset was obtained from the Genomic Data Commons (GDC) Data Portal. Samples were queried from the TCGA Colon Adenocarcinoma (COAD) and Rectum Adenocarcinoma (READ) databases using the TCGAbiolinks R package, filtered to keep only samples annotated as Primary Solid Tumor, and subsequently further filtered to remove duplicated samples from individual patients. Fragments Per Kilobase of transcript per Million mapped reads (FPKM) RNA-Seq counts for the remaining 622 samples were quantile-normalized to allow between-sample comparisons using the preprocessCore package in R.

C10	CCO7	HCT116	HT55	OUMS23	SW1116
C106	CL14	HDC111	ISCEROL	OXCO1	SW1222
C125PM	CL40	HDC114	JHCOLOY1	OXCO2	SW1417
C32	COLO201	HDC135	LIM1863	OXCO3	SW403
C70	COLO320DM	HDC142	LOVO	PCJW	SW48
C75	COLO678	HDC143	LS1034	PMFK014	SW480
C80	CW2	HDC57	LS123	RCM1	SW837
C84	DLD1	HDC73	LS174T	RKO	SW948
C99	Gp2D	HDC8	LS411	RW2982	T84
CACO2	HCA46	HDC82	LS513	RW7213	VACO10MS
CAR1	HCA7	HDC9	NCIH508	SKCO1	VACO429
CC20	HCC2998	HRA19	NCIH716	SNUC1	VACO4A
CCK81	HCC56	HT29	NCIH747	SNUC2B	VACO5

Table 5.1: Colorectal cancer cell lines making up the 78-cell line panel used for GMM analysis.

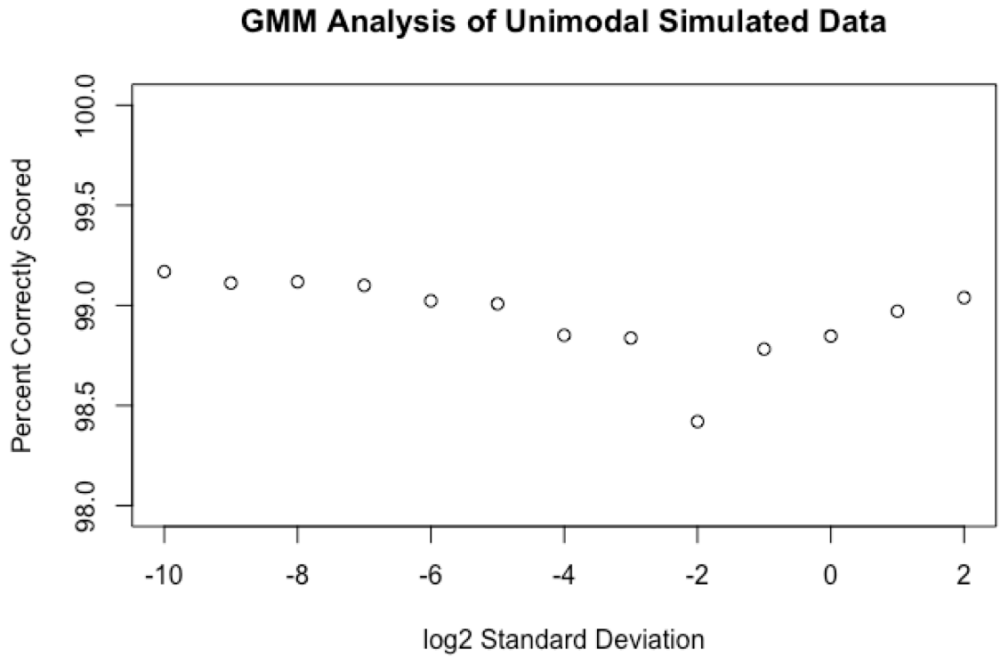


Figure 5.3: Accuracy in scoring a simulated unimodal normal distribution with varying standard deviations as unimodal. Note the minimum attained at standard deviation 0.25, suggesting that GMM detects unimodality somewhat less accurately for distributions with intermediate standard deviation compared to those with very low or very high standard deviation. 100,000 iterations of 80 samplings of random numbers were performed for each standard deviation.

5.3.1 GMM Validation and Examples

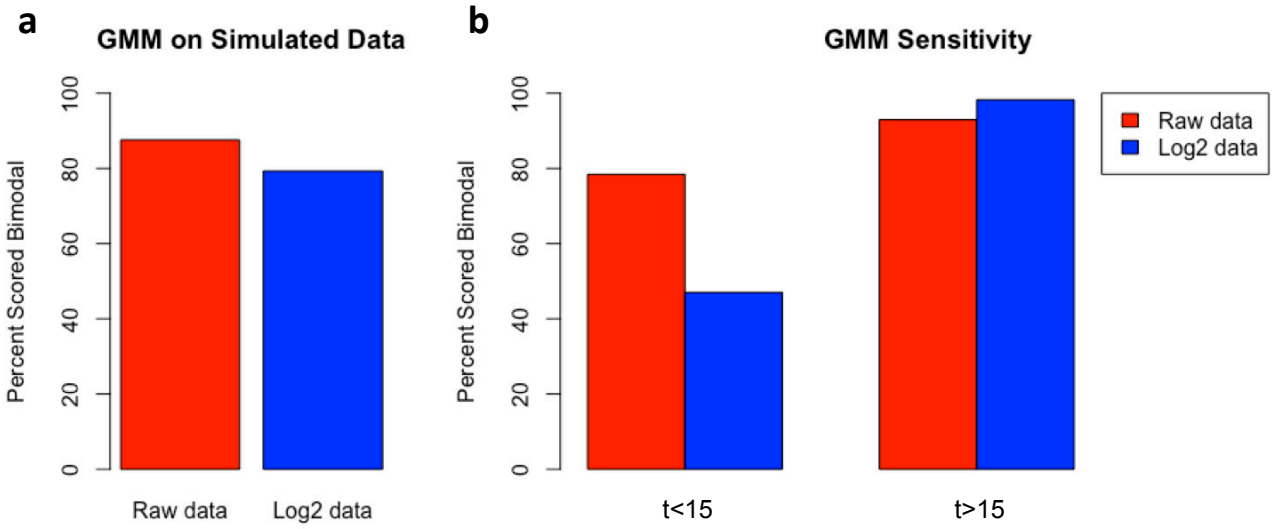
Before the GMM step was applied across the gene expression datasets, the procedure was tested on simulated data to obtain several performance metrics. First, the ability of GMM to correctly score unimodal distributions (i.e. the false-call rate of bimodality) was tested. For each given standard deviation as shown in **Figure 5.3**, 100,000 samplings of 80 random numbers from a normal distribution of that standard deviation with mean 0 were analyzed by GMM. The sample size of 80 was chosen to mimic the expression distributions seen across the colorectal cancer cell line panel. As it can be seen from **Figure 5.3**, the GMM step correctly identified the distribution as unimodal over 98% of the time for every standard deviation tested.

Next, GMM was tested on simulated bimodal distributions. Two means were each chosen from a uniform distribution between 3 and 15, roughly the range of $\log_2$-transformed microarray gene expression values across the cell lines (**Figure 5.8a**), and two standard deviations were each chosen from a uniform distribution between 0 and 1. A random integer n between 0 and 80 was chosen to be the number of samples in the first subgroup, and 80-n samples were then picked for the second subgroup. The overall distribution was the union of a set of n random numbers chosen from a normal distribution with the first mean and first standard deviation, and a set of 80-n random numbers chosen from a normal with the second mean and second standard deviation. GMM was then applied to that distribution, as well as to the inverse-$\log_2$-transformed distribution (2 raised to the power of the distribution), and the outputs were compared to each other and to the known underlying subgrouping; this process was iterated 1,000,000 times. In subsequent figures, the original simulated

distribution is referred to as "Log_2-transformed Data," while the inverse-transformed distribution is called the "Raw Data." **Figure 5.4** compares the ability of GMM to detect bimodality in log_2-transformed and raw data. As shown in **Figure 5.4a**, GMM exhibited an overall higher detection rate of bimodality in raw data compared to log_2-transformed data, with 87.6% and 79.3% of simulated distributions scored as bimodal, respectively. In order to further understand the performance of GMM on the simulated bimodal distributions, Welch's t-value was calculated for each pair of means and standard deviations corresponding to individual simulated distributions; this is a combined value that takes into account both the separation of the means of the two subgroups as well as the variances of the subdistributions, resulting in an overall quantification of the overlap between the subdistributions:

$$t = \frac{|\overline{X_1} - \overline{X_2}|}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Here, $\overline{X_1}$ and $\overline{X_2}$ correspond to the means of the two subgroups, s_1 and s_2 correspond to their standard deviations, and N_1 and N_2 correspond to the numbers of elements in each subgroup. As seen in **Figure 5.4b**, it turns out that while GMM is substantially more sensitive on raw data for distributions with small t-values less than 15, it actually performs slightly better on log_2-transformed data for t-values greater than 15. **Figure 5.4c** displays the overall breakdown of the 1,000,000 simulated distributions based on the form of the data (raw or log_2-transformed) in which GMM detected bimodality; notably, while GMM detects bimodality in raw but not log_2-transformed data for 14.8% of the distributions, the reverse is also true for 6.5% of distributions.



c

Classification of raw data: / Classification of log_2 data:	Unimodal	Bimodal
Unimodal	5.9%	14.8%
Bimodal	6.5%	72.8%

Figure 5.4: GMM detection of bimodality in raw and log_2-transformed simulated data. (**a**) When applied to simulated data, GMM detects bimodality in 87.6% of simulated bimodal distributions when applied to raw data and in 79.3% of the same distributions when they are in the log_2-transformed form. (**b**) When comparing the success of GMM at detecting bimodality in log_2-transformed and raw data across Welch's t-values, the method detects 78.4% of simulated bimodal distributions in raw data for t<15 and only 47.0% in log_2-transformed data. On the other hand, for t>15, GMM detects bimodality for 92.9% of simulated distributions in raw data and a striking 98.3% in log_2-transformed data. (**c**) The 1,000,000 simulated distributions fell into four categories based on whether or not they were scored as bimodal by GMM in raw and/or in log_2-transformed form. Note that substantial numbers of both discordant scorings are present, amounting to over 20% of the total.

In order to further investigate the dependence of GMM bimodality detection on t-value, the t-values of successfully detected distributions were plotted as a subset of all simulated t-values for raw and $\log_2$-transformed data in **Figure 5.5a**. Notably, and as expected, GMM is more successful at detecting bimodality in raw data for small t-values. Further breakdown in **Figure 5.5b**, paralleling the table in **Figure 5.4c**, confirms that the improved bimodality detection in raw data mainly derives from distributions with small t-values, while the distributions detected as bimodal in $\log_2$-transformed but not raw data show a relatively mild asymmetry across t-values.

Next, the ability of GMM to detect bimodality in distributions with small subgroups was examined. **Figure 5.6a** shows the sizes of the subgroups with smaller mean in successfully detected distributions plotted as a subset of all smaller-mean subgroup sizes for raw and $\log_2$-transformed data. While raw data exhibits an increased level of bimodality detection for larger smaller-mean subgroup sizes compared to $\log_2$-transformed data, there is a notable dip in detection for smaller subgroup sizes. **Figure 5.6b** again breaks the results down as in **Figure 5.4c**, and clearly shows that while raw data exhibits an improved detection across all smaller-mean subgroup sizes, $\log_2$-transformed data and not raw data allows for a substantially improved detection of small subgroups of smaller mean.

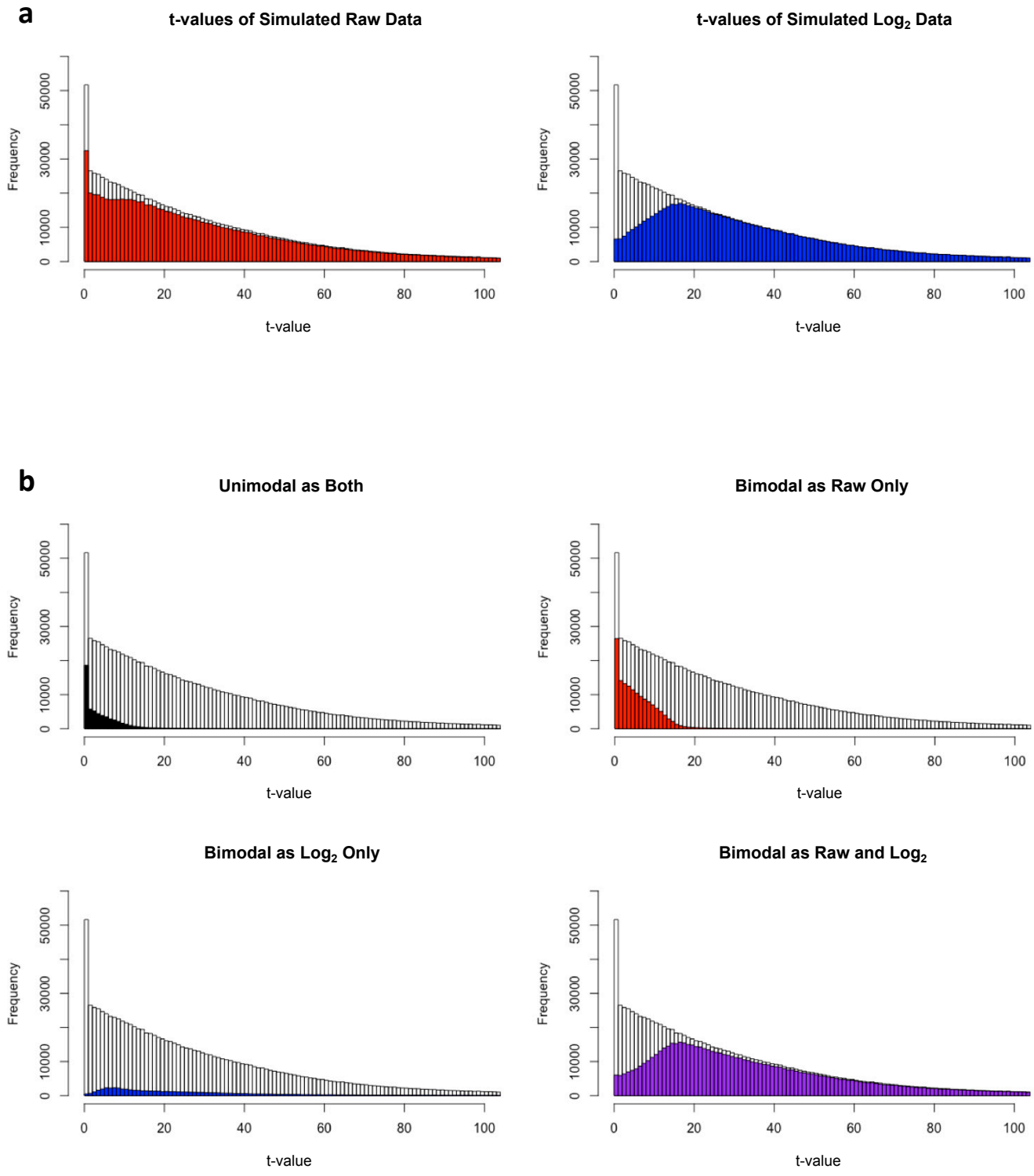


Figure 5.5: GMM detection of bimodality in raw and $\log_2$-transformed simulated data by t-value. The white background histogram summarizes all 1,000,000 t-values of the simulated distributions; the colored distribution corresponds specifically to the t-values of the simulated distributions detected as bimodal in the indicated form (red: raw data, blue: $\log_2$-transformed data, purple: both forms, black: neither form). (**a**) Overall detection of bimodality in raw and $\log_2$-transformed data, by t-value. Note that GMM seems to detect more bimodal distributions with low t-value in raw data compared to $\log_2$-transformed. (**b**) Paired raw and $\log_2$-transformed bimodality detection, paralleling **Figure 5.4c**. Note that the distributions detected as bimodal in raw but not $\log_2$-transformed data predominantly have t<20, while those detected as bimodal in $\log_2$-transformed but not raw data show a milder asymmetry across t-values.

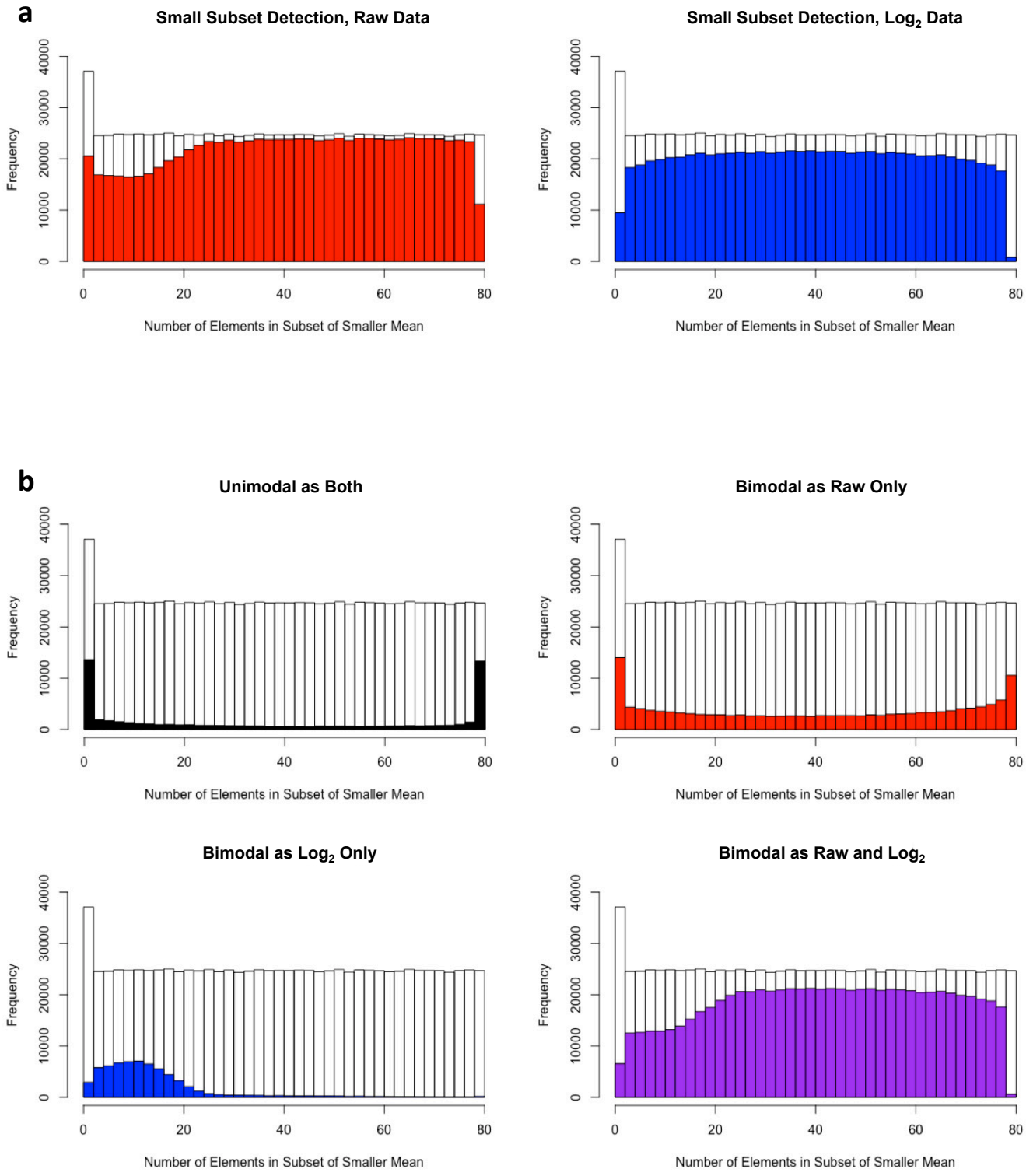


Figure 5.6: GMM detection of bimodality in raw and $\log_2$-transformed simulated data by size of the simulated subgroup with smaller mean. The white background distribution is the histogram of all 1,000,000 sizes of the smaller-mean subgroup in the simulated distributions; the colored distribution corresponds specifically to the smaller-mean subgroup sizes of the simulated distributions detected as bimodal in the indicated form (red: raw data, blue: $\log_2$-transformed data, purple: both forms, black: neither form). (**a**) Overall detection of bimodality in raw and $\log_2$-transformed data, by size of smaller mean subgroup. Note that GMM detects more bimodal distributions across the range of smaller mean subgroup sizes in raw data compared to $\log_2$-transformed, but shows a distinct dip for subgroups of size less than 20. (**b**) Paired raw and $\log_2$-transformed bimodality detection, paralleling **Figure 5.4c**. Note that the distributions detected as bimodal in $\log_2$-transformed but not raw data predominantly have smaller-mean subgroup size less than 20, but no mirror advantage for large subsets of smaller mean is seen when scoring raw data.

Finally, for the distributions correctly detected as bimodal in either raw or log_2-transformed data, concordance between the predicted subgrouping and the simulated subgrouping was examined. Log_2-transformed data (**Figure 5.7a**) exhibited nearly 100% concordance between simulated and predicted subgroup membership for distributions with t-values over 100, while raw data (**Figure 5.7b**) still exhibited concordances as low as 60% nearly up to t=200 (concordance cannot be below 50% because subgroup labels are interchangeable). Note from **Figures 5.4-5.5** that GMM is almost exactly as successful at detecting bimodality in raw data as in log_2-transformed data for t>20, meaning that the observed difference in **Figure 5.7** is not due to an increased ability of GMM to detect bimodality in raw data compared to log_2-transformed, but rather a true difference in accuracy.

Combining the insights above, it was determined that log_2-transformed data should be used for GMM due to its accentuation of small subsets of small mean and overall improved subgrouping accuracy, while raw data exhibited improved bimodality detection only for distributions with small t-values and very closely overlapping subgroups.

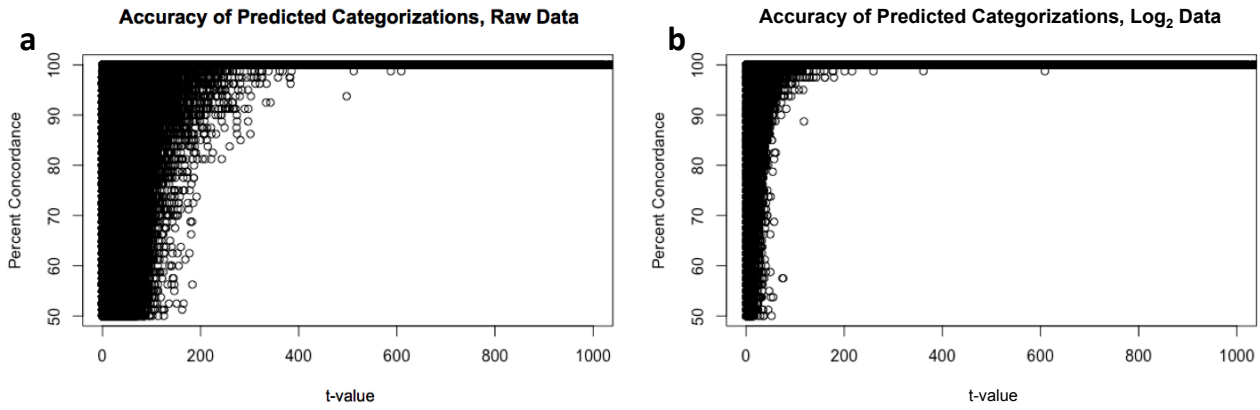


Figure 5.7: Concordance of simulated and GMM-predicted sample subgroup membership in $\log_2$-transformed and raw data. (**a**) Note the wide extent of poorly concordant predictions for raw data, with >95% concordance not fully reached until t>300. (**b**) Note the concentration of poorly concordant predicted subgroups in the t<100 range for $\log_2$-transformed data, and >95% concordance for a vast majority of distributions with t>100.

5.3.2 GMM and Tail Trimming in Real Data

Following the above verification, GMM was applied to the $\log_2$-transformed in-house cell line gene expression dataset. Of the 54,675 probesets on the HG-U133 Plus 2.0 chip, 12,719 (23.3%) were scored as bimodal; out of these, 3,518 (27.7%) exhibited two cutoffs as in **Figure 5.2a**. Following tail removal and repeat application of GMM for these 3,518 probesets, 3,329 (94.6%) were deemed bimodal with a single cutoff, 172 (4.9%) were deemed unimodal, and 17 (0.5%) still exhibited two cutoffs and were deemed uncategorizable, resulting in their manual scoring as unimodal as described above. Thus the final output was 12,530 of 54,675 (22.9%) probesets split into two subgroups by a single cutoff, with the rest of the probesets exhibiting unimodal distributions across the cell line panel.

GMM was also applied to $\log_2$-transformed quantile-normalized FPKM RNA-Seq data from TCGA. Of the 53,393 genes in the dataset with nonzero FPKM values for at

least two samples, 19,929 (37.3%) were scored as bimodal; out of these, 10,436 (52.4%) exhibited two cutoffs. Following tail trimming for these 10,436 genes, 10,084 (96.6%) were deemed bimodal with a single cutoff, 292 (2.8%) were deemed unimodal, and 60 (0.6%) still exhibited two cutoffs, were deemed uncategorizable, and were manually scored as unimodal. The final output consisted of 19,571 of 53,393 (36.7%) genes split into two subgroups by a single cutoff, with the remaining genes exhibiting unimodal distributions across the 622 tumor samples. All zero values in the original FPKM dataset were marked as NA upon $\log_2$ transformation and excluded from the GMM step, meaning that some genes had normal distributions fitted only to a subset of the 622 tumor samples.

5.3.3 Example Outputs

In order to showcase the power of this GMM approach to the categorization of continuous gene expression data from both microarray and RNA-Seq, various outputs are shown in **Figures 5.8-5.11** for a representative selection of probesets and genes, comparing $\log_2$-transformed and raw data as well as examining various results of the tail trimming process. Rankit plots, comparing of the actual expression values of the particular gene or probeset across samples compared to the values predicted by the fitted Gaussian mixture, are shown to confirm the quality of the fit.

Figure 5.8 shows the overall distributions of observed microarray gene expression values in raw and $\log_2$-transformed form, as well as GMM results for two well-known genes active in the colonic epithelium. *CDX1*, a transcription factor known to control the differentiation of enterocytes, is evidently bimodally expressed across the cell line panel, with a convincing cutoff made between subgroups in $\log_2$-transformed data and a marginally less compelling cutoff in raw data (**Figure 5.8b**) (Chan et al. 2009). *CDH1* encodes E-Cadherin, a key structural protein in differentiated epithelial cells. **Figure 5.8c** shows that a majority of cell lines express high levels of *CDH1*, but that a small proportion produces less of the mRNA; this subgroup is not detected in raw data but clearly delineated in $\log_2$-transformed data. Examples of the three possible tail-trimming outcomes in $\log_2$-transformed microarray data are shown in **Figure 5.9**.

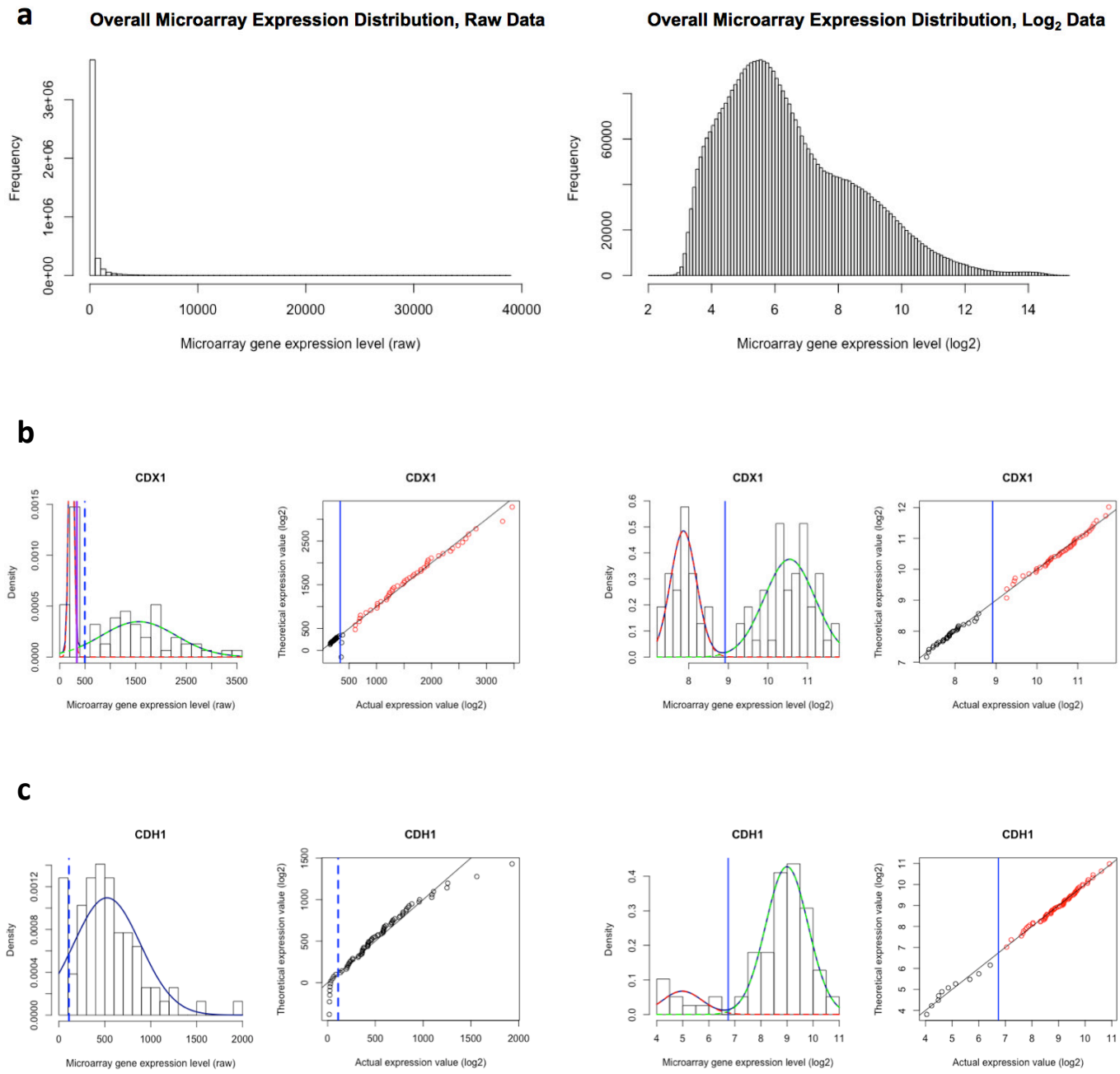


Figure 5.8: GMM application to real raw and $\log_2$-transformed microarray data from 78 colorectal cancer cell lines. (**a**) Overall expression level distributions across all cell lines and probesets. Note that the raw expression values have a lognormal distribution with a broad positive tail, while the $\log_2$-transformed values contain two visible normal-like subdistributions. (**b,c**) Comparisons between raw (left) and $\log_2$-transformed (right) data for two representative probesets. (**b**) shows the probeset for the *CDX1* gene, for which GMM detects bimodality in both raw and $\log_2$-transformed data, but makes different cutoffs as shown. (**c**) shows the probeset for *CDH1*, which exhibits a small subgroup of smaller mean that is clearly detected in $\log_2$-transformed data but obscured by the larger subgroup of larger mean in raw data. Blue vertical lines show the cutoff made by GMM on $\log_2$-transformed data, and the purple line is the cutoff made on raw data. Near-linearity of rankit plots confirms that two Gaussian distributions fit the $\log_2$-transformed data well for both probesets; the rankit plots for the raw data are also not bad, but the subdistributions are more skewed and the *CDH1* fit does not detect the small subset of small mean.

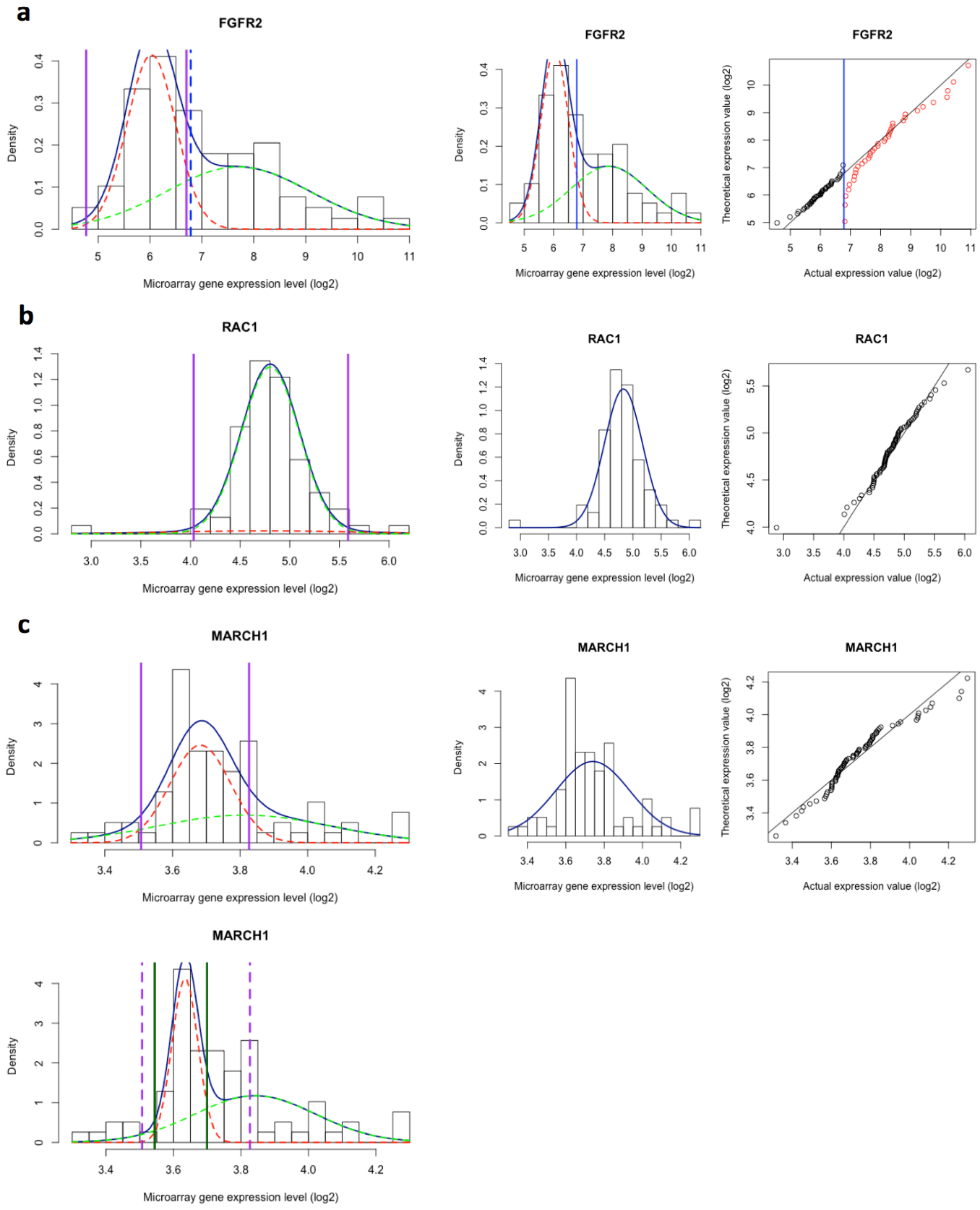


Figure 5.9: Tail trimming on $\log_2$-transformed microarray data from 78 colorectal cancer cell lines. (**a**) This *FGFR2* probeset was deemed bimodal with a single cutoff between subgroups following tail trimming. The first GMM result is shown on the left with purple lines indicating the two cutoffs; upon removal of the shorter left tail, the second GMM iteration resulted in the cutoff shown on the left plot. (**b**) This *RAC1* probeset was deemed unimodal following tail trimming, with original cutoffs as shown on the left. (**c**) Tail trimming failed for this *MARCH1* probeset. As shown on the left, the first GMM iteration (top) resulted in two cutoffs (purple), and upon removal of the shorter left tail, the second GMM iteration (bottom) again gave two cutoffs (green). Thus the probeset was manually scored as unimodal, as shown on the right. Rankit plots are shown for the final fits; linearity is not perfect but reasonable considering the initial GMM results.

Analogously to the treatment of microarray data above, **Figure 5.10** examines the overall distributions of observed RNA-Seq expression values in raw and $\log_2$-transformed form across the 622 patient tumor samples. *MLH1*, a gene encoding a key DNA mismatch repair protein, is known to be mutated or inactivated by hypermethylation in 10-15% of colorectal cancers (Boland & Goel 2010). A subset of tumors with low *MLH1* expression is detected in both $\log_2$-transformed and raw data as shown, following tail trimming in the latter case. Notably, the subset consists of slightly over 10% of the 622 primary tumors, in concordance with previously reported rates of *MLH1*-deficient colorectal cancers. *CDX1* is strongly expressed in a vast majority of tumors, though a small tail of low-expressing tumors is detected in $\log_2$-transformed data. GMM also detects bimodality in the raw *CDX1* expression distribution, though the cutoff it makes there is not at all reasonable. As for the microarray data, **Figure 5.11** shows the three possible tail-trimming outcomes in the $\log_2$-transformed RNA-Seq data.

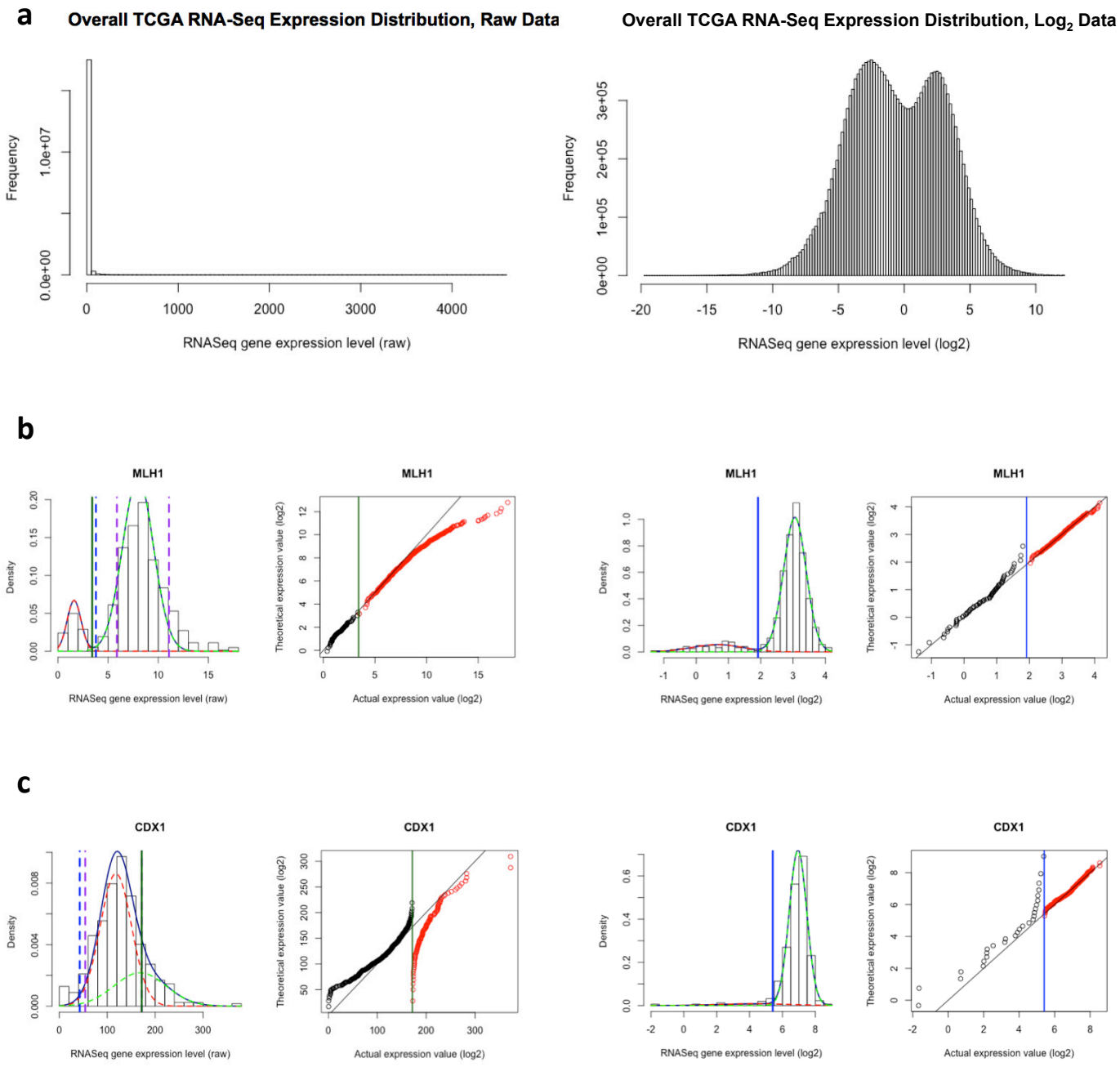


Figure 5.10: GMM application to real raw and $\log_2$-transformed RNA-Seq data from 622 primary colorectal tumors in the TCGA database. (**a**) Overall expression level distributions across all samples and genes. Akin to the microarray data, the raw expression values have a lognormal distribution with a broad positive tail, while the $\log_2$-transformed values contain two visible normal-like subdistributions. (**b,c**) Comparisons between raw (left) and $\log_2$-transformed (right) data for two representative genes. (**b**) GMM detects bimodality in both raw and $\log_2$-transformed data for the *MLH1* gene, but makes different cutoffs as shown and requires tail trimming in the raw case. (**c**) GMM also detects bimodality in both raw and $\log_2$-transformed data for *CDX1*. The raw data again requires tail trimming, and the final cutoffs vary significantly between the two types of data. Blue vertical lines show the cutoff made by GMM on $\log_2$-transformed data, the purple lines are the cutoffs on raw data after the first GMM iteration, and the green lines show the cutoffs on raw data following tail trimming. Note that for both of these genes, the rankit plots for the $\log_2$-transformed data are much more clearly linear than for the raw data, further supporting its use in further analyses.

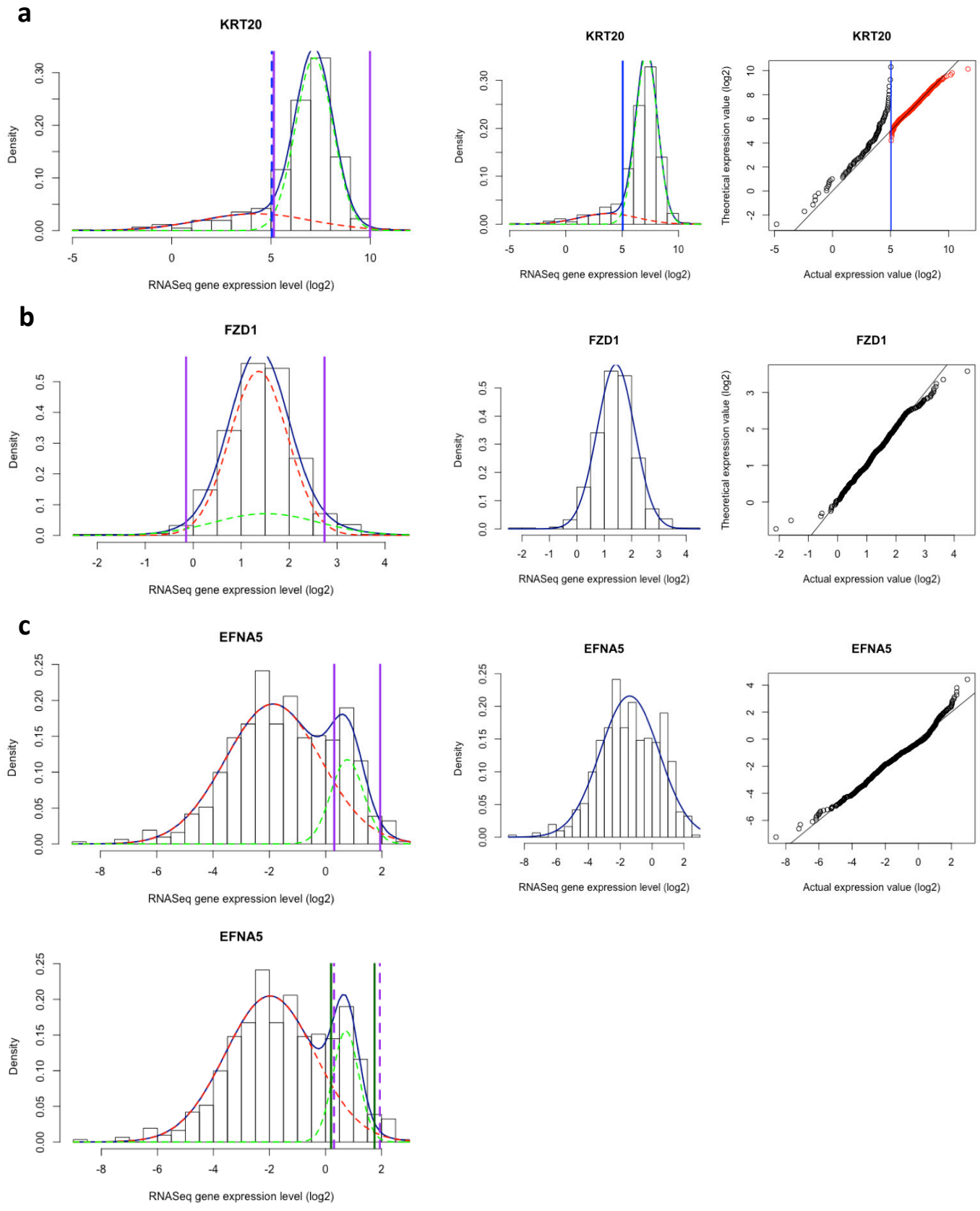


Figure 5.11: Tail trimming on $\log_2$-transformed RNA-Seq data from 622 primary colorectal tumors. (**a**) The *KRT20* gene was deemed bimodal with a single cutoff following tail trimming. Purple vertical lines on the left indicate the two cutoffs determined by the first GMM iteration; upon removal of the shorter right tail, the second GMM iteration predicted the single cutoff shown on the right (blue). (**b**) The *FZD1* gene was deemed unimodal following tail trimming, with original cutoffs shown on the left. (**c**) Tail trimming failed for the *EFNA5* gene. As shown on the left, the first GMM iteration (top) resulted in two cutoffs (purple), and removal of the shorter right tail led to the second GMM step (bottom) again resulting in two cutoffs (green). The probeset was manually scored as unimodal, shown on the right. Rankit plots for the final fits show reasonable linearity, confirming that tail-trimming was successful.

5.4 Discussion

Carcinogenesis is an evolutionary process, characterized by the accumulation of genetic changes in cancer cells and the retention of advantageous changes in the population by natural selection. Selection acts upon phenotypic heterogeneity, which presents as a consequence of genetic variation. In order to affect selective advantage, genetic changes must therefore be striking enough to significantly alter the phenotype. This explains why tumor evolution is often driven by specific pivotal mutations, which are either strikingly deleterious or strikingly activating at a certain genetic locus, while other mutations barely contribute to the cells' selective advantage. It is therefore expected that genetic changes critical to specific events along the evolutionary path of a certain type of cancer should be present in a clearly heterogeneous distribution across a panel of such cancers, separating the samples that have passed the evolutionary event in which this gene is involved from those that have not yet reached this event or evolved along a different path.

In this chapter, Gaussian Mixture Modeling (GMM) was applied as an unsupervised computational means of detecting bimodal gene expression patterns across a panel of colorectal cancer cell lines and primary tumor samples. Genes with clearly bimodal expression patterns would be expected to participate in or vary as a consequence of impactful events in the cancers' evolution, possibly with substantial effects on the tumors' selective advantage.

As discussed at length in **Section 5.3.1**, GMM is a powerful tool for the detection of subgroups of bimodal distributions, especially when the data is $\log_2$-transformed. This

seems to be due to two benefits of logarithmic transformation. First of all, this transformation condenses wide distributions of large numbers onto a smaller domain, meaning that their variances are narrowed and overlaps between subgroups are limited. On the other hand, the transformation widens the domain for small values, allowing for a clearer detection of small subsets with low expression of the gene at hand. This explains why although one may expect more accurate subgroup detection in raw data due to its wider domain, $\log_2$-transformed data actually gives better predictions of subgroup membership for individual samples when subjected to GMM (**Figure 5.7**). The increased performance of GMM on $\log_2$-transformed data is also confirmed on real gene expression data from both microarray and RNA-Seq experiments (**Figures 5.8, 5.10**). In fact, it is already evident from both the $\log_2$-transformed microarray and RNA-Seq profiles across all samples and genes that two component subgroups corresponding to low and high gene expression arise from both platforms (**Figures 5.8a, 5.10a**).

Although the real gene expression data shown in **Figures 5.8-5.11** is piecemeal and chosen to exemplify representative distributions of expression rather than gene interrelationships or functional associations, it nevertheless provides insights for further analysis. First of all, while a small fraction of genes exhibit obvious bimodality such as *CDX1* in the cell lines (**Figure 5.8b**), most heterogeneities are subtler. Some bimodal distributions exhibit substantial subgroup overlap akin to *FGFR2* in the cell lines or *KRT20* in the tumor samples (**Figures 5.9a, 5.11a**). Subgroups may vary significantly in their sizes, and certain characteristics may only be found in small subsets of samples, as for *CDH1* in the cell lines or *MLH1* in the tumors (**Figures 5.8c, 5.10b**). Also, cell lines and primary tumors seem to exhibit

vastly different proportions of samples with certain characteristics – while about half of the cell lines express low amounts of *CDX1*, a vast majority of tumors exhibit high levels of this gene (**Figures 5.8b, 5.10c**). These differences between cell lines and primary tumors, as well as associations between the expression patterns of different genes and the variations of subgroup shapes within individual gene expression distributions, will be discussed further in Chapter 6.

Chapter Six

Association Analysis of Categorical Gene Expression Data

6.1 Introduction

During natural selection, different alleles for the same gene may provide quantifiably different advantages to the overall phenotype in a given environment, and the fitness of the individual is then a function of these advantage weights across all genes that its genome encodes. As alluded to in the previous chapter, cancer types are known to be defined by limited sets of driver mutations, which dominate the selective advantages of the cells in which they arise and determine the primary characteristics of patient tumors (Stratton et al. 2009). While many mutations may have slight effects on fitness, it is these driver mutations that induce the "phase transitions" accounting for major steps in the punctuated equilibrium of cancer progression (Cross et al. 2016). To orchestrate such cardinal shifts in phenotype, driver mutations must significantly alter the activity of one or several genes, either by affecting expression or by fundamentally disturbing the function of encoded proteins through alterations in coding regions. Conversely, sharp discontinuities in gene expression delineating subpopulations within large collections of cancers are expected to arise as consequences of such switch-like changes. Thus, while combinations of many alleles usually result in ostensibly continuous phenotypes, obscuring the Mendelian basis of inheritance, these underlying discrete characteristics can be inferred from downstream properties when single alleles have especially strong effects.

In Chapter 5, a method was devised to detect and characterize such gene expression discontinuities in panels of colorectal cancer cell lines and publically available patient tumor datasets. Specifically, it was determined whether each gene's expression pattern was better fitted by one or a mixture of two Gaussian distributions, and in the

latter case, a cutoff was made to separate the samples into two groups. Besides the binary expression datasets derived in this manner, means and variances for the subdistributions were predicted, allowing for robust quantification of the extent of separation/overlap between the subdistributions. This chapter expands on the previous to take advantage of the Gaussian Mixture Model outputs. Specifically, the binary datasets are first used to calculate associations between gene expressions based on 2x2 tables, and χ^2 tests are used to determine significance of these associations. Next, an interactive interface is built using the shiny package in R to allow for optimal analysis of these associations. Specific examples of the application of this program to the identification and gene expression characterization of differentiating and goblet cell-producing cell lines and tumors, as well as those exhibiting signs of epithelial-mesenchymal transition, are reported. Implications of the findings to various tumor behaviors and heterogeneities are discussed.

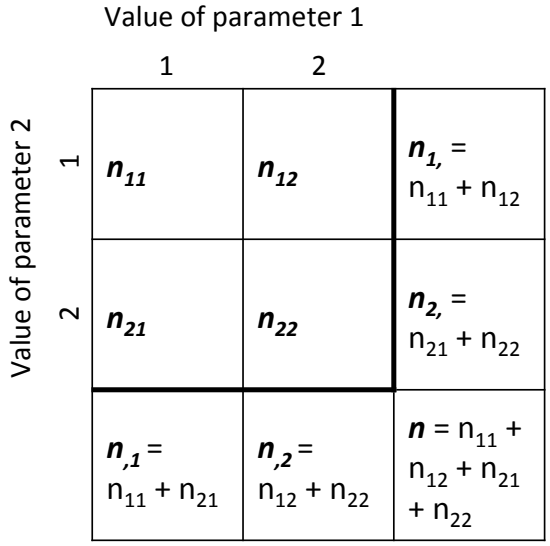
6.2 Analytical Approach

6.2.1 2x2 Analysis of Gene Expression and Mutation

The two datasets of binary expression values for cell lines and TCGA colorectal tumors from the previous chapter consist of every sample coded as either a 1 (low expression) or 2 (high expression) for every probeset or gene determined to be bimodal by the GMM algorithm. In the case of the cell lines, these datasets were additionally augmented with known mutation status for genes known to be involved in colorectal carcinogenesis, namely *APC*, *TP53*, *CTNNB1*, *KRAS*, *BRAF*, *PIK3CA*, *FBXW7*, *TGFBR2*, and *SMAD4*, as well as mismatch repair status, lumen formation in Matrigel due to differentiation, goblet cell formation assayed by immunostaining for secreted mucin MUC2, myofibroblast recruitment ability (Chapter 4), and WNT pathway activity. These characteristics were derived from published literature or determined in-house using various means, and binarization was performed either using GMM as previously for continuous data (WNT activity), or by presence/absence for the mutations and other variables.

In order to calculate association values between bimodal probesets, characteristics, or genes, 2x2 tables were formed for every pair of such variables in both datasets, quantifying the number of samples falling into each combination of values 1 and 2 for either of the variables. For a given 2x2 table, two outputs were calculated as described in **Figure 6.1**: the *r* coefficient, also known as ϕ, which quantifies the strength and direction of the association (positive or negative), and the p-value from Pearson's χ^2 Test with Yates's correction for continuity. Some of the non-expression data added to

the cell line dataset were incomplete, and 2x2 tables involving these data included only the cell lines with complete data for both variables at hand; thus, the total number of samples in the 2x2 table (*n*) was also extracted.



$$r = \frac{n_{11} * n_{22} - n_{12} * n_{21}}{\sqrt{n_{1,} * n_{2,} * n_{,1} * n_{,2}}}$$

$$\chi^2 = \frac{n * (\max\left(0, |n_{11} * n_{22} - n_{12} * n_{21}| - \frac{n}{2}\right))^2}{n_{1,} * n_{2,} * n_{,1} * n_{,2}}$$

Figure 6.1: Example 2x2 table, with *r* and χ^2 calculations shown. Note that Yates's correction introduces the subtracted $\frac{n}{2}$ term in the numerator of the χ^2, and if the corrected diagonal product difference is negative following that subtraction, a value of 0 is returned.

6.2.2 Interactive Association Mining Interface

Datasets for the cell line and TCGA data containing the *r*, p, and *n* values were returned from the 2x2 analysis. In order to facilitate the interpretation of these outputs, an interactive program was built to query these datasets and display their contents in a meaningful way. The primary objective of the program was to take as input either one or a pair of variables in the datasets (probesets, characteristics, or genes), and to display either the direct comparison between the pair or a list of the top most associated other variables in case only one is queried. In addition, options were included to only list outputs with t-values in a certain range, calculated from the predicted means and standard deviations of the two normal subsets derived by GMM, or with at least one sample with expression value above a chosen cutoff. As seen in the previous chapter, bimodal distributions with low t-values have poor separation between the subgroups, and the predicted cutoff between the subgroups is not expected to be very reliable. On the other hand, if a certain gene has very low expression for every single sample in the original dataset, it is expected to be negative in all samples and thus any bimodality in its expression is not considered meaningful. For every continuous query and output variable, the program was designed to return rank, rankit, and histogram plots as seen in the previous chapter to give a direct visualization of the variable's distribution across the sample. Additionally, every 2x2 table is displayed, along with the *r*, p, and *n* values, and a data table containing the values of the query and output variables for every sample is also shown. The program was implemented in R using the shiny package; the full code is presented in the appendix, and the program is also available online (Oxford VPN Required for Access): rstudio.molbiol.ox.ac.uk/shinyApp/pkalugin/caa2/

6.3 Analytical Outputs

In order to test its functionality, this program was applied to three characteristics known to be heterogeneous across the cell line panel and clinical tumors, namely differentiation, goblet cell formation, and epithelial-mesenchymal transition (EMT). Queries were performed either by selecting a single gene/probeset, in which case a list of top genes associated with the query, ordered by $|r|$, was returned, or by choosing two genes/probesets, in which case only a direct comparison between them was made. Additionally, in the case of the in-house microarray dataset, one of several probesets could be chosen from a dropdown menu for each gene that mapped to multiple probesets on the microarray chip. Also, the in-house dataset was further extended by the addition of various phenotypes and mutations as previously described; these variables could be chosen from a dropdown menu. Several example screenshots of the interface are shown in **Figure 6.2**.

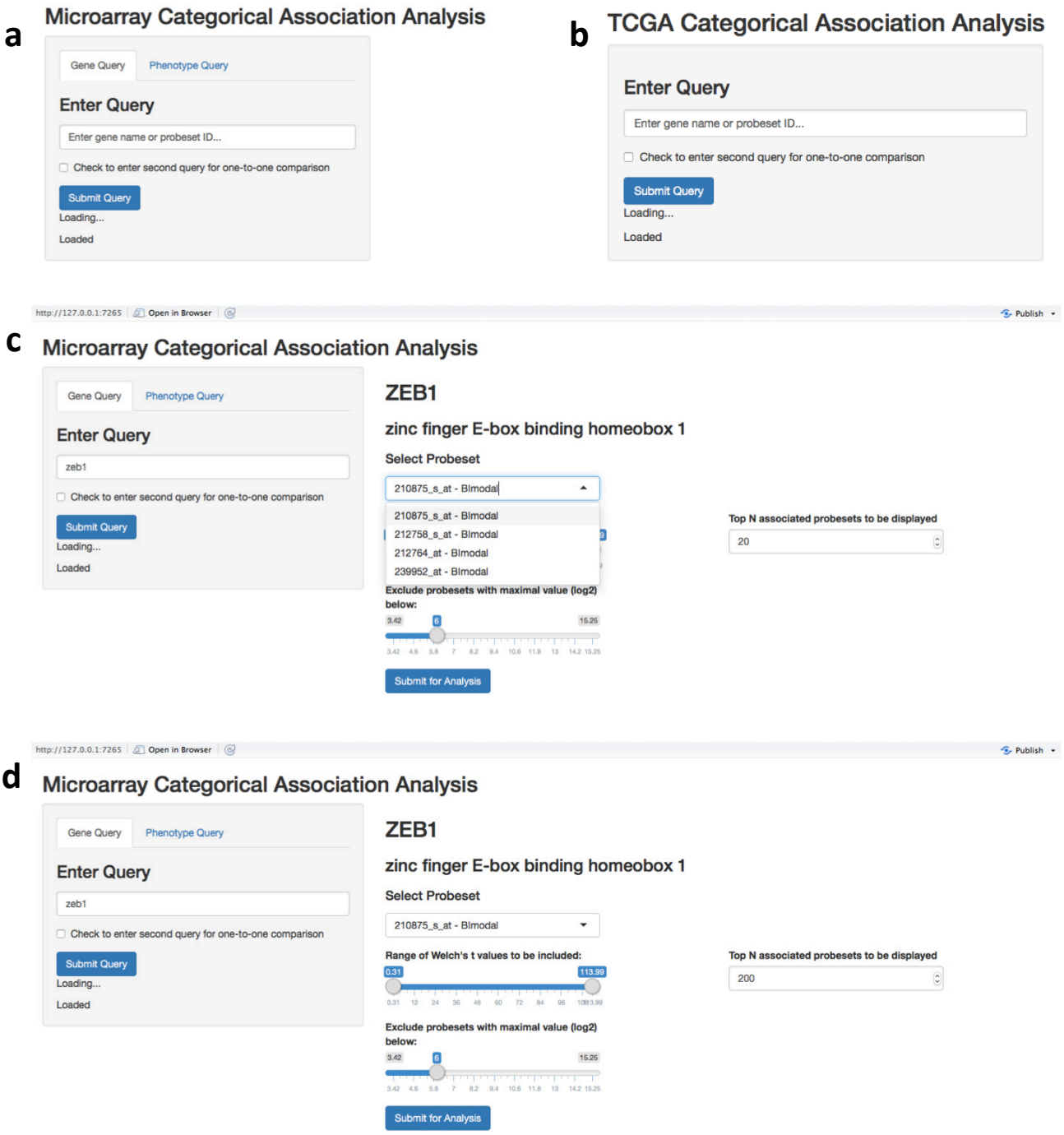


Figure 6.2: Example screenshots of the interface to query the outputs of the GMM analysis. The initial query is entered as shown in (**a**) and (**b**) in the interfaces for cell line microarray and TCGA RNA-Seq data, respectively. (**c**) When a query is found in the dataset, a dropdown of associated probesets is displayed in the case of the microarray data (RNA-Seq has only one output for each gene). (**d**) Other options to specify desired outputs include choosing a range of t-values, excluding all genes or probesets with maximal value across the samples below a certain value, and picking how many of the top outputs most associated with the query are to be displayed.

6.3.1 Differentiation in Colorectal Cell Lines and Primary Tumors

Colorectal cancers commonly retain some of the characteristics of differentiating colonic epithelium, and often express a range of differentiation markers both in patients and in cell culture (Ashley et al. 2013). It has previously been reported that the primary transcription factor responsible for the differentiation of the enterocytes in the colon is caudal-related homeobox protein CDX1, possibly in concert with the family member CDX2 (Chan et al. 2009; Grainger et al. 2013). To begin with, the cell line dataset was queried for the lumen formation phenotype, in the hopes of picking up these proteins. As shown in **Figure 6.3a**, *CDX1* and *CDX2* expression patterns are indeed significantly positively associated with the lumen formation phenotype. Additionally, two other well-known differentiation markers are shown to be significantly associated with lumen formation – carcinoembryonic antigen-related cell adhesion molecule 5 (*CEACAM5*) and villin 1 (*VIL1*). In order to verify the association of *CDX1* expression with differentiation, this gene was now queried, and highly significant results were obtained for its association with the other three genes from before (**Figure 6.3b**). As confirmation of the associations found in cell lines, the same query for *CDX1* was performed on the primary tumor data (**Figure 6.4**). It is evident that the same strong associations for these genes were found again. Notably, while 39.7% (31/78) of cell lines exhibited low *CDX1* expression, only 4.5% (28/622) of primary tumors were characterized as low expressers by the GMM algorithm, suggesting that the cell line panel overrepresents poorly differentiated cancers, possibly because such cells are more capable of surviving in culture (Masters 2000).

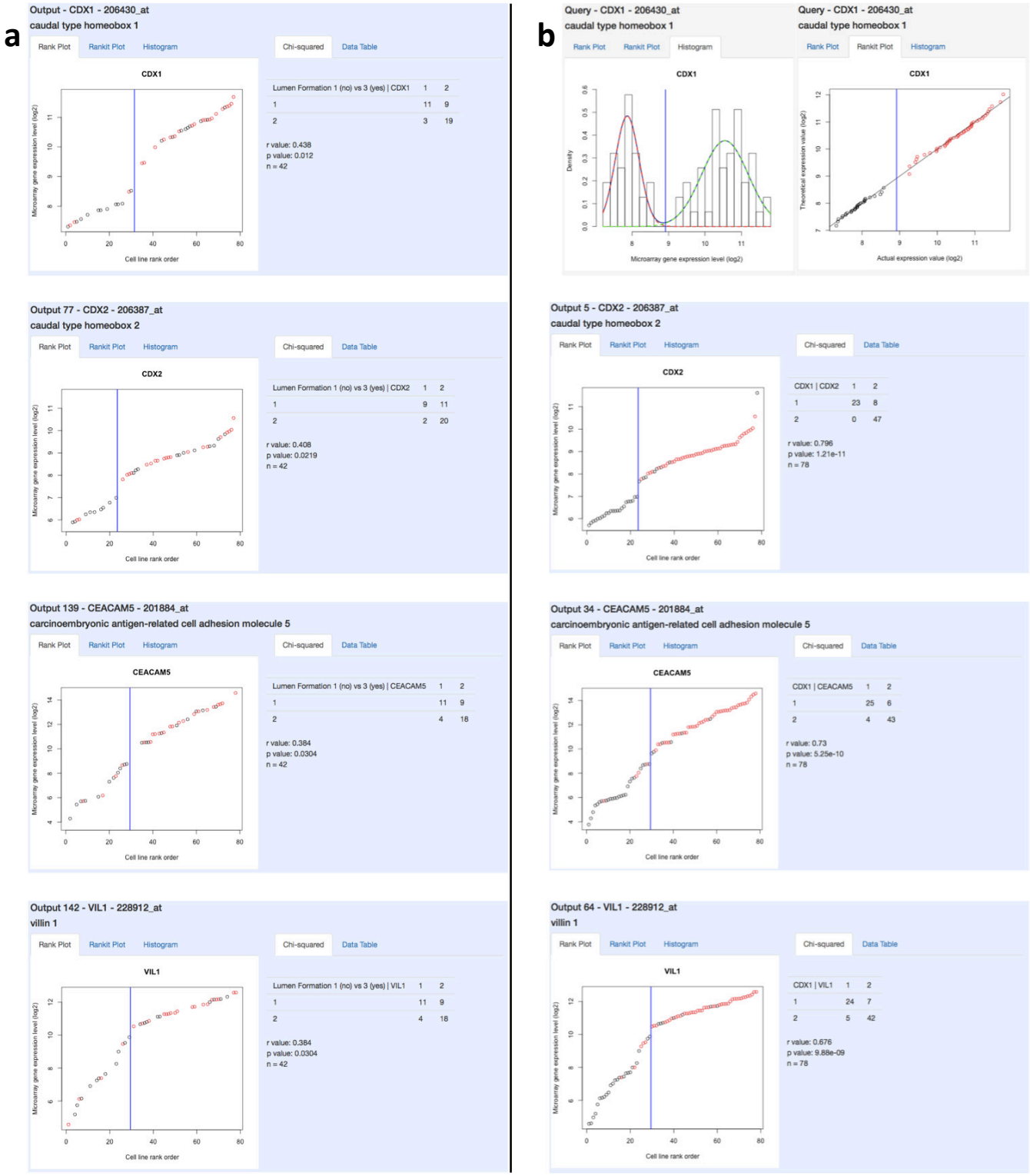


Figure 6.3: Differentiation in colorectal cancer cell lines. (**a**) Lumen formation in Matrigel is significantly positively associated with high expression of *CDX1*, *CDX2*, *CEACAM5*, and *VIL1*. Red points on the plot correspond to cell lines that form lumens. (**b**) *CDX1* expression is bimodal across the cell line panel, and well-fitted by two normal curves as evidenced by the near-linear rankit plot. *CDX2*, *CEACAM5*, and *VIL1* expression is strongly positively associated with *CDX1* expression. Red points on the plot correspond to cell lines with high *CDX1* expression. A table of the top 200 probesets most associated with lumen formation and *CDX1* expression can be found in the appendix.

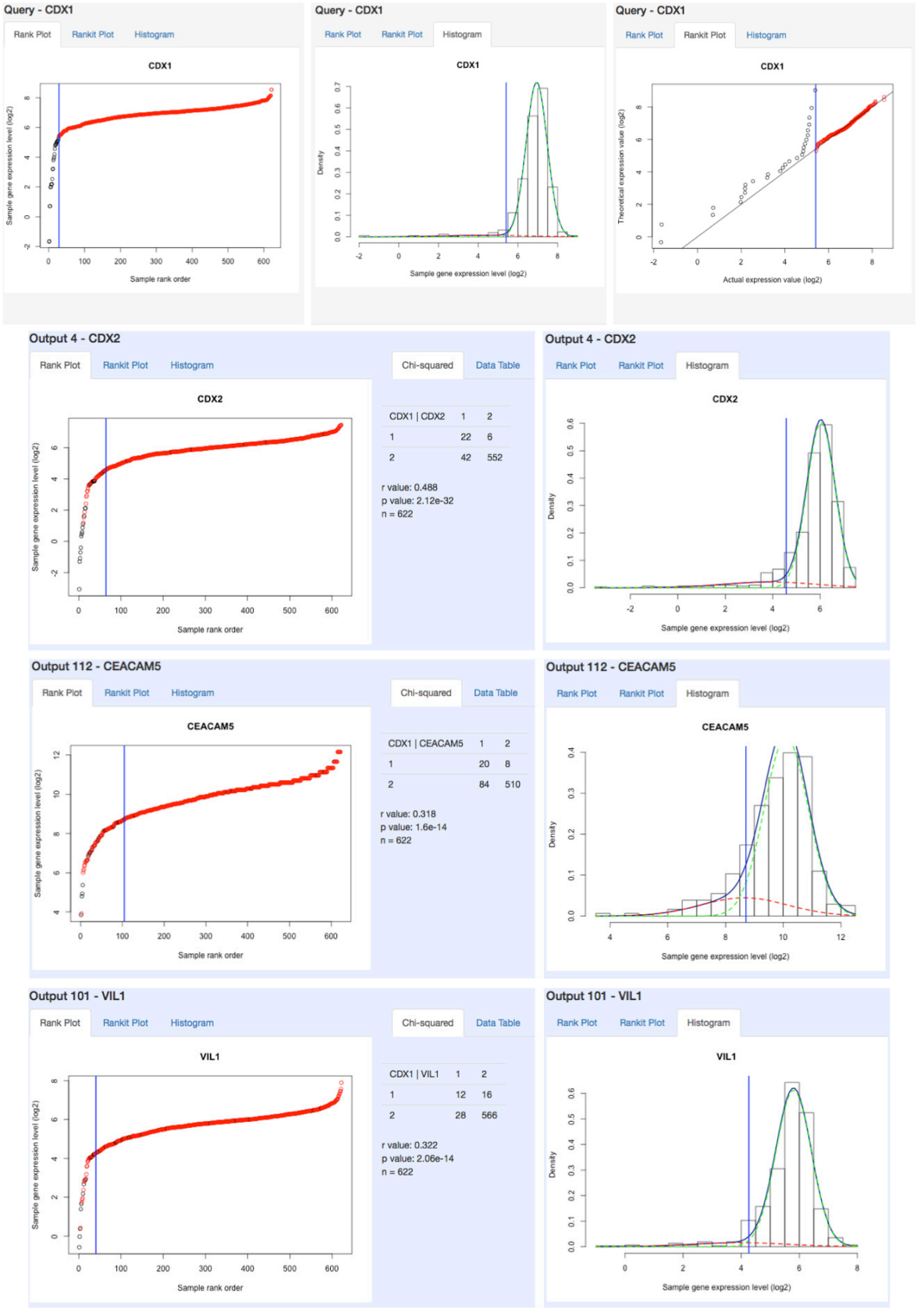


Figure 6.4: Differentiation in primary colorectal tumors. *CDX1* expression is bimodal across the tumor panel, with reasonable fit by two normal distributions as per the rankit plot. *CDX2*, *CEACAM5*, and *VIL1* expression is strongly positively associated with *CDX1* expression. Red points on the plot correspond to tumors with high *CDX1* expression. A table of the top 200 genes most associated with *CDX1* expression can be found in the appendix.

Two genes with potential implications to cancer interaction with stromal cells were found to be negatively associated with *CDX1* expression – *TGFB2*, encoding transforming growth factor-β2, and *PDGFC*, encoding platelet-derived growth factor-C, previously mentioned in Chapter 4 (**Figure 6.5a**). As seen in **Figure 4.4**, the cell lines incapable of recruiting myofibroblasts seemed to express higher levels of several known differentiation markers, including *CEACAM5*, suggesting that undifferentiated cells might be the ones to secrete the ligands attracting the mesenchymal cells. However, *TGFB2* expression was not significantly associated with *CDX1* in the primary tumors, and *PDGFC* was deemed to have a unimodal expression pattern with no apparent enrichment of *CDX1*-low samples at either the low or high ends of its expression (**Figure 6.5b**).

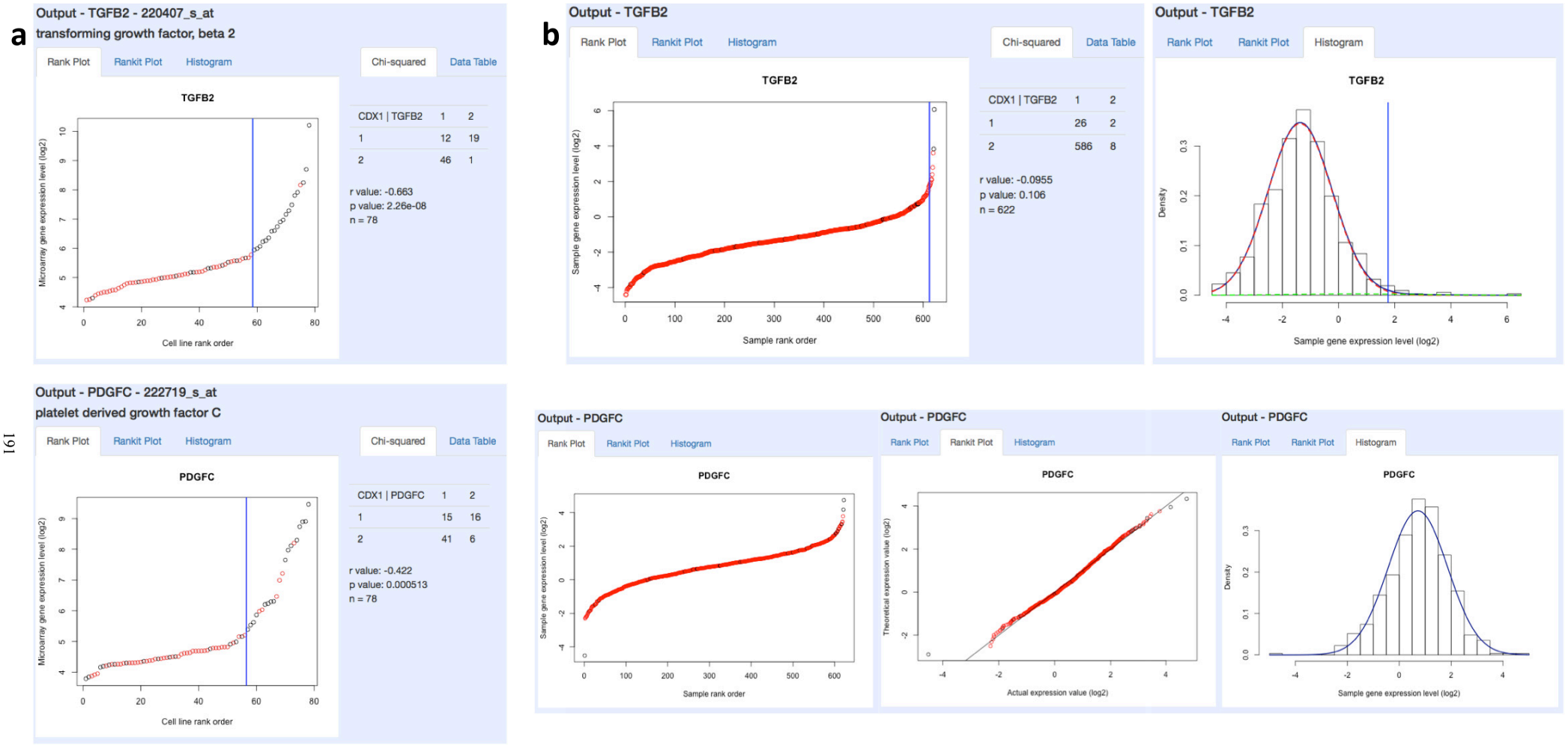


Figure 6.5: *TGFB2* and *PDGFC* in cell lines and primary tumors. (**a**) In the cell line panel, *TGFB2* and *PDGFC* expression is significantly negatively associated with *CDX1* expression, suggesting that these ligands may be predominantly secreted by undifferentiated cells. (**b**) However, no such association is found in the primary tumors, suggesting either that the expression of these genes is an artifact of culture or that the relationship between these genes and *CDX1* is obscured by stromal contamination of the primary tumor. Red points on the plot correspond to high *CDX1* expression in the cell lines or tumors.

6.3.2 Goblet Cell Formation

Goblet cells constitute one of the three differentiated epithelial cell populations in the colonic epithelium, and their primary function is the production of intestinal mucus. MUC2 is the primary mucin secreted by goblet cells, and is commonly used as a marker of these cells (Kim & Ho 2010). With the functionality and applicability of the 2x2 association approach to gene expression analysis supported by the findings regarding *CDX1*, the program was next applied to goblet cell formation. As expected, *MUC2* expression was strongly and significantly associated with the goblet cell formation phenotype in the colorectal cancer cell lines (**Figure 6.6a**). Well-known goblet cell markers *FCGBP* and *REG4* also strongly associated with goblet cell formation, as did *CDX1* (Pelaseyed et al. 2014; Sasaki et al. 2016). Querying the dataset directly for *MUC2* returned the same significant associations (**Figure 6.6b**). Notably, all three 2x2 tables for the associations with *MUC2* were strongly asymmetrical along the antidiagonal – *MUC2* was almost never expressed unless *FCGBP*, *REG4*, and *CDX1* were also present. This is in line with evidence suggesting that *MUC2* expression is a relatively late event in goblet cell maturation (McIntire et al. 2011). As before, these associations were confirmed in the primary tumor data, with a similar antidiagonal asymmetry for *CDX1* (**Figure 6.7**). 61.6% (383/622) primary tumors were deemed to have high *MUC2* expression by GMM, while only 19.2% (15/78) of cell lines did the same, again underscoring the underrepresentation of differentiating cancers *in vitro*.

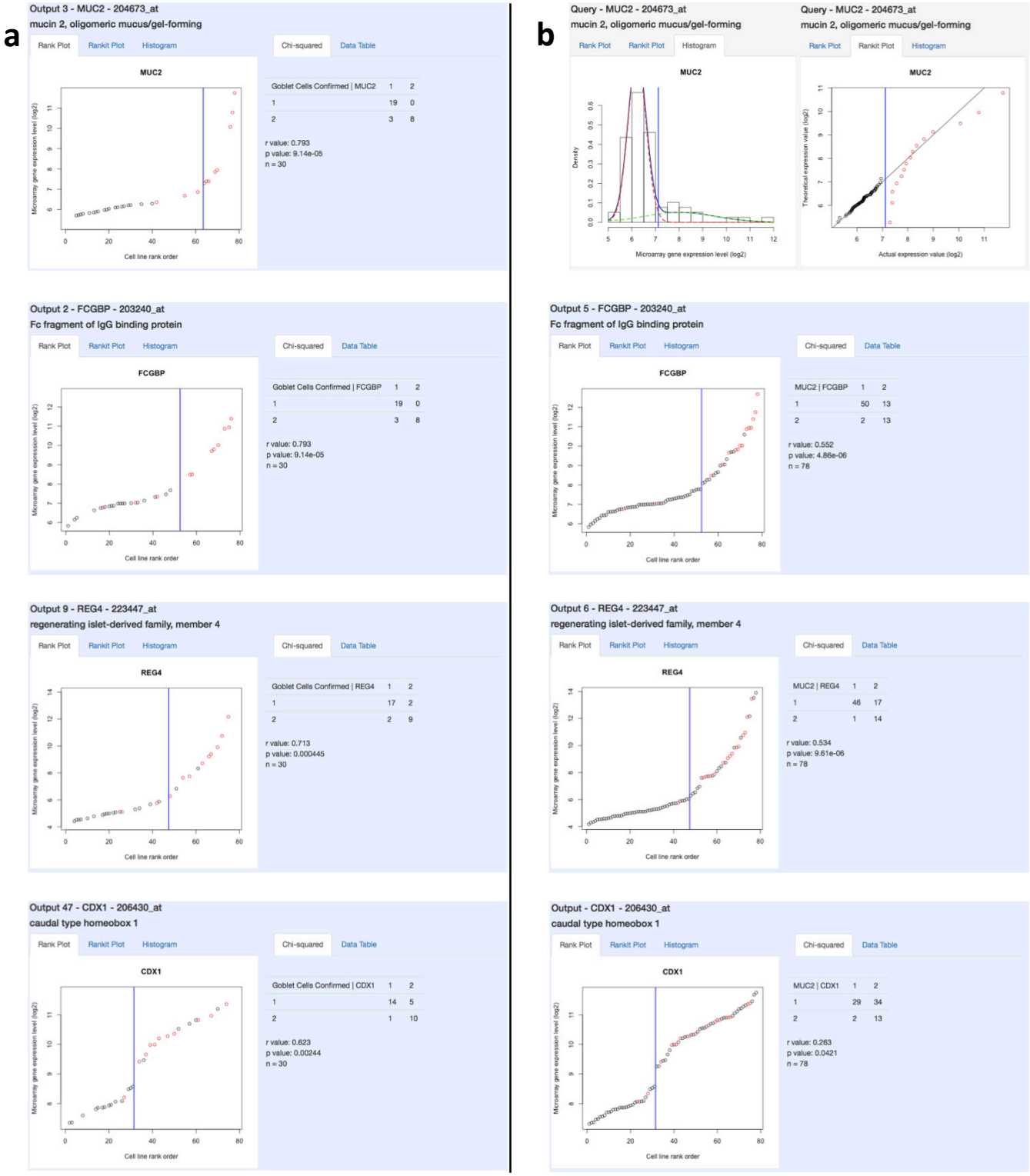


Figure 6.6: Goblet cell formation in colorectal cancer cell lines. (**a**) Goblet cell formation as assayed by MUC2 immunostaining is significantly positively associated with high expression of *MUC2*, *FCGBP*, *REG4*, and *CDX1*. Red points on the plot correspond to cell lines that form goblet cells. (**b**) *MUC2* expression is bimodal across the cell line panel, and reasonably fitted by two normal curves as per rankit plot. Notably, the low and high *MUC2* distributions overlap substantially. *FCGBP*, *REG4*, and *CDX1* expression is strongly positively associated with *MUC2* expression; almost no cell lines express *MUC2* in the absence of any one of these factors. Red points on the plot correspond to cell lines with high *MUC2* expression. A table of the top 200 probesets most associated with goblet cell formation and *MUC2* expression can be found in the appendix.

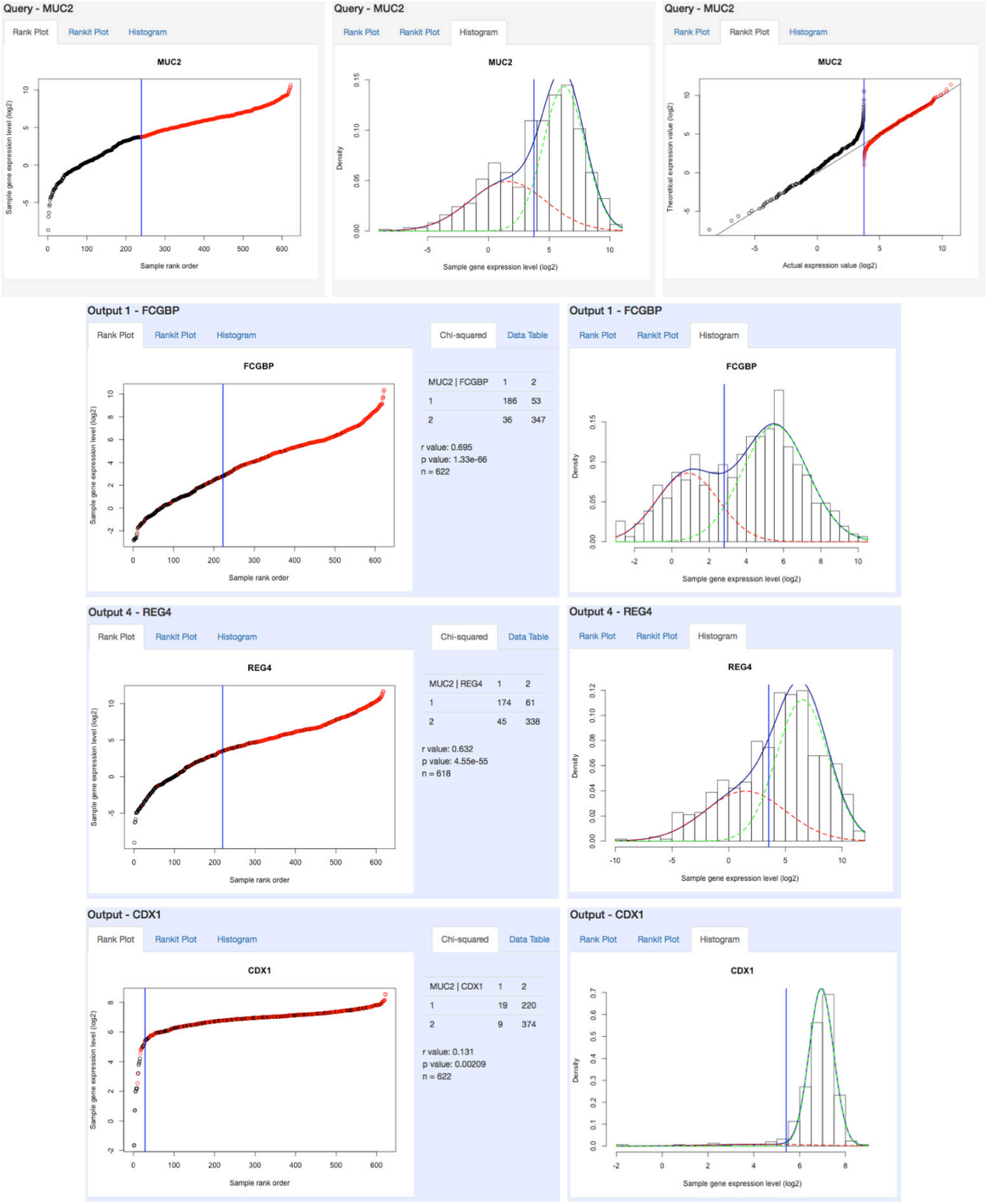


Figure 6.7: Goblet cell formation in primary colorectal tumors. *MUC2* expression is bimodal across the tumor panel, with reasonable fit by two normal distributions as per the rankit plot. *FCGBP*, *REG4*, and *CDX1* expression is strongly positively associated with *MUC2* expression. Red points on the plot correspond to tumors with high *MUC2* expression. A table of the top 200 genes most associated with *MUC2* expression can be found in the appendix.

The differentiation of the secretory epithelial cells in the intestine is thought to be governed by the action of two transcription factors, ATOH1 and SPDEF (Lo et al. 2017). Both of these were found to be significantly positively associated with *MUC2* expression, as well as with each other, in both the cell line and primary tumor datasets (**Figure 6.8**). However, none of these relationships are obviously direct. Most primary tumors with low *MUC2* express *ATOH1*, and a substantial number express *SPDEF*. On the other hand, the possibility that *ATOH1* and *SPDEF* must both be present for *MUC2* to be expressed is contradicted by the cell lines, since only 8 lines express high levels of both *ATOH1* and *SPDEF* while 15 express *MUC2*. Notably, unlike *CDX1* expression, which exhibits a very clear bimodal distribution across the cell lines, the expression for *MUC2* is less separated, with a large peak of *MUC2*-negative cell lines and a smaller overlapping distribution of *MUC2*-positive lines. Very few tumors containing goblet cells are made up exclusively of these cells; in fact, goblet cells are usually a minor population of the bulk colony, which also includes stem cells and other differentiated cells. Thus it is expected that bulk RNA collection for expression analysis substantially diluted the goblet cell-derived *MUC2* signal, resulting in the heavy-tailed distribution shown in **Figure 6.6**. In fact, some cell lines known to contain a subset of goblet cells from *MUC2* immunostaining are actually scored as negative by the GMM algorithm due to the dilution of the expression signal from the small subgroup of cells by the remaining non-goblet cells. The same effect also complicates the interpretation of the *ATOH1* and *SPDEF* associations.

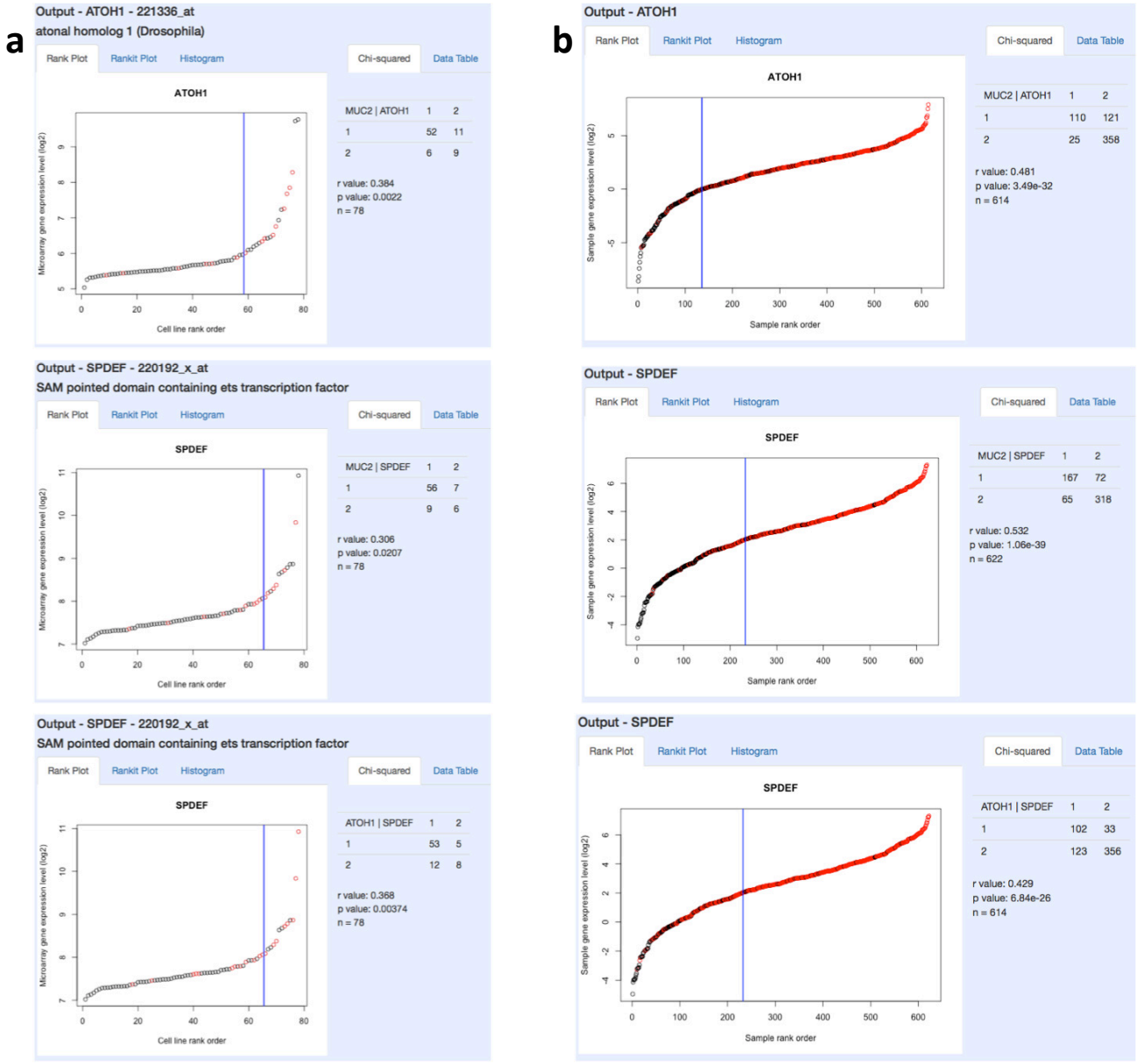


Figure 6.8: *ATOH1* and *SPDEF* in cell lines and primary tumors. (**a**) In the cell line panel, *ATOH1* and *SPDEF* significantly positively associate with *MUC2* and with each other. (**b**) The same is found in the primary tumors. However, neither *ATOH1* nor *SPDEF* completely control *MUC2* expression, as the protein is still expressed in samples in which either transcription factor is low in both cell lines and primary tumors.

6.3.3 Epithelial-Mesenchymal Transition

The epithelial-mesenchymal transition (EMT) is a process in which a polarized, nonmotile epithelial cell takes on stromal characteristics, beginning to express some mesenchymal proteins and losing adhesion to the other cells in the epithelial sheet (Serrano-Gomez et al. 2016). EMT is known to occur in normal development and wound healing, and has also been suggested to influence cancer invasiveness. The primary sign that a cell has undergone EMT is the loss of the expression of E-Cadherin (*CDH1*), a critical epithelial adhesion protein. When *CDH1* was analyzed for 2x2 associations in the cell lines and primary tumors, several consistent positive associations were found (**Figures 6.9-10**). Specifically, epithelial cell adhesion molecule (*EPCAM*), another adhesion protein, was found to be strongly downregulated whenever *CDH1* was. Metastasis-associated in colon cancer-1 (*MACC1*), a recently characterized protein implicated in the transcription of the *MET* receptor of hepatocyte growth factor, was also found to be positively associated with *CDH1*, as was *ESRP1*, previously implicated in the inhibition of RAC1B-mediated cell motility (Stein et al. 2009; Ishii et al. 2014; Singh et al. 2004).

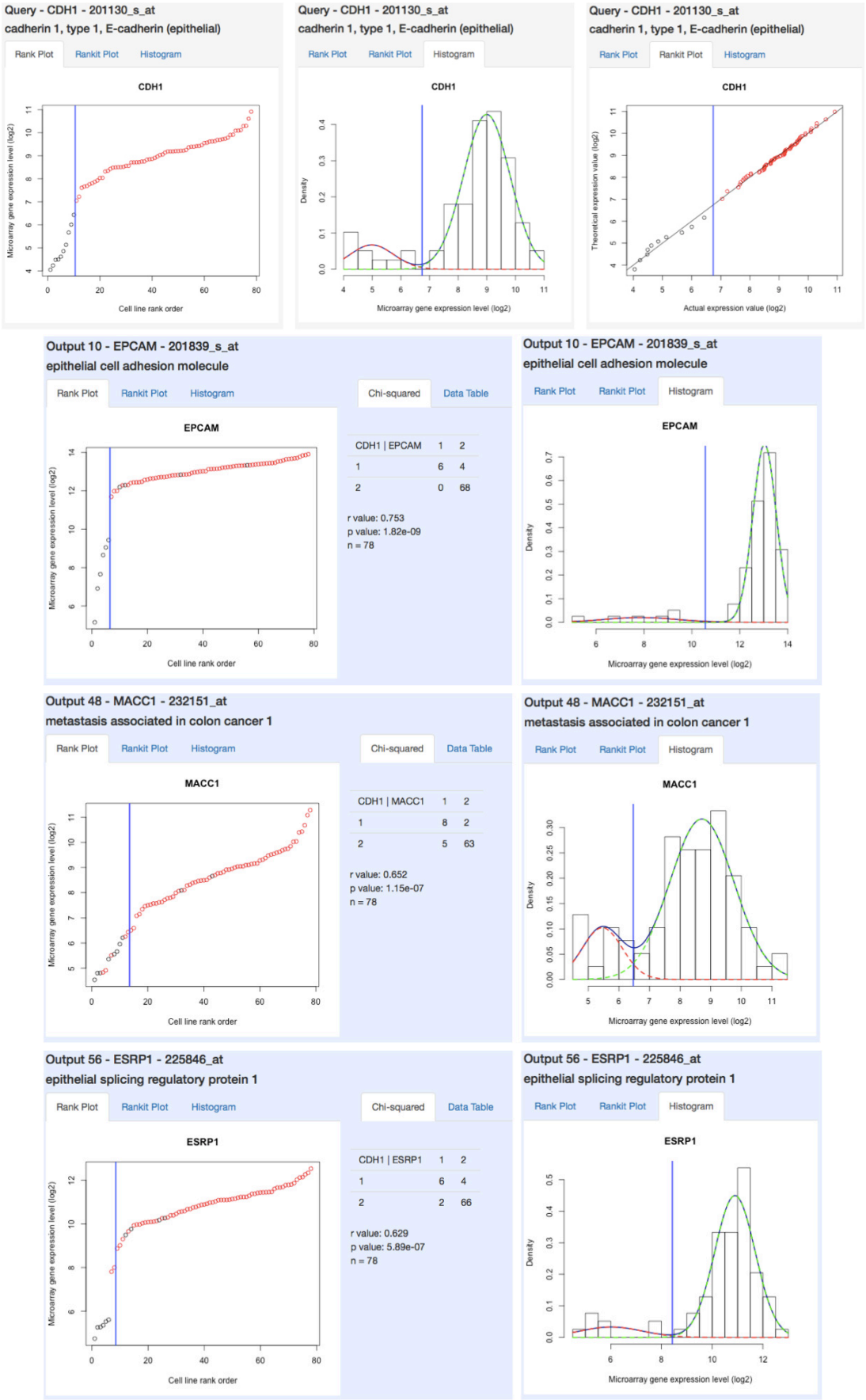


Figure 6.9: EMT in colorectal cancer cell lines. *CDH1* expression is bimodal across the cell line panel, and well-fitted by two normal curves as evidenced by the near-linear rankit plot. *EPCAM*, *MACC1*, and *ESRP1* expression is strongly positively associated with *CDH1* expression. Red points on the plot correspond to cell lines with high *CDH1* expression. A table of the top 200 probesets most associated with *CDH1* expression can be found in the appendix.

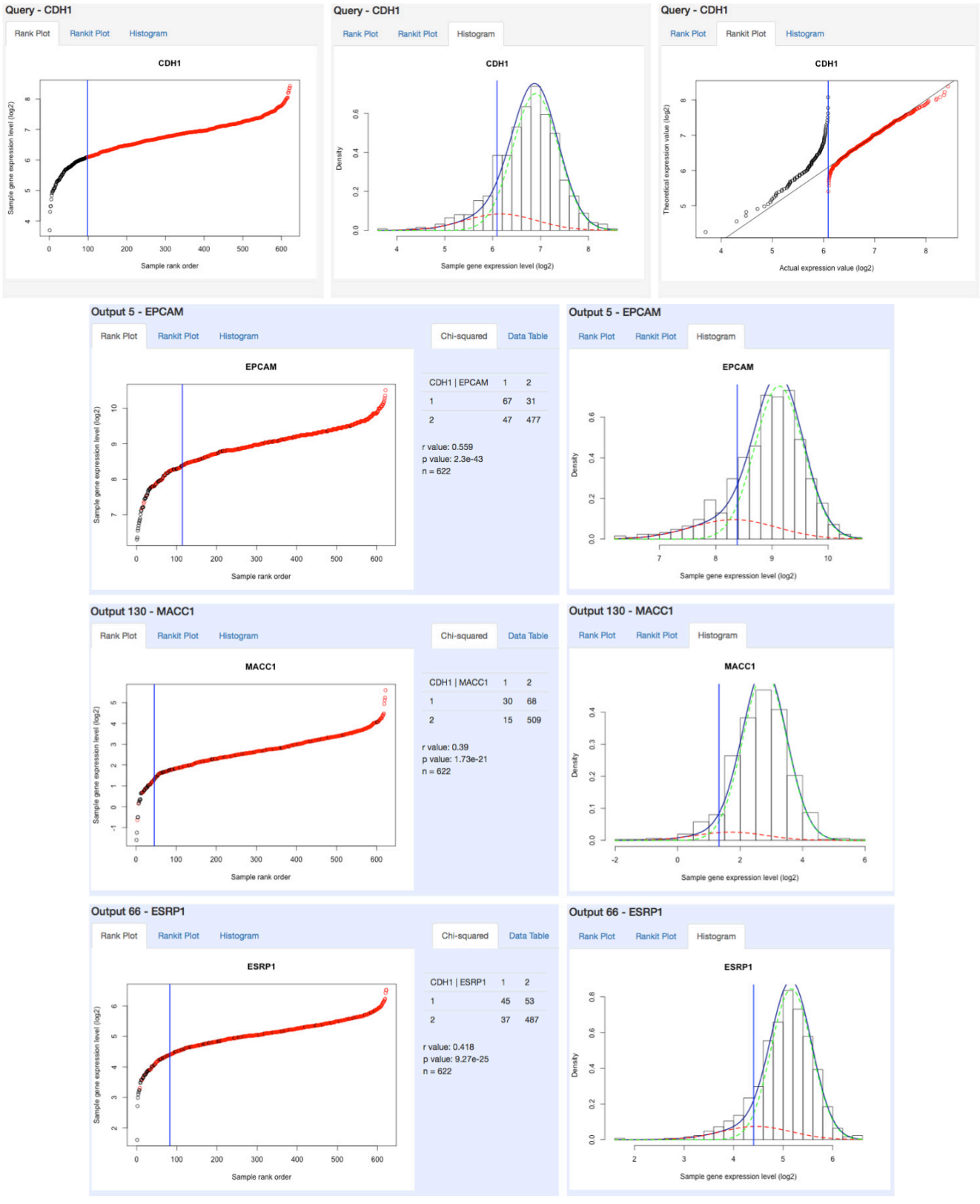


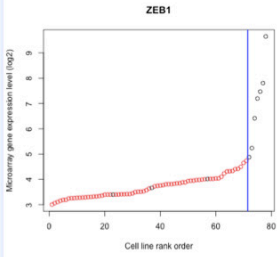
Figure 6.10: EMT in primary colorectal tumors. *CDH1* expression is bimodal across the tumor panel, and reasonably fitted by two normal curves as per rankit plot. *EPCAM*, *MACC1*, and *ESRP1* expression is strongly positively associated with *CDH1* expression. Red points on the plot correspond to tumors with high *CDH1* expression. A table of the top 200 genes most associated with *CDH1* expression can be found in the appendix.

More striking, however, was the set of genes associated with *CDH1* expression in the cell lines that were not confirmed to do this in the primary tumors (**Figure 6.11**). Specifically, *ZEB1*, a transcription factor long implicated in inducing EMT, was significantly negatively associated with *CDH1* in the cell lines, as was vimentin (*VIM*), an intermediate filament protein present in mesenchymal cells (Liu et al. 2008). On the other hand, *TWIST1*, another transcription factor reported to induce EMT, was not significantly associated with *CDH1*, and *SNAI1*, also known to stimulate EMT, exhibited a unimodal expression distribution across the cell lines. *ZEB1* and *VIM* had an almost completely associated pattern of expression. However, all of these genes had unimodal expression across the primary tumors, confirmed by almost perfectly linear rankit plots. Vimentin had an average expression level of around $2^{5.5}\approx45$ FPKM, corresponding to high expression (compare, for instance, to the expression pattern of *CDX2* in **Figure 6.4**). This suggests that most primary tumors contain vimentin, even though most of them do not exhibit EMT, meaning that non-cancerous vimentin-expressing cells, namely local stromal cells, substantially contribute to the gene expression patterns in the primary tumors (Guinney et al. 2015; Dunne et al. 2017). As epithelial cells undergo EMT and take on mesenchymal characteristics, they become less and less distinguishable from this contaminating stromal component. This explains why epithelial markers downregulated along with *CDH1* are detected as bimodal in both cell lines and primary tumors, while upregulated mesenchymal markers can only be detected *in vitro*, in the absence of a mesenchymal stroma that obscures the signal.

a

Output 3 - ZEB1 - 210875_s_at
zinc finger E-box binding homeobox 1

Rank Plot | Rankit Plot | Histogram | Chi-squared | Data Table

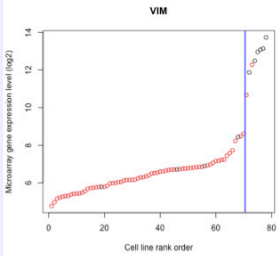


CDH1 | ZEB1	1	2
1	3	7
2	68	0

r value: -0.819
p value: 3.16e-11
n = 78

Output 50 - VIM - 201426_s_at
vimentin

Rank Plot | Rankit Plot | Histogram | Chi-squared | Data Table

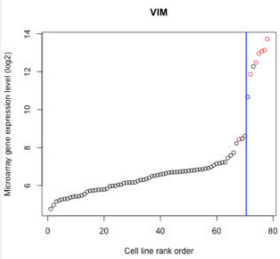


CDH1 | VIM	1	2
1	4	6
2	66	2

r value: -0.629
p value: 5.89e-07
n = 78

Output 43 - VIM - 201426_s_at
vimentin

Rank Plot | Rankit Plot | Histogram | Chi-squared | Data Table

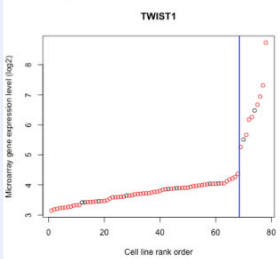


ZEB1 | VIM	1	2
1	69	2
2	1	6

r value: 0.781
p value: 4.26e-10
n = 78

Output - TWIST1 - 213943_at
twist homolog 1 (Drosophila)

Rank Plot | Rankit Plot | Histogram | Chi-squared | Data Table

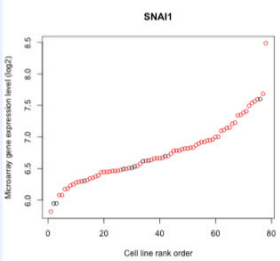


CDH1 | TWIST1	1	2
1	8	2
2	60	8

r value: -0.0824
p value: 0.825
n = 78

Output - SNAI1 - 219480_at
snail homolog 1 (Drosophila)

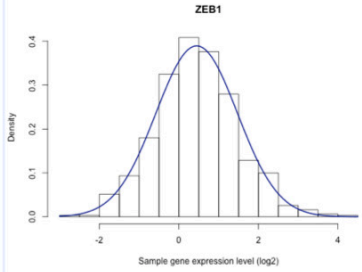
Rank Plot | Rankit Plot | Histogram



b

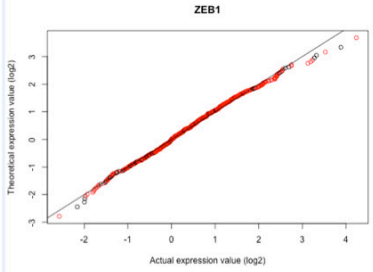
Output - ZEB1

Rank Plot | Rankit Plot | Histogram



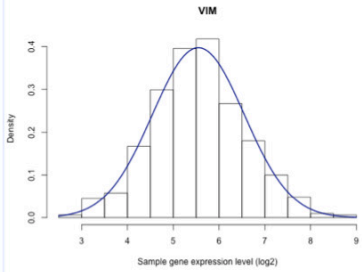
Output - ZEB1

Rank Plot | Rankit Plot | Histogram



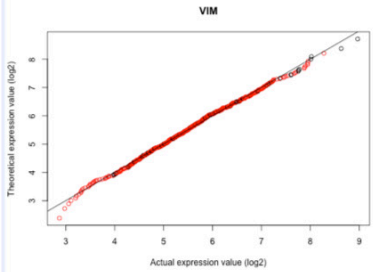
Output - VIM

Rank Plot | Rankit Plot | Histogram



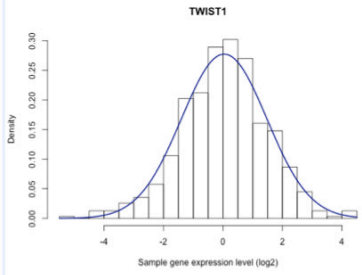
Output - VIM

Rank Plot | Rankit Plot | Histogram



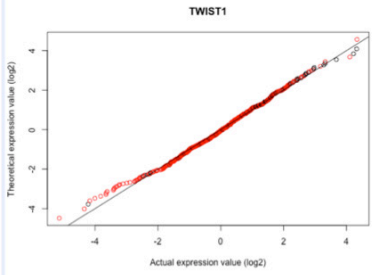
Output - TWIST1

Rank Plot | Rankit Plot | Histogram



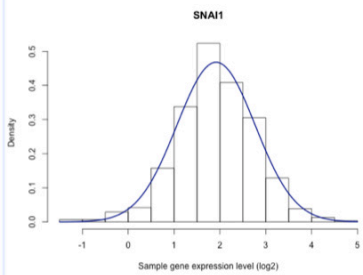
Output - TWIST1

Rank Plot | Rankit Plot | Histogram



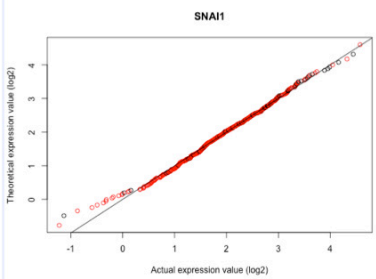
Output - SNAI1

Rank Plot | Rankit Plot | Histogram



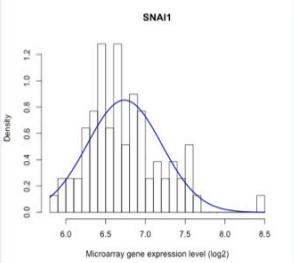
Output - SNAI1

Rank Plot | Rankit Plot | Histogram



Output - SNAI1 - 219480_at
snail homolog 1 (Drosophila)

Rank Plot | Rankit Plot | Histogram



Output - SNAI1 - 219480_at
snail homolog 1 (Drosophila)

Rank Plot | Rankit Plot | Histogram

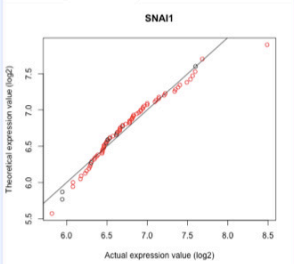


Figure 6.11 (previous page): EMT markers and transcription factors in colorectal cancer cell lines and primary tumors. (**a**) *ZEB1* expression is very strongly negatively associated with *CDH1*, and *VIM* follows almost exactly the same expression pattern as *ZEB1*. Other EMT-related transcription factors *TWIST1* and *SNAI1* show no association with *CDH1*. (**b**) On the other hand, none of *ZEB1*, *VIM*, *TWIST1*, or *SNAI1* is scored as bimodal in the primary tumors, and no relationship with *CDH1* can be seen for any of these genes. Red points on the plot correspond to high *CDH1* expression in the cell lines or tumors, with exception of the lower *VIM* plot in (**a**), in which the red points show cell lines with high *ZEB1* expression. For the cell lines, a table of the top 200 genes most associated with *ZEB1* expression can be found in the appendix.

6.3.4 Myofibroblast Contamination of Primary Tumors

As discussed previously in Chapter 4, and again in the previous section, stromal involvement is a fundamental characteristic of cancer. Myofibroblasts are critical components of the colorectal cancer stroma, and are known to accumulate near and in both primary and metastatic tumors (Yeung et al. 2013). As described in Chapter 1, the most recent definition of colorectal pericryptal myofibroblasts by marker expression is αSMA^+, $AOC3^+$, $NKX2.3^+$, $LRRC17^+$, $EDNRB^+$, VIM^+, $EPCAM^-$, $SHOX2^-$, $TBX5^-$, with NKX2-3 and AOC3 as the clearest positive markers (Hsia et al. 2016). As shown in **Figure 6.12**, *NKX2-3* expression was deemed bimodal across the primary tumors. Several proteins known to be expressed in myofibroblasts, including *PDGFRA* (Chapter 4) and *EDNRB*, were found to strongly positively associate with *NKX2-3*. Meanwhile, although no significant association was found between *NKX2-3* and *AOC3*, a striking asymmetry along the antidiagonal of their 2x2 table suggests that *AOC3* is almost never highly expressed in the absence of *NKX2-3*. Accordingly, a smaller fraction of primary tumors was deemed to have high *AOC3* expression (**Figure 6.13**). *AOC3* expression significantly positively associated with several proteins known to be involved in myofibroblast contractility, such as myosin light chain kinase (*MYLK*) and myosin-11 (*MYH11*), as well as secreted cytokine

CXCL12 (Chapter 3) (Hsia et al. 2016). This suggests that *AOC3* marks a distinct subset of *NKX2-3*-positive myofibroblasts, possibly with certain unique contractile and secretory activity.

α-smooth muscle actin (*ACTA2*), the classic myofibroblast marker, was not deemed bimodal across the primary cancers by the GMM algorithm. However, samples with higher *ACTA2* expression were visibly enriched for high *AOC3* expression (**Figure 6.14**). Further exploration of this sort found a similar pattern for *VIM*. It thus seems that although the expression of mesenchymal markers in the primary tumors is ostensibly unimodally normal, there is an underlying heterogeneity in the extent of stromal involvement and invasion in these tumors, or contamination of the samples with nontumor material. With both stromal components and EMT expected to affect the levels of epithelial and mesenchymal markers in the bulk expression analysis, it is not surprising that only the strongest and most specific markers, namely *CDH1* downregulation in EMT, as well as *AOC3* and *NKX2-3* appearance during myofibroblast invasion, exhibit enough expression asymmetry to be classified as bimodal. These observations further support the previously suspected caveat of stromal contamination of primary tumor gene expression analyses, and additional study is required to understand how this may affect the interpretation of such data.

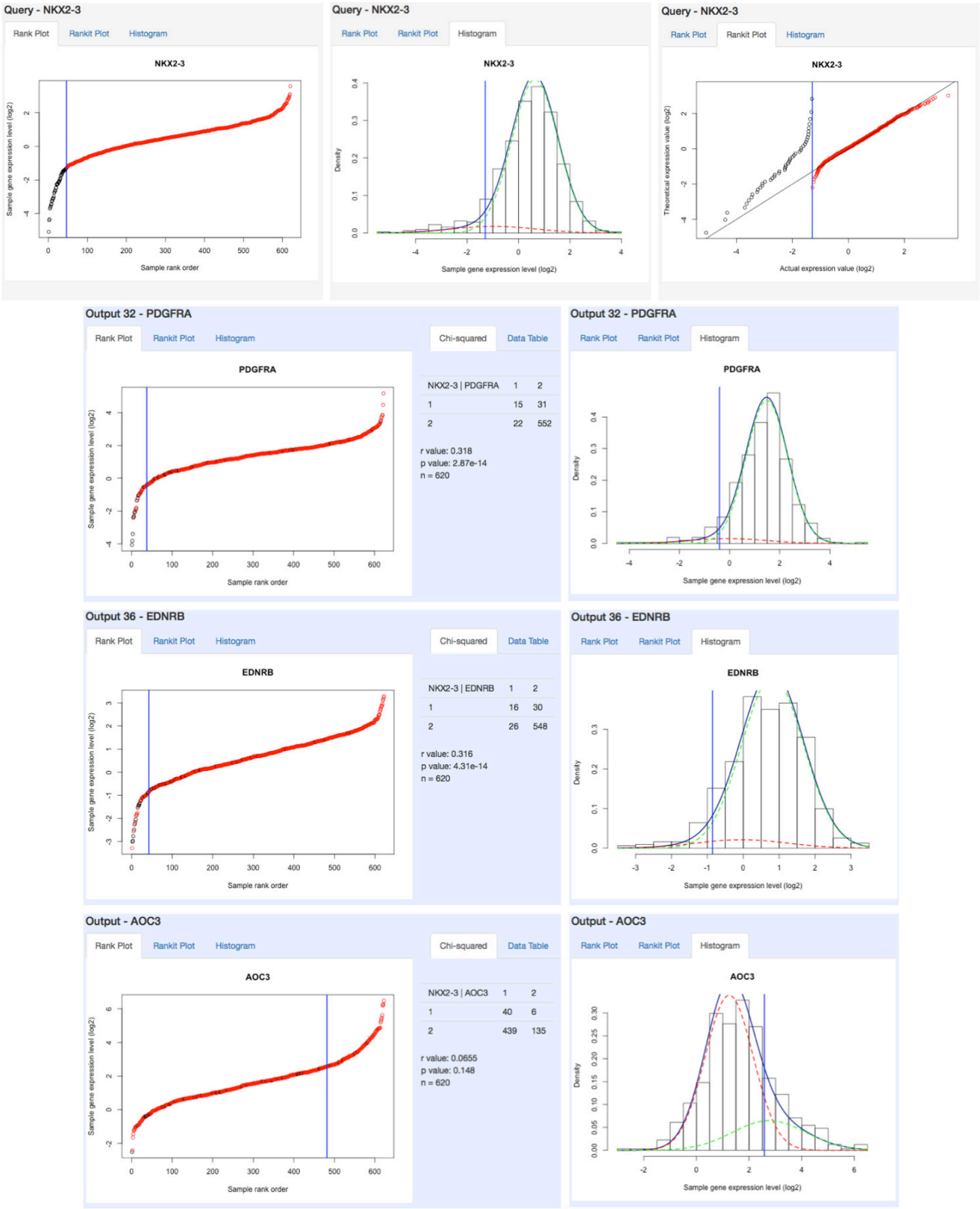


Figure 6.12: *NKX2-3* expression in the primary tumors. *NKX2-3* expression is bimodal across the tumor panel, and reasonably fitted by two normal curves as per rankit plot. *PDGFRA* and *EDNRB* expression is significantly positively associated with *NKX2-3*. While *AOC3* is not significantly associated with *NKX2-3*, it is an almost complete inclusion – *AOC3* is almost never expressed in the absence of *NKX2-3*. Red points on the plot correspond to tumors with high *NKX2-3* expression. A table of the top 200 genes most associated with *NKX2-3* expression can be found in the appendix.

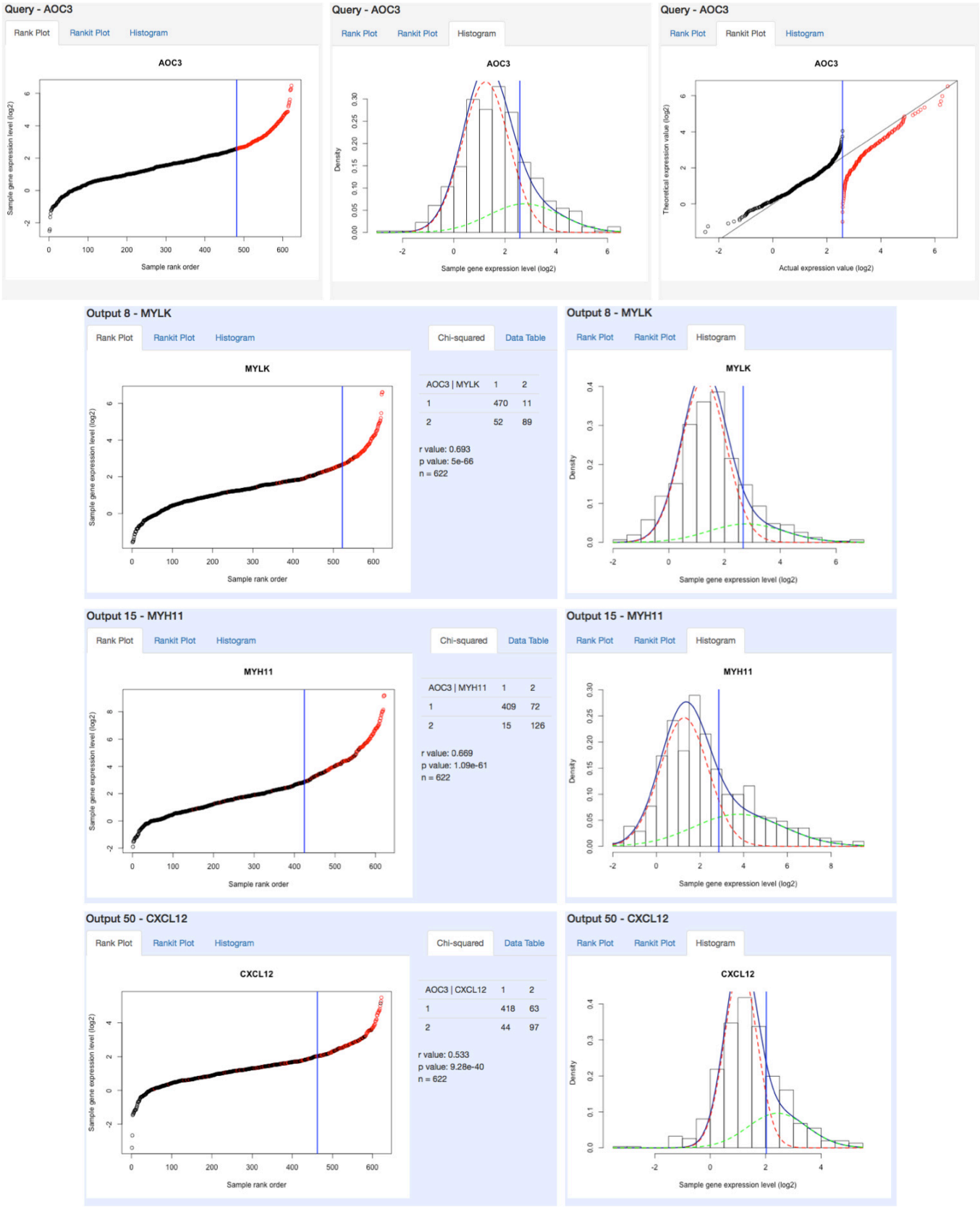


Figure 6.13: *AOC3* expression in the primary tumors. *AOC3* expression is bimodal across the tumor panel and reasonably fitted by two normal curves as per rankit plot. *MYLK*, *MYH11*, and *CXCL12* expression is significantly positively associated with *AOC3*. Red points on the plot correspond to tumors with high *AOC3* expression. A table of the top 200 genes most associated with *AOC3* expression can be found in the appendix.

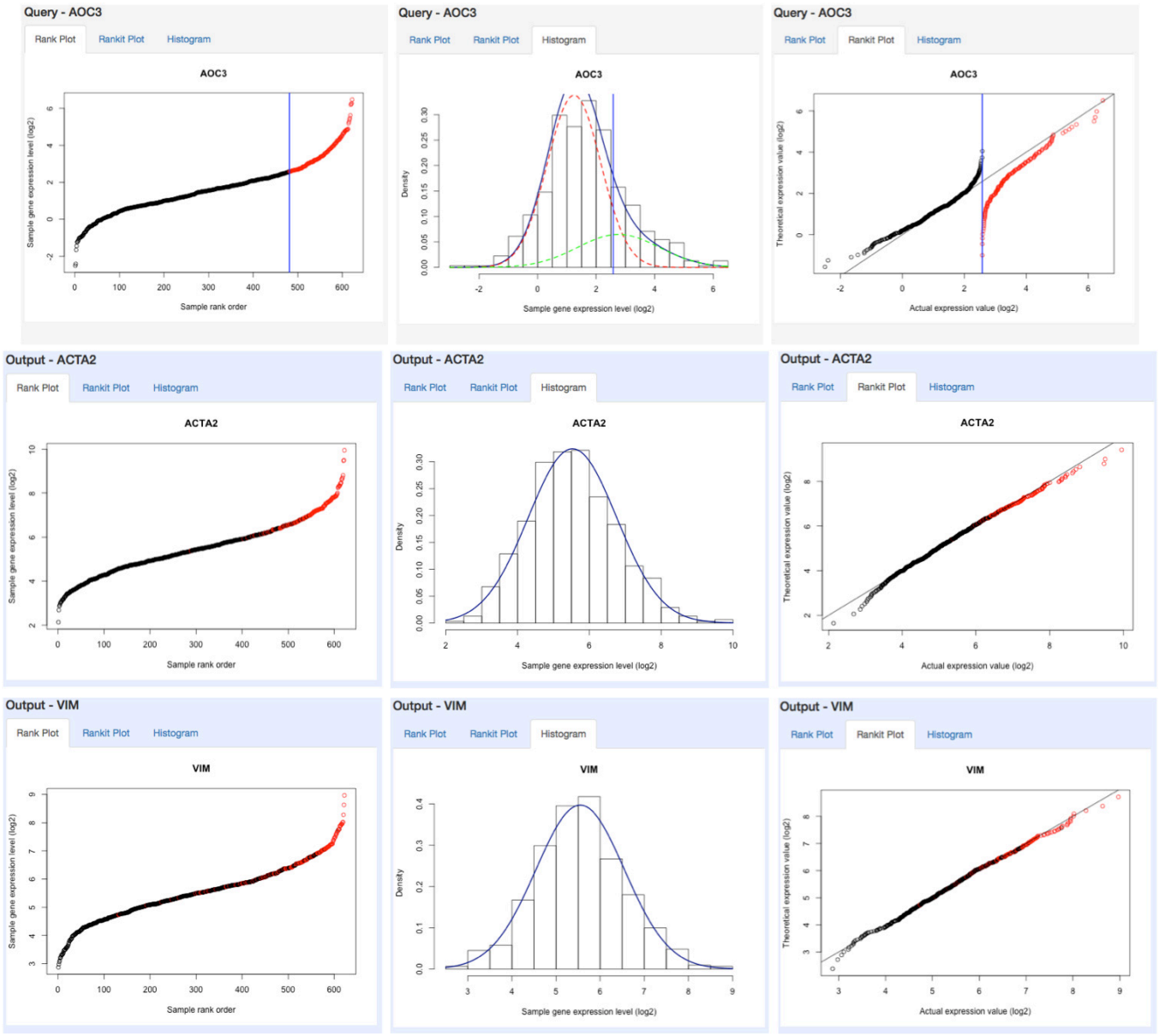


Figure 6.14: Other genes co-varying with *AOC3* in the primary tumors. While *ACTA2* and *VIM* are not scored as bimodal in the primary tumors by the GMM algorithm, tumors high for *AOC3* also tend to express higher levels of these genes. In all cases, red points on the plot correspond to tumors with high *AOC3* expression.

6.4 Discussion

The goal of this chapter was to test the ability of the GMM algorithm to adequately separate colorectal cancer cell lines and primary tumors into binary subgroups based on the expression patterns of individual genes, and to explore the associations between genes using categorical rather than continuous data. The advantages of this approach were twofold. Mapping the expression of every gene to a simple binary measure meant that equal attention was paid to genes with large and small expression ranges, thus ensuring that strong structural markers did not overshadow less variable but more consequential transcription factors. Additionally, technical and other noise is commonly assumed to be normally distributed, and modeling the expression patterns with normal curves also took some account of these factors. As shown in this chapter, a number of well-known gene regulatory relationships involved in the differentiation of the colonic epithelium, the formation and maturation of goblet cells, and epithelial-mesenchymal transition were successfully replicated in the categorized data, supporting the primary rationale of this analytical approach.

Colorectal cancers have long been subdivided based on their mutational and epigenetic characteristics, and recent work has also suggested a subgrouping based on genome-wide expression patterns (Vogelstein et al. 2013; Guinney et al. 2015). However, it has also long been known that expression analyses performed on bulk tumor suffer from contamination by noncancerous stroma. While more refined methods focusing only on genes thought to be exclusively present in cancer cells and minimally affected by such contamination have been applied with some success, the inherent noisiness of bulk tumor expression data remains a confounding factor

(Dunne et al. 2017). Expression analysis using GMM confirms these caveats. While a subset of colorectal cancer cell lines exhibited clear characteristics of EMT, with good definition of associated transcription factors and structural markers, these relationships were not obvious in primary tumors due to the expression of mesenchymal genes in invading/contaminating stromal cells. With some reports suggesting that the EMT process is itself affected by the tumor microenvironment, it becomes even less clear how any conclusions about it can be made from bulk tumor data (Mele et al. 2014).

On the other hand, the presence of mesenchyme-derived gene expression in the bulk tumor can also be advantageous, as it reports on the extent and quality of stromal involvement in the cancer. In this chapter, a clear subset of primary tumors was shown to express strong levels of *NKX2-3* and *AOC3*, encoding two well-characterized proteins known to specifically mark myofibroblasts (Hsia et al. 2016). Accordingly, the expression of these genes was found to associate or substantially co-vary with a number of other myofibroblast markers. Although it is unclear how much of this signal was due to myofibroblast invasion, and how much to contamination during the experimental procedure, this is a compelling observation that requires further study.

Studies of intratumoral heterogeneity have traditionally focused on genetically distinct subclones within the tumor as well as stromal contribution. Less attention has been paid to differentiation, the natural process by which more or less homogenous stem cells give rise to progeny with a wide spectrum of characteristics and activities (Marusyk et al. 2012). This is because normal differentiation is generally thought to

break down in cancers, with only a limited form of the original process remaining. However, this chapter shows that a vast majority (>95%) of primary colorectal tumors in the TCGA database express *CDX1*, the a key controller of enterocyte differentiation, and that over 60% of tumors form goblet cells as evidenced by high *MUC2* expression. This is despite the fact that usually only a fraction of cancer cells are goblet cells, if any, an observation confirmed by the *MUC2* gene expression pattern in colorectal cancer cells. Specifically, *MUC2* expression across the cell line panel takes the shape of a normal distribution with a heavy tail on the positive end, corresponding to a broad distribution of *MUC2*-positive cell lines overlapping with a narrower distribution of negative cell lines. With such clear effects present in clonal cell lines, it is evident that primary bulk tumor samples must suffer from the same problems.

The previous two chapters have proposed, tested, applied, and confirmed a novel approach to gene expression analysis based on Gaussian Mixture Modeling. This algorithm effectively detects heterogeneity in gene expression patterns across panels of colorectal cancer cell lines and primary tumors, and divides the samples into meaningful subsets. The applications shown in this chapter rely exclusively on the sample categorizations determined by this algorithm for every gene, and the direct associations between pairs of genes. Other avenues yet to be explored include using the gene pair association scores as the basis for expression clustering, and using the predicted subdistribution parameters (means and standard deviations) to more finely quantify expression heterogeneity and possibly improve the detection of transcripts from cell subsets akin to *MUC2*.

Chapter Seven

General Discussion

7.1 Myofibroblast-derived HGF Modulates Differentiation in a Subset of Colorectal Cancers

Hepatocyte growth factor (HGF) was originally discovered by two independent avenues – as scatter factor, a fibroblast-derived protein that enacted a breakdown of cell-cell contacts and dispersal in cultures of a range of epithelial cell types, and as HGF, a mitogen for hepatocytes isolated from platelets; subsequent structural studies determined that these are in fact identical factors (Stoker et al. 1987; Nakamura et al. 1986; Naldini et al. 1991). It was later found that plasma HGF levels increase in response to a range of injuries of different tissues, and inhibition of HGF activity reduces the rate of healing of injuries of the intestinal mucosa in mice (Nakamura & Mizuno 2010; Itoh et al. 2004). In colorectal cancers, Vermeulen *et al*. report that HGF derived from myofibroblasts in the tumor microenvironment helps maintain cancer stem cells in their undifferentiated state, and possibly even reverts differentiated cells back to a stem-like phenotype (Vermeulen et al. 2010). The results of Chapter 3 confirm that HGF derived from pericryptal colonic myofibroblasts negatively influences the differentiation of colorectal cancer cells, although this only occurs in a subset of cancers.

In this thesis, *HGF* was identified as a secreted factor with strongly increased expression in colonic myofibroblasts compared to skin fibroblasts (**Table 3.2**). Subsequent blocking of HGF signaling in cocultures of myofibroblasts and colorectal cancer cells in Matrigel using both PHA665752, an inhibitor of phosphorylation of HGF receptor c-Met, and siRNA-mediated knockdown of *HGF* in myofibroblasts, resulted in morphological changes in the colonies of one colorectal cancer cell line

indicating increased differentiation compared to untreated coculture (**Figures 3.7-11**). While c-Met is highly expressed in every colorectal cancer cell line of over 100 held by the lab (data not shown), four other colorectal cancer cell lines exhibited no apparent changes in colony morphology upon HGF blocking, forming luminal structures indicative of differentiation under all conditions. No appreciable mitogenic effect of HGF on colorectal cancer cells was determined.

The most recent models of cancer stem cells in colorectal cancer identify WNT signaling as a primary determinant of this phenotype (de Sousa E Melo & Vermeulen 2016). This model is especially attractive as high WNT activity reportedly defines stem cells in normal colon, and activating mutations in the WNT pathway are by far the most common initiators of colorectal tumors. HGF signaling has been implicated in regulating the WNT pathway in a range of tissues, both by inducing β-catenin translocation into the nucleus and by directly activating production of its target transcription factors (Monga et al. 2002). Accordingly, the mechanism proposed by Vermeulen *et al*. is that myofibroblast-derived HGF induces nuclear accumulation of β-catenin in colorectal cancer cells and thereby maintains their stemness (Vermeulen et al. 2010). However, this study worked mainly with primary cultures of human colorectal tumors, and it is unclear how many patients' samples were used. In the small panel of five cell lines tested in Chapter 3, only one responded to HGF signaling inhibition in coculture with myofibroblasts. This seems to indicate that there may be a subpopulation of colorectal tumors, missed by the Vermeulen study due to limited patient material, which is insensitive to HGF and differentiates in the same way regardless of the presence of that ligand, maintaining its cancer stem cell pool in a microenvironment-independent, or at least HGF-independent manner. Future

expansion on the findings of Chapter 3 should begin by testing the effects of HGF inhibition on colony morphology in coculture with myofibroblasts in a substantially larger panel of colorectal cancer cell lines. Next, gene expression comparisons between HGF responder and non-responder cell lines should be performed to determine what characteristics predict a cell line's responsiveness to HGF. If confirmed in a large panel of primary tumor-derived cultures, these genetic correlates could form the basis for patient stratification according to likelihood of response to HGF-targeted therapy.

7.2 PDGF Ligands from Colorectal Cancers Activate and Recruit Local Myofibroblasts

Transforming growth factor-β (TGF-β) and platelet-derived growth factor (PDGF) ligands are reported as the two primary signals responsible for intestinal myofibroblast activation in a range of contexts (Powell et al. 1999). Ligands of both of these families are known to be expressed by a subset of colorectal cancer cell lines (data not shown), and are reportedly secreted by primary colorectal tumors (Calon et al. 2012; Sundberg et al. 1997). Both primary colorectal cancers and their lymph node metastases are known to activate myofibroblasts in their environments and recruit them into the tumor bulk (Tsujino et al. 2007; Yeung et al. 2013). Accordingly, TGF-β and PDGF signals have long been the primary candidates suspected to mediate these interactions, although limitations in available *in vitro* assays have made it difficult to thoroughly characterize the full set of colorectal cancer-myofibroblast communications. In Chapter 4, PDGF-AA and PDGF-CC are identified as two

ligands possibly responsible, in a redundant fashion, for the induction of myofibroblast invasion into colorectal tumors.

To begin with, an assay was devised to mimic the spatial arrangement of colorectal cancer cells and myofibroblasts within an extracellular matrix (ECM), closely proximal to allow meaningful exchange of both soluble and potentially membrane-bound signals (**Figure 4.1**). This assay was used to test the ability of 43 colorectal cancer cell lines to recruit CCD-18Co colonic myofibroblasts, and *PDGFC* was found to be a secreted protein-coding gene with increased expression in the cell lines that recruited the myofibroblasts (**Figure 4.4**). Subsequent testing with imatinib, an inhibitor of the phosphorylation of the platelet-derived growth factor receptor, confirmed that this pathway is involved in the recruitment phenotype (**Figure 4.5**). More thorough genetic analysis implicated *PDGFA* and *PDGFC* as possibly interchangeably acting ligands that bring about this process (**Figure 4.6**).

Cancers in general, and colorectal cancers in particular, modulate the stromal cellular content in both primary and metastatic sites by secreting a range of cytokines (Conti & Thomas 2011; Gout & Huot 2008). PDGF receptors are widely expressed in stromal cells of the colorectum, as well as in the liver, lung, and lymph nodes, the most common sites of colorectal cancer metastasis; according to co-staining for α-smooth muscle actin, these cells are local myofibroblasts, although more thorough characterization would be required to verify this (Kitadai et al. 2006; Tejada et al. 2006; Chai et al. 2013). Incidentally, each of four myofibroblast cell lines with available in-house gene expression profiles expresses the PDGF receptors (**Figure 4.6c**). Several studies report that blocking PDGF receptor signaling using imatinib

inhibits growth and metastasis of colorectal cancers by reducing stromal involvement, and PDGF-CC has specifically been implicated as a cancer cell-derived signal that stimulates local stroma and promotes metastatic colonization in liver (Yuge et al. 2015; Stahtea et al. 2007; Bandapalli et al. 2012). The findings in this thesis seem to confirm the role of PDGF signaling in myofibroblast activation and recruitment by colorectal cancer cells. Notably, in an in-house panel of over 100 colorectal cancer cell lines, only *PDGFA* and *PDGFC* showed substantial expression in subsets of cell lines, suggesting that PDGF-BB, which has previously been suggested to modulate tumor stroma in a range of cancer types including colorectal, is derived from stromal or endothelial cells rather than cancer cells (Hosaka et al. 2016; Nakamura et al. 2008). As shown in Chapter 4, expression of either *PDGFA* or *PDGFC* significantly associated with the ability of a cancer cell line to recruit myofibroblasts, a characteristic abolished by the addition of imatinib. Further study is required to fully confirm the specificity of this interaction. First of all, genetic downregulation of *PDGFA* and *PDGFC* using siRNA or CRISPR knockout should be performed in the colorectal cancer cells to show that these genes are indeed responsible for the factors stimulating the myofibroblasts; alternatively, PDGF receptors could be similarly removed in the myofibroblasts. Additionally, a verification set of cell lines not tested in the original screen should be selected based on *PDGFA* and *PDGFC* expression, and the assay should be repeated to check that the expression of one of these genes indeed predicts myofibroblast recruitment. Finally, these tests should be repeated using a larger set of myofibroblast cell lines to ensure that the effect is not specific to the CCD-18Co line. If confirmed, this would constitute strong mechanistic evidence of a central role of PDGF ligands in myofibroblast activation and recruitment by colorectal cancers.

7.3 Gaussian Mixture Modeling Extracts Meaningful Heterogeneity in Colorectal Cancer Expression Data

Statistical analysis of continuous genome-wide gene expression data has become a staple technique across biological disciplines, and is expected to soon find clinical applications. When comparing sets of samples with variable phenotypic characteristics, it is informative to pay special attention to the genes whose expression is associated with the presence or absence of the phenotype in an effort to explain it mechanistically. However, it is often difficult to establish clear-cut definitions for certain phenotypes, and in such cases it would be beneficial to extract the nature of the heterogeneity within a set of samples in an unsupervised manner. In the specific case of gene expression analysis across a set of cancers from a certain tissue, it is desirable to uncover the genes exhibiting clear and well-separated bi- or multi-modal distributions across the samples, as these are most likely to report on strong genetic changes that were subsequently acted upon by natural selection. Such distributions would be expected to delineate the most impactful mutations defining the evolutionary pathways of the tumors. In Chapters 5 and 6, a method to detect such distributions is established and applied to sets of colorectal cancer cell lines and primary tumors.

The analytical approach is based on Gaussian Mixture Modeling (GMM), a process in which the data is modeled as one or several normal distributions (**Figure 5.1**). Analyses of simulated data found that $\log_2$-transformed expression values result in more reliable results as this allows for improved detection of sample subsets with low expression for the gene at hand. When applied to actual expression data from

colorectal cancer cell lines and primary tumors, derived from both expression microarrays and RNA-Seq, the algorithm detected known functionally relevant heterogeneity across the samples. In Chapter 6, association analyses were performed on the categorical data derived from GMM using 2x2 table statistics, uncovering previously reported associations between genes and also suggesting various novel relationships. Importantly, these analyses also detected myofibroblast-derived expression patterns in a subset of primary tumors, confirming that bulk tumor samples are frequently contaminated with stromal cells.

The primary aim of the GMM approach was to uncover intertumoral heterogeneity in an effort to understand what genetic changes distinguish tumors that have followed different evolutionary pathways, or advanced along these pathways to different extents. As discussed in Chapter 6, both the cell lines and the primary tumors showed stratification by extent of differentiation marker expression, goblet cell formation, and evidence of epithelial-mesenchymal transition (EMT). The proportions of the samples exhibiting various states of these characteristics varied between the cell lines and the primary tumors; for instance, while almost 40% of cell lines exhibited low expression of *CDX1*, the transcription factor controlling enterocyte differentiation, less than 5% of primary tumors did the same. However, matching subsets were represented in both datasets, suggesting that while cell lines may not represent the proportions of different tumor types seen in patients, they nevertheless exhibit the full range of tumor characteristics. Additionally, tumor-stromal interactions not present in culture may alter the expression patterns in the cancer cells, further skewing the comparison between primary tumors and cell lines. Another finding from these analyses concerns intratumoral heterogeneity, relating to the different subsets of cells within a single

tumor. Most notably, myofibroblast markers and genetic associations were found in a large subset of primary colorectal tumors, suggesting that stromal components are routinely included in gene expression analyses of bulk tumors. While this is an undesirable and confounding factor on its surface, it may also be useful as it could report on the tumors that are more able to recruit local myofibroblasts into their bulk. Intratumoral heterogeneity also appears as a consequence of differentiation, as seen most clearly in the analysis of *MUC2* in the cell lines (**Figure 6.6**). Clonal cell lines gave rise to heterogenous cell populations, with usually only a fraction of them, if any, differentiating into goblet cells. This gave the characteristic "heavy-tailed" bimodal distribution of *MUC2* expression across the set of cell lines. Future work should focus on further characterizing such "heavy-tailed" distributions, which could be used to infer that certain genes are only present in subsets of cells as with *MUC2*. It would also be beneficial to use the association statistics between genes as a distance measure for clustering, and to compare the resulting gene groupings with those determined through methods for continuous data; presumably, the removal of Gaussian noise and the projection of ranges of expression onto a binary measure should result in more clear-cut associations and gene clusters. Finally, further study on the stromal component in the primary tumor expression data should be performed to properly extract the exact level and nature of the contamination, and to understand how it affects the interpretation of the cancer cell-derived signal.

7.4 Summary

This thesis focused on the bidirectional communication between colorectal cancer cells and myofibroblasts, and on the effects that these closely proximal cell types have on each other in the tumor microenvironment. Using various *in vitro* assays, it has been shown that HGF is a key signaling molecule derived from myofibroblasts that modulates the differentiation of colorectal cancer cells, while PDGF ligands seem to be the primary cancer-derived effectors of myofibroblast invasion into the tumor bulk. Additionally, a new approach to unsupervised detection of heterogeneity in gene expression profiles across samples using Gaussian Mixture Modeling has been devised and verified in this thesis, confirming well-known colorectal cancer subsets and hinting at other underexplored characteristics. It is hoped that these studies can form the groundwork for an improved understanding of the colorectal cancer microenvironment and gene expression spectrum, and eventually result in the development of novel ways to manage and treat this prevalent and deadly disease.

References

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.

Anderberg, C. et al., 2008. Paracrine Signaling by Platelet-Derived Growth Factor-CC Promotes Tumor Growth by Recruitment of Cancer-Associated Fibroblasts. *Cancer Research*, 69(1).

Andoh, A. et al., 2002. Interleukin (IL)-4 and IL-17 synergistically stimulate IL-6 secretion in human colonic myofibroblasts. *International Journal of Molecular Medicine*, 10(5), pp.631–634.

Andoh, A. et al., 2005. Interleukin-22, a member of the IL-10 subfamily, induces inflammatory responses in colonic subepithelial myofibroblasts. *Gastroenterology*, 129(3), pp.969–984.

Armaka, M. et al., 2008. Mesenchymal cell targeting by TNF as a common pathogenic principle in chronic inflammatory joint and intestinal diseases. *The Journal of Experimental Medicine*, 205(2), pp.331–337.

Ashley, N., Yeung, T.M. & 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.

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.

Bach, S.P., 2000. Stem cells: the intestinal stem cell as a paradigm. *Carcinogenesis*, 21(3), pp.469–476.

Bamba, S. et al., 2003. Regulation of IL-11 expression in intestinal myofibroblasts: role of c-Jun AP-1- and MAPK-dependent pathways. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, 285(3), pp.G529–G538.

Bandapalli, O.R. et al., 2012. Paracrine signalling in colorectal liver metastases involving tumor cell-derived PDGF-C and hepatic stellate cell-derived PAK-2. *Clinical and Experimental Metastasis*, 29(5), pp.409–417.

Barker, N. et al., 2009. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature*, 457(7229), pp.608–11.

Barker, N. et al., 2007. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*, 449(7165), pp.1003–1007.

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.

Baum, J. & Duffy, H.S., 2011. Fibroblasts and myofibroblasts: what are we talking about? *Journal of Cardiovascular Pharmacology*, 57(4), pp.376–9.

Bejcek, B.E. et al., 1992. The v-sis oncogene product but not platelet-derived growth factor (PDGF) a homodimers activate PDGF alpha and beta receptors intracellularly and initiate cellular transformation. *Journal of Biological Chemistry*, 267(5), pp.3289–3293.

Bodmer, W.F. et al., 1987. Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature*, 328(6131), pp.614–6.

Boland, C.R. & Goel, A., 2010. Microsatellite Instability in Colorectal Cancer. *Gastroenterology*, 138(6), p.2073–2087.e3.

Bracht, K. et al., 2010. 5-Fluorouracil response in a large panel of colorectal cancer cell lines is associated with mismatch repair deficiency. *British Journal of Cancer*, 103(3), pp.340–346.
Bradley, C.A. et al., 2016. Transcriptional upregulation of c-MET is associated with invasion and tumor budding in colorectal cancer. *Oncotarget*, 7(48), pp.78932–78945.
Brand, S. et al., 2005. CXCR4 and CXCL12 are inversely expressed in colorectal cancer cells and modulate cancer cell migration, invasion and MMP-9 activation. *Experimental Cell Research*, 310(1), pp.117–130.
Brookes, M.J. et al., 2006. Modulation of iron transport proteins in human colorectal carcinogenesis. *Gut*, 55(10), pp.1449–1460.
Burke, J.P. et al., 2010. Bacterial lipopolysaccharide promotes profibrotic activation of intestinal fibroblasts. *British Journal of Surgery*, 97(7), pp.1126–1134.
Bussolino, F. et al., 1992. Hepatocyte growth factor is a potent angiogenic factor which stimulates endothelial cell motility and growth. *The Journal of Cell Biology*, 119(3), pp.629–641.
Calon, A. et al., 2012. Dependency of Colorectal Cancer on a TGF-β-Driven Program in Stromal Cells for Metastasis Initiation. *Cancer Cell*, 22(5), pp.571–584.
Cao, R. et al., 2002. Angiogenesis stimulated by PDGF-CC, a novel member in the PDGF family, involves activation of PDGFR-alphaalpha and -alphabeta receptors. *FASEB J*, 16(12), pp.1575–1583.
Chai, Q. et al., 2013. Maturation of Lymph Node Fibroblastic Reticular Cells from Myofibroblastic Precursors Is Critical for Antiviral Immunity. *Immunity*, 38(5), pp.1013–1024.
Chambers, A.F., Groom, A.C. & MacDonald, I.C., 2002. Metastasis: Dissemination and growth of cancer cells in metastatic sites. *Nature Reviews Cancer*, 2(8), pp.563–572.
Chan, C.W.M. et al., 2009. Gastrointestinal differentiation marker Cytokeratin 20 is regulated by homeobox gene CDX1. *Proceedings of the National Academy of Sciences of the United States of America*, 106(6), pp.1936–1941.
Chen, T. et al., 2012. Model-based multidimensional clustering of categorical data. *Artificial Intelligence*, 176(1), pp.2246–2269.
Conti, J. & Thomas, G., 2011. The role of tumour stroma in colorectal cancer invasion and metastasis. *Cancers*, 3(2), pp.2160–2168.
Costa, T., Boccignone, G. & Ferraro, M., 2012. Gaussian mixture model of heart rate variability. *PLoS ONE*, 7(5), p.e37731.
Costea, D.E. et al., 2013. Identification of two distinct carcinoma-associated fibroblast subtypes with differential tumor-promoting abilities in oral squamous cell carcinoma. *Cancer Research*, 73(13), pp.3888–3901.
Cox, T.R. et al., 2013. LOX-mediated collagen crosslinking is responsible for fibrosis-enhanced metastasis. *Cancer Research*, 73(6), pp.1721–1732.
Cox, T.R. & Erler, J.T., 2014. Molecular pathways: Connecting fibrosis and solid tumor metastasis. *Clinical Cancer Research*, 20(14), pp.3637–3643.
Cross, W.C., Graham, T.A. & Wright, N.A., 2016. New paradigms in clonal evolution: punctuated equilibrium in cancer. *The Journal of Pathology*, 240(2), pp.126–136.
D'Haeseleer, P., 2005. How does gene expression clustering work? *Nature Biotechnology*, 23(12), pp.1499–1501.

Dai, X. et al., 2009. A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data. *BMC Bioinformatics*, 10(1), p.165.
Darby, I.A. et al., 2014. Fibroblasts and myofibroblasts in wound healing. *Clinical, Cosmetic and Investigational Dermatology*, 7, pp.301–311.
de Barrios, O. et al., 2017. ZEB1-induced tumourigenesis requires senescence inhibition via activation of DKK1/mutant p53/Mdm2/CtBP and repression of macroH2A1. *Gut*, 66(4), pp.666–682.
De L. Karlson, T., Whiting, C. V. & Bland, P.W., 2006. Proinflammatory cytokine synthesis by mucosal fibroblasts from mouse colitis is enhanced by interferon-gamma-mediated up-regulation of CD40 signalling. *Clinical and Experimental Immunology*, 147(2), pp.313-23.
de Sousa e Melo, F. & Vermeulen, L., 2016. Wnt signaling in cancer stem cell biology. *Cancers*, 8(7).
De Wever, O. et al., 2004. Tenascin-C and SF/HGF produced by myofibroblasts in vitro provide convergent pro-invasive signals to human colon cancer cells through RhoA and Rac. *FASEB J*, 18(9), pp.1016–1018.
Derynck, R., Akhurst, R.J. & Balmain, A., 2001. TGF-beta signaling in tumor suppression and cancer progression. *Nature Genetics*, 29(2), pp.117–29.
Desmouliere, A. et al., 1993. Transforming growth factor-β1 induces α-smooth muscle actin expression in granulation tissue myofibroblasts and in quiescent and growing cultured fibroblasts. *Journal of Cell Biology*, 122(1), pp.103–111.
Desmouliere, Alexis; Gabbiani, G., 2013. The Role of the Myofibroblast in Wound Healing and Fibrocontractive Disease. In *The Molecular and Cellular Biology of Wound Repair*. Boston, MA: Springer US, pp. 9–11.
Di Mari, J.F. et al., 2003. IL-1alpha-induced COX-2 expression in human intestinal myofibroblasts is dependent on a PKCzeta-ROS pathway. *Gastroenterology*, 124(7), pp.1855–1865.
Doolittle, R.F. et al., 1983. Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, 221(4607), pp.275–7.
Dunne, P.D. et al., 2017. Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification. *Nature Communications*, 8(May), p.15657.
El-Gendi, S. & Al-Gendi, A., 2011. Assessment of tumor budding in colorectal carcinoma: Correlation with β-catenin nuclear expression. *Journal of the Egyptian National Cancer Institute*, 23(1), pp.1–9.
Fahlgren, A. et al., 2003. Increased expression of antimicrobial peptides and lysozyme in colonic epithelial cells of patients with ulcerative colitis. *Clinical and Experimental Immunology*, 131(1), pp.90–101.
Farin, H.F., Van Es, J.H. & Clevers, H., 2012. Redundant sources of Wnt regulate intestinal stem cells and promote formation of paneth cells. *Gastroenterology*, 143(6), p.1518–1529.e7.
Fearon, E.R. & Vogelstein, B., 1990. A genetic model for colorectal tumorigenesis. *Cell*, 61(5), pp.759–767.

Ffrench-Constant, C. et al., 1989. Reappearance of an embryonic pattern of fibronectin splicing during wound healing in the adult rat. *Journal of Cell Biology*, 109(2), pp.903–14.
Fisher, R.A., 1919. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2), pp.399–433.
Fodde, R. & Brabletz, T., 2007. Wnt/β-catenin signaling in cancer stemness and malignant behavior. *Current Opinion in Cell Biology*, 19(2), pp.150–158.
Fodde, R. et al., 2001. Mutations in the APC tumour suppressor gene cause chromosomal instability. *Nature Cell Biology*, 3(4), pp.433–438.
Gotzmann, J. et al., 2006. A crucial function of PDGF in TGF-beta-mediated cancer progression of hepatocytes. *Oncogene*, 25(22), pp.3170–3185.
Gould, S. jay & Eldredge, N., 1993. Punctuated equilibrium comes of age. *Nature*, 366(6452), pp.223–227.
Gout, S. & Huot, J., 2008. Role of cancer microenvironment in metastasis: Focus on colon cancer. *Cancer Microenvironment*, 1(1), pp.69–83.
Grainger, S., Hryniuk, A. & Lohnes, D., 2013. Cdx1 and Cdx2 Exhibit Transcriptional Specificity in the Intestine S. R. Singh, ed. *PLoS ONE*, 8(1), p.e54757.
Grant, D.S. et al., 1993. Scatter factor induces blood vessel formation in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 90(5), pp.1937–1941.
Grigore, A. et al., 2016. Tumor Budding: The Name is EMT. Partial EMT. *Journal of Clinical Medicine*, 5(5), p.51.
Grugan, K.D. et al., 2010. Fibroblast-secreted hepatocyte growth factor plays a functional role in esophageal squamous cell carcinoma invasion. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24), pp.11026–31.
Guinney, J. et al., 2015. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21(11), pp.1350–1356.
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.
Hardwick, J.C. et al., 2008. Bone morphogenetic protein signalling in colorectal cancer. *Nature Reviews Cancer*, 8(10), pp.806–812.
Hata, K. et al., 2002. IL-17 stimulates inflammatory responses via NF-kappaB and MAP kinase pathways in human colonic myofibroblasts. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, 282(6), pp.G1035-44.
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.
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.
Heldin, C.-H., 2013. Targeting the PDGF signaling pathway in tumor treatment. *Cell Communication and Signaling*, 11, p.97.
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.
Hill, M.J., Morson, B.C. & Bussey, H.J.R., 1978. Aetiology of adenoma-carcinoma sequence in large bowel. *The Lancet*, 311(8058), pp.245–247.

Hinz, B. et al., 2001. Alpha-Smooth Muscle Actin Expression Upregulates Fibroblast Contractile Activity. *Molecular Biology of the Cell*, 12(9), pp.2730–2741.
Hinz, B. et al., 2012. Recent developments in myofibroblast biology: Paradigms for connective tissue remodeling. *American Journal of Pathology*, 180(4), pp.1340–1355.
Hoey, T. et al., 2009. DLL4 Blockade Inhibits Tumor Growth and Reduces Tumor-Initiating Cell Frequency. *Cell Stem Cell*, 5(2), pp.168–177.
Hosaka, K. et al., 2016. Pericyte-fibroblast transition promotes tumor growth and metastasis. *Proceedings of the National Academy of Sciences of the United States of America*, 113(38), pp.E5618-27.
Hosgood, G., 1993. Wound healing. The role of platelet-derived growth factor and transforming growth factor beta. *Veterinary Surgery*, 22(6), pp.490–495.
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.E2162-71.
Hsia, L.-T. & Bodmer, W.F., 2015. *Identification and characterisation of myofibroblast markers in human colon and colorectal cancer microenvironment* (DPhil thesis).
Huang, F.-I. et al., 2012. Hepatocyte growth factor activates Wnt pathway by transcriptional activation of LEF1 to facilitate tumor invasion. *Carcinogenesis*, 33(6), pp.1142–8.
Huang, J. et al., 1996. APC mutations in colorectal tumors with mismatch repair deficiency. *Proceedings of the National Academy of Sciences of the United States of America*, 93(17), pp.9049–54.
Huang, Y. et al., 2005. A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses. *IEEE Transactions on Biomedical Engineering*, 52(11), pp.1801–1811.
Humphries, A. & Wright, N.A., 2008. Colonic crypt organization and tumorigenesis. *Nature Reviews Cancer*, 8(6), pp.415–424.
Iizuka, M. & Konno, S., 2011. Wound healing of intestinal epithelial cells. *World Journal of Gastroenterology*, 17(17), pp.2161–2171.
Ishii, H. et al., 2014. Epithelial splicing regulatory proteins 1 (ESRP1) and 2 (ESRP2) suppress cancer cell motility via different mechanisms. *Journal of Biological Chemistry*, 289(40), pp.27386–27399.
Itoh, H. et al., 2004. Regeneration of injured intestinal mucosa is impaired in hepatocyte growth factor activator-deficient mice. *Gastroenterology*, 127(5), pp.1423–1435.
Jass, J.R. et al., 2003. APC mutation and tumour budding in colorectal cancer. *Journal of Clinical Pathology*, 56(1), pp.69–73.
Jechlinger, M. et al., 2006. Autocrine PDGFR signaling promotes mammary cancer metastasis. *Journal of Clinical Investigation*, 116(6), pp.1561–1570.
Jia, C.C. et al., 2013. Cancer-Associated Fibroblasts from Hepatocellular Carcinoma Promote Malignant Cell Proliferation by HGF Secretion S. G. Zheng, ed. *PLoS ONE*, 8(5), p.e63243
Jiang, W.G. et al., 1993. Regulation of spreading and growth of colon cancer cells by hepatocyte growth factor. *Clinical & Experimental Metastasis*, 11(3), pp.235–242.

Johnston, M.D. et al., 2007. Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 104(10), pp.4008–4013.
Johnston, M.D. et al., 2010. On the proportion of cancer stem cells in a tumour. *Journal of Theoretical Biology*, 266(4), pp.708–711.
Kabiri, Z. et al., 2014. Stroma provides an intestinal stem cell niche in the absence of epithelial Wnts. *Development*, 141(11), pp.2206–15.
Karp, S.L. et al., 1994. Epithelial differentiation of metanephric mesenchymal cells after stimulation with hepatocyte growth factor or embryonic spinal cord. *Proceedings of the National Academy of Sciences of the United States of America*, 91(12), pp.5286–5290.
Kastanis, G.J. et al., 2011. The role of dystroglycan in PDGF-BB-dependent migration of activated hepatic stellate cells/myofibroblasts. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, 301(3), pp.G464-74.
Kawamoto, M. et al., 1997. Selective migration of alpha-smooth muscle actin-positive myofibroblasts toward fibronectin in the Boyden's blindwell chamber. *Clinical Science*, 93(4), pp.355–362.
Kaye, G.I., Lane, N. & Pascal, R.R., 1968. Colonic pericryptal fibroblast sheath: replication, migration, and cytodifferentiation of a mesenchymal cell system in adult tissue. II. Fine structural aspects of normal rabbit and human colon. *Gastroenterology*, 54(5), pp.852–65.
Kermorgant, S. et al., 2001. Hepatocyte growth factor induces colonic cancer cell invasiveness via enhanced motility and protease overproduction. Evidence for PI3 kinase and PKC involvement. *Carcinogenesis*, 22(7), pp.1035–1042.
Kernochan, L.E. et al., 2002. Endothelin-1 stimulates human colonic myofibroblast contraction and migration. *Gut*, 50(1), pp.65–70.
Kim, B.R. et al., 2015. BMP-2 induces motility and invasiveness by promoting colon cancer stemness through STAT3 activation. *Tumor Biology*, 36(12), pp.9475–9486.
Kim, S.J. et al., 1990. Autoinduction of transforming growth factor beta 1 is mediated by the AP-1 complex. *Molecular and Cellular Biology*, 10(4), pp.1492–1497.
Kim, Y.S. & 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.
Kim, Y.G. et al., 2011. The Nod2 Sensor Promotes Intestinal Pathogen Eradication via the Chemokine CCL2-Dependent Recruitment of Inflammatory Monocytes. *Immunity*, 34(5), pp.769–780.
Kinzler, K.W. & Vogelstein, B., 1996. Lessons from hereditary colorectal cancer. *Cell*, 87(2), pp.159–170.
Kitadai, Y. et al., 2006. Targeting the expression of platelet-derived growth factor receptor by reactive stroma inhibits growth and metastasis of human colon carcinoma. *American Journal of Pathology*, 169(6), pp.2054–2065.
Klingberg, F., Hinz, B. & White, E.S., 2013. The myofibroblast matrix: Implications for tissue repair andfibrosis. *Journal of Pathology*, 229(2), pp.298–309.
Klockars, M. et al., 1977. Immunohistochemical identification of lysozyme in intestinal lesions in ulcerative colitis and Crohn's disease. *Gut*, 18(5), pp.377–81.

Ko, J.Y. et al., 2008. Increased IL-1β expression and myofibroblast recruitment in subacromial bursa is associated with rotator cuff lesions with shoulder stiffness. *Journal of Orthopaedic Research*, 26(8), pp.1090–1097.
Ko, Y., Zhai, C. & Rodriguez-Zas, S.L., 2007. Inference of Gene Pathways Using Gaussian Mixture Models. *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)*, (1), pp.362–367.
Koliaraki, V., Roulis, M. & Kollias, G., 2012. Tpl2 regulates intestinal myofibroblast HGF release to suppress colitis-associated tumorigenesis. *Journal of Clinical Investigation*, 122(11), pp.4231–4242.
Kondo, T. et al., 2016. The impact of hepatic fibrosis on the incidence of liver metastasis from colorectal cancer. *British Journal of Cancer*, 115(1), pp.34–9.
Konishi, M. et al., 1996. Molecular nature of colon tumors in hereditary nonpolyposis colon cancer, familial polyposis, and sporadic colon cancer. *Gastroenterology*, 111(2), pp.307–317.
Korinek, V. et al., 1998. Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nature Genetics*, 19(4), pp.379–383.
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.15418–23.
Kosinski, C. et al., 2010. Indian hedgehog regulates intestinal stem cell fate through epithelial-mesenchymal interactions during development. *Gastroenterology*, 139(3), pp.893–903.
Krasinskas, A.M., 2011. EGFR Signaling in Colorectal Carcinoma. *Pathology Research International*, 2011, p.932932.
Ku, H.H., 1966. Notes on the use of propagation of error formulas. *Journal of Research of the National Bureau of Standards, Section C: Engineering and Instrumentation*, 70C(4), p.263.

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.266–71.
Lakatos, P.-L. & Lakatos, L., 2008. Risk for colorectal cancer in ulcerative colitis: changes, causes and management strategies. *World Journal of Gastroenterology*, 14(25), pp.3937–47.
Lee, W. et al., 2012. Multiple response regression for gaussian mixture models with known labels. *Statistical Analysis and Data Mining*, 5(6), pp.493–508.
Leeb, S.N. et al., 2002. Regulation of migration of human colonic myofibroblasts. *Growth Factors*, 20(2), pp.81–91.
Lei, N.Y. et al., 2014. Intestinal subepithelial myofibroblasts support the growth of intestinal epithelial stem cells J. Sun, ed. *PLoS ONE*, 9(1), p.e84651.
Leshem, Y. et al., 2000. Hepatocyte growth factor (HGF) inhibits skeletal muscle cell differentiation: A role for the bHLH protein Twist and the cdk inhibitor p27. *Journal of Cellular Physiology*, 184(1), pp.101–109.

Liang, P. et al., 2005. Myofibroblasts correlate with lymphatic microvessel density and lymph node metastasis in early-stage invasive colorectal carcinoma. *Anticancer Research*, 25(4), pp.2705–2712.
Liska, D. et al., 2011. HGF rescues colorectal cancer cells from EGFR inhibition via MET activation. *Clinical Cancer Research*, 17(3), pp.472–482.
Liu, L. et al., 2016. Stromal Myofibroblasts Are Associated with Poor Prognosis in Solid Cancers: A Meta-Analysis of Published Studies. *PLoS ONE*, 11(7), p.e0159947.
Liu, Y., Li, Q. & Zhu, L., 2012. Expression of the hepatocyte growth factor and c-Met in colon cancer: Correlation with clinicopathological features and overall survival. *Tumori*, 98(1), pp.105–112.
Liu, Y. et al., 2008. Zeb1 links epithelial-mesenchymal transition and cellular senescence. *Development*, 135(3), pp.579–88.
Lo, Y.H. et al., 2017. Transcriptional Regulation by ATOH1 and its Target SPDEF in the Intestine. *Cellular and Molecular Gastroenterology and Hepatology*, 3(1), pp.51–71.
Lobb, R.J., Lima, L.G. & Möller, A., 2017. Exosomes: Key mediators of metastasis and pre-metastatic niche formation. *Seminars in Cell and Developmental Biology*, 67, pp.3–10.
Mahale, J. et al., 2016. The role of stromal fibroblasts in lung carcinogenesis: A target for chemoprevention? *International Journal of Cancer*, 138(1), pp.30–44.
Majno, G., 1979. The story of the myofibroblasts. *The American Journal of Surgical Pathology*, 3(6), pp.535–42.
Manzat Saplacan, R.M. et al., 2017. The Role of PDGFs and PDGFRs in Colorectal Cancer. *Mediators of Inflammation*, 2017, pp.1–9.
Markowitz, S.D. & Bertagnolli, M.M., 2009. Molecular Basis of Colorectal Cancer. *New England Journal of Medicine*, 361(25), pp.2449–60.
Marshman, E., Booth, C. & Potten, C.S., 2002. The intestinal epithelial stem cell. *BioEssays*, 24(1), pp.91–98.
Martin, P., 1997. Wound healing-aiming for perfect skin regeneration. *Science*, 276(5309), pp.75–81.
Marusyk, A., Almendro, V. & Polyak, K., 2012. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5), pp.323–34.
Marx, V., 2013. Biology: The big challenges of big data. *Nature*, 498(7453), pp.255–260.
Massagué, J. & Obenauf, A.C., 2016. Metastatic colonization by circulating tumour cells. *Nature*, 529(7586), pp.298–306.
Masters, J.R., 2000. Human cancer cell lines: fact and fantasy. *Nature Reviews Molecular Cell biology*, 1(3), pp.233–236.
Matsuda, Y., Ueda, J. & Ishiwata, T., 2012. Fibroblast growth factor receptor 2: Expression, roles, and potential as a novel molecular target for colorectal cancer. *Pathology Research International*, 2012, pp.1–8.
Mayr, E. & Provine, W.B., 1998. *The Evolutionary Synthesis: Perspectives on the Unification of Biology*, Harvard University Press.
McAnulty, R.J., 2007. Fibroblasts and myofibroblasts: Their source, function and role in disease. *International Journal of Biochemistry and Cell Biology*, 39(4), pp.666–671.

McCarty, M.F. et al., 2007. Overexpression of PDGF-BB decreases colorectal and pancreatic cancer growth by increasing tumor pericyte content. *Journal of Clinical Investigation*, 117(8), pp.2114–2122.
McIntire, M.G. et al., 2011. MUC2 is a highly specific marker of goblet cell metaplasia in the distal esophagus and gastroesophageal junction. *The American Journal of Surgical Pathology*, 35(7), pp.1007–1013.
McKaig, B. et al., 1998. Human colonic subepithelial myofibroblasts enhance epithelial restitution. *Gastroenterology*, 114, p.A1036.
Mclachlan, G.J. & 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.
McMillan, D.H. et al., 2011. Lung-targeted overexpression of the NF-κB member RelB inhibits cigarette smoke-induced inflammation. *American Journal of Pathology*, 179(1), pp.125–133.
McNicholas, P.D., 2016. Model-Based Clustering. *Journal of Classification*, 33, pp.331–373.
Medema, J.P. & Vermeulen, L., 2011. Microenvironmental regulation of stem cells in intestinal homeostasis and cancer. *Nature*, 474(7351), pp.318–26.
Mele, V. et al., 2014. Mesenchymal stromal cells induce epithelial-to-mesenchymal transition in human colorectal cancer cells through the expression of surface-bound TGF-β. *International Journal of Cancer*, 134(11), pp.2583–2594.
Melnykov, V. & Maitra, R., 2010. Finite mixture models and model-based clustering. *Statistics Surveys*, 4, pp.80–116.
Merlo, L.M.F. et al., 2006. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6(12), pp.924–935.
Meyer-ter-Vehn, T. et al., 2006. Contractility as a Prerequisite for TGF-β–Induced Myofibroblast Transdifferentiation in Human Tenon Fibroblasts. *Investigative Ophthalmology and Visual Science*, 47(11), p.4895.
Meyer, R.C. et al., 2013. GPR37 and GPR37L1 are receptors for the neuroprotective and glioprotective factors prosaptide and prosaposin. *Proceedings of the National Academy of Sciences of the United States of America*, 110(23), pp.9529–34.
Mifflin, R.C. et al., 2011. Intestinal myofibroblasts: targets for stem cell therapy. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, 300(5), pp.G684–G696.
Mirabelli-Primdahl, L. et al., 1999. β-Catenin mutations are specific for colorectal carcinomas with microsatellite instability but occur in endometrial carcinomas irrespective of mutator pathway. *Cancer Research*, 59(14), pp.3346–3351.
Mitrovic, B. et al., 2012. Tumor budding in colorectal carcinoma: time to take notice. *Modern Pathology*, 25(10), pp.1315–25.
Monga, S.P.S. et al., 2002. Hepatocyte growth factor induces Wnt-independent nuclear translocation of beta-catenin after Met-beta-catenin dissociation in hepatocytes. *Cancer Research*, 62(7), pp.2064–2071.
Moore, R., Carlson, S. & Madara, J.L., 1989. Villus contraction aids repair of intestinal epithelium after injury. *American Journal of Physiology*, 257(2 Pt 1), pp.G274-83.

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.
Nakamura, T., Teramoto, H. & Ichihara, A., 1986. Purification and characterization of a growth factor from rat platelets for mature parenchymal hepatocytes in primary cultures. *Proceedings of the National Academy of Sciences of the United States of America*, 83(17), pp.6489–93.
Nakamura, T. & Mizuno, S., 2010. The discovery of Hepatocyte Growth Factor (HGF) and its significance for cell biology, life sciences and clinical medicine. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, 86(6), pp.588–610.
Nakamura, Y. et al., 2008. PDGF-BB is a Novel Prognostic Factor in Colorectal Cancer. *Annals of Surgical Oncology*, 15(8), pp.2129–2136.
Naldini, L. et al., 1991. Scatter factor and hepatocyte growth factor are indistinguishable ligands for the MET receptor. *The EMBO Journal*, 10(10), pp.2867–2878.
Normand, S. et al., 2011. Nod-like receptor pyrin domain-containing protein 6 (NLRP6) controls epithelial self-renewal and colorectal carcinogenesis upon injury. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23), pp.9601–9606.
Novelli, M.R. et al., 1996. Polyclonal origin of colonic adenomas in an XO/XY patient with FAP. *Science*, 272(5265), pp.1187–1190.
Nowell, P.C., 1976. The clonal evolution of tumor cell populations. *Science (New Yor*, 194(4260), pp.23–8.
Okuno, T. et al., 2002. Interleukin-1β and Tumor Necrosis Factor-α Induce Chemokine and Matrix Metalloproteinase Gene Expression in Human Colonic Subepithelial Myofibroblasts. *Scandinavian Journal of Gastroenterology*, 37(3), pp.317–324.
Olaso, E. et al., 1997. Tumor-dependent activation of rodent hepatic stellate cells during experimental melanoma metastasis. *Hepatology*, 26(3), pp.634–42.
Olschwang, S. et al., 1997. Alternative genetic pathways in colorectal carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(22), pp.12122–7.
Otte, J.M. et al., 2000. Functional expression of HGF and its receptor in human colorectal cancer. *Digestion*, 61(4), pp.237–246.
Otte, J.M., Rosenberg, I.M. & Podolsky, D.K., 2003. Intestinal myofibroblasts in innate immune responses of the intestine. *Gastroenterology*, 124(7), pp.1866–1878.
Owens, B.M.J. & Simmons, A., 2013. Intestinal stromal cells in mucosal immunity and homeostasis. *Mucosal immunology*, 6(2), pp.224–34.
Paget, S., 1889. The distribution of secondary growth in cancer of breast. *The Lancet*, 133(3421), pp.571–573.
Pang, G. et al., 1994. GM-CSF, IL-1 alpha, IL-1 beta, IL-6, IL-8, IL-10, ICAM-1 and VCAM-1 gene expression and cytokine production in human duodenal fibroblasts stimulated with lipopolysaccharide, IL-1 alpha and TNF-alpha. *Clinical and Experimental Immunology*, 96(3), pp.437–43.
Pascal, R.R., Kaye, G.I. & Lane, M.D, N., 1968. Colonic Pericryptal Fibroblast Sheath: Replication, Migration, and Cytodifferentiation of a Mesenchymal Cell System in

Adult Tissue: I. Autoradiographic studies of normal rabbit colon. *Gastroenterology*, 54(5), pp.835–851.
Pelaseyed, T. et al., 2014. The mucus and mucins of the goblet cells and enterocytes provide the first defense line of the gastrointestinal tract and interact with the immune system. *Immunological Reviews*, 260(1), pp.8–20.
Pena, C. et al., 2013. STC1 expression by cancer-associated fibroblasts drives metastasis of colorectal cancer. *Cancer Research*, 73(4), pp.1287–1297.
Pierce, G.F. et al., 1991. Role of platelet-derived growth factor in wound healing. *Journal of Cellular Biochemistry*, 45(4), pp.319–326.
Pietras, K. et al., 2008. Functions of paracrine PDGF signaling in the proangiogenic tumor stroma revealed by pharmacological targeting C. L. Sawyers, ed. *PLoS Medicine*, 5(1), pp.0123–0138.
Pinchuk, I. V et al., 2007. Monocyte Chemoattractant Protein-1 Production by Intestinal Myofibroblasts in Response to Staphylococcal Enterotoxin A: Relevance to Staphylococcal Enterotoxigenic Disease. *The Journal of Immunology*, 178(12), pp.8097–8106.
Pinto, D. et al., 2003. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium. *Genes and Development*, 17(14), pp.1709–1713.
Plaks, V., Kong, N. & Werb, Z., 2015. The cancer stem cell niche: How essential is the niche in regulating stemness of tumor cells? *Cell Stem Cell*, 16(3), pp.225–238.
Potente, M., Gerhardt, H. & Carmeliet, P., 2011. Basic and therapeutic aspects of angiogenesis. *Cell*, 146(6), pp.873–887.
Poulogiannis, G., Frayling, I.M. & Arends, M.J., 2010. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology*, 56(2), pp.167–179.
Powell, D.W. et al., 1999. Myofibroblasts. II. Intestinal subepithelial myofibroblasts. *The American Journal of Physiology*, 277(1), pp.C183–C201.
Powell, D.W. et al., 2011. Mesenchymal cells of the intestinal lamina propria. *Annual Review of Physiology*, 73, pp.213–37.
Powell, D.W. et al., 2005. Epithelial cells and their neighbors I. Role of intestinal myofibroblasts in development, repair, and cancer. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, 289(1), pp.G2-7.
Powell, S.M. et al., 1992. APC mutations occur early during colorectal tumorigenesis. *Nature*, 359(6392), pp.235–237.
Prajapati, R.T. et al., 2000. Mechanical loading regulates protease production by fibroblasts in three-dimensional collagen substrates. *Wound Repair and Regeneration*, 8(3), pp.226–237.
Preston, S.L. et al., 2003. Bottom-up Histogenesis of Colorectal Adenomas. *Cancer Research*, 63(13).
Reigstad, L.J., Varhaug, J.E. & Lillehaug, J.R., 2005. Structural and functional specificities of PDGF-C and PDGF-D, the novel members of the platelet-derived growth factors family. *FEBS Journal*, 272(22), pp.5723–5741.
Reya, T. & Clevers, H., 2005. Wnt signalling in stem cells and cancer. *Nature*, 434(7035), pp.843–850.
Richman, P.I. et al., 1987. Colonic pericrypt sheath cells: characterisation of cell type with new monoclonal antibody. *Journal of Clinical Pathology*, 40(6), pp.593–600.

Rieder, F. et al., 2007. Wound healing and fibrosis in intestinal disease. *Gut*, 56(1), pp.130–139.
Riihimäki, M. et al., 2016. Patterns of metastasis in colon and rectal cancer. *Scientific Reports*, 6, p.29765.
Ringnér, M., 2008. What is principal component analysis? *Nature Biotechnology*, 26(3), pp.303–304.
Roulis, M. et al., 2011. Intestinal epithelial cells as producers but not targets of chronic TNF suffice to cause murine Crohn-like pathology. *Proceedings of the National Academy of Sciences of the United States of America*, 108(13), pp.5396–5401.
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.
Saada, J.I. et al., 2006. Subepithelial myofibroblasts are novel nonprofessional APCs in the human colonic mucosa. *Journal of Immunology*, 177(9), pp.5968–5979.
Santamaria-Martínez, A. & Huelsken, J., 2013. The niche under siege: Novel targets for metastasis therapy. *Journal of Internal Medicine*, 274(2), pp.127–136.
Sappino, A.P. et al., 1989. Colonic pericryptal fibroblasts - Differentiation pattern in embryogenesis and phenotypic modulation in epithelial proliferative lesions. *Virchows Archiv A Pathological Anatomy and Histopathology*, 415(6), pp.551–557.
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), p.1607327113-.
Sato, T. et al., 2009. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature*, 459(7244), pp.262–5.
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.
Sceneay, J., Smyth, M.J. & Möller, A., 2013. The pre-metastatic niche: finding common ground. *Cancer and Metastasis Reviews*, 32(3–4), pp.449–464.
Schürch, W., Seemayer, T. & Gabbiani, G., 1998. The Myofibroblast : A Quarter Century After Its Discovery Back to Top. *The American Journal of Surgical Pathology*, 22(2), pp.141–147.
Scrucca, L. et al., 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1), pp.289–317.
Serini, G. et al., 1998. The fibronectin domain ED-A is crucial for myofibroblastic phenotype induction by transforming growth factor-β1. *Journal of Cell Biology*, 142(3), pp.873–881.
Serrano-Gomez, S.J., Maziveyi, M. & Alahari, S.K., 2016. Regulation of epithelial-mesenchymal transition through epigenetic and post-translational modifications. *Molecular Cancer*, 15(1), p.18.
Shao, J. et al., 2006. Roles of myofibroblasts in prostaglandin E2-stimulated intestinal epithelial proliferation and angiogenesis. *Cancer Research*, 66(2), pp.846–855.
Shen, B.Q. et al., 1997. Hepatocyte growth factor stimulates the differentiation of human tracheal epithelia in vitro. *American Journal of Physiology*, 272(6 Pt 1), pp.L1115-20.

Shimizu, A. et al., 1999. The dermatofibrosarcoma protuberans-associated collagen type Iα1/platelet-derived growth factor (PDGF) B-chain fusion gene generates a transforming protein that is processed to functional PDGF-BB. *Cancer Research*, 59(15), pp.3719–3723.
Silen, W. & Ito, S., 1985. Mechanisms for rapid re-epithelialization of the gastric mucosal surface. *Annual Review of Physiology*, 47(2A), pp.217–29.
Singer, A.J.. & Clark, R.A., 1999. Cutaneous Wound Healing. *The New England Journal of Medicine*, 341(10), pp.738–746.
Singh, A. et al., 2004. Rac1b, a tumor associated, constitutively active Rac1 splice variant, promotes cellular transformation. *Oncogene*, 23(58), pp.9369–9380.
Slonim, D.K. & Yanai, I., 2009. Getting started in gene expression microarray analysis. *PLoS Computational Biology*, 5(10), p.e1000543.
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.
Song, Z.Y., Gao, Z.H. & Qu, X.J., 2014. A review of CXCR4/CXCL12 axis in colorectal cancer. *Biomedicine and Aging Pathology*, 4(3), pp.285–290.
Stahtea, X.N. et al., 2007. Imatinib inhibits colorectal cancer cell growth and suppresses stromal-induced growth stimulation, MT1-MMP expression and pro-MMP2 activation. *International Journal of Cancer*, 121(12), pp.2808–2814.
Stallmach, A. et al., 1992. Increased collagen type III synthesis by fibroblasts isolated from strictures of patients with Crohn's disease. *Gastroenterology*, 102(6), pp.1920–9.
Stein, U. et al., 2009. MACC1, a newly identified key regulator of HGF-MET signaling, predicts colon cancer metastasis. *Nature Medicine*, 15(1), pp.59–67.
Steller, E.J.A. et al., 2013. PDGFRB promotes liver metastasis formation of mesenchymal-like colorectal tumor cells. *Neoplasia*, 15(2), pp.204–17.
Stoker, M. et al., 1987. Scatter factor is a fibroblast-derived modulator of epithelial cell mobility. *Nature*, 327(6119), pp.239–242.
Stratton, M.R., Campbell, P.J. & Andrew F, P., 2009. The cancer genome. *Nature*, 458(7239), pp.719–724.
Sun, H.Q. et al., 1999. Gelsolin, a multifunctional actin regulatory protein. *Journal of Biological Chemistry*, 274(47), pp.33179–33182.
Sundberg, C. et al., 1997. Tumor cell and connective tissue cell interactions in human colorectal adenocarcinoma. Transfer of platelet-derived growth factor-AB/BB to stromal cells. *The American Journal of Pathology*, 151(2), pp.479–92.
Surawska, H., Ma, P.C. & Salgia, R., 2004. The role of ephrins and Eph receptors in cancer. *Cytokine and Growth Factor Reviews*, 15(6), pp.419–433.
Takaku, K. et al., 1998. Intestinal tumorigenesis in compound mutant mice of both Dpc4 (Smad4) and Apc genes. *Cell*, 92(5), pp.645–656.
Takatsuna, M. et al., 2016. Myofibroblast distribution is associated with invasive growth types of colorectal cancer. *Oncology Reports*, 36(6), pp.3154–3160.
Tamaoki, M. et al., 2005. Tenascin-C regulates recruitment of myofibroblasts during tissue repair after myocardial injury. *The American Journal of Pathology*, 167(1), pp.71–80.

Tejada, M.L. et al., 2006. Tumor-Driven Paracrine Platelet-Derived Growth Factor Receptor Signaling Is a Key Determinant of Stromal Cell Recruitment in a Model of Human Lung Carcinoma. *Clinical Cancer Research*, 12(9), pp.2676–2688.
Than, B.L.N. et al., 2016. CFTR is a tumor suppressor gene in murine and human intestinal cancer. *Oncogene*, 35(32), pp.4179–4187.
Theiss, A.L. et al., 2005. Tumor necrosis factor (TNF) alpha increases collagen accumulation and proliferation in intestinal myofibroblasts via TNF receptor 2. *The Journal of Biological Chemistry*, 280(43), pp.36099–109.
Tomasek, J.J. et al., 2002. Myofibroblasts and mechano-regulation of connective tissue remodelling. *Nature Reviews Molecular Cell Biology*, 3(5), pp.349–63.
Tomasek, J.J. et al., 1992. Fibroblast contraction occurs on release of tension in attached collagen lattices: Dependency on an organized actin cytoskeleton and serum. *The Anatomical Record*, 232(3), pp.359–368.
Tommelein, J. et al., 2015. Cancer-associated fibroblasts connect metastasis-promoting communication in colorectal cancer. *Frontiers in Oncology*, 5(March), p.63.
Tsujino, T. et al., 2007. Stromal myofibroblasts predict disease recurrence for colorectal cancer. *Clinical Cancer Research*, 13(7), pp.2082–2090.
Tyan, S.W. et al., 2011. Breast cancer cells induce cancer-associated fibroblasts to secrete hepatocyte growth factor to enhance breast tumorigenesis X. Guan, ed. *PLoS ONE*, 6(1), p.e15313.
Uehara, Y., 1992. Expression of a Human Hepatocyte Growth Factor/Scatter Factor eDNA in MDCK Epithelial Cells Influences Cell Morphology, Motility, and Anchorage-independent Growth. *The Journal of Cell Biology*, 117(4), pp.889–894.
Uhlén, M. et al., 2015. Tissue-based map of the human proteome. *Science*, 347(6220), p.1260419.
Uhrbom, L. et al., 2000. Dependence of autocrine growth factor stimulation in platelet-derived growth factor-B-induced mouse brain tumor cells. *International Journal of Cancer*, 85(3), pp.398–406.
Ustach, C. V. et al., 2010. A novel signaling axis of matriptase/PDGF-D/ss-PDGFR in human prostate cancer. *Cancer Research*, 70(23), pp.9631–9640.
Valenta, T. et al., 2016. Wnt Ligands Secreted by Subepithelial Mesenchymal Cells Are Essential for the Survival of Intestinal Stem Cells and Gut Homeostasis. *Cell Reports*, 15(5), pp.911–918.
van den Brink, G.R. et al., 2004. Indian Hedgehog is an antagonist of Wnt signaling in colonic epithelial cell differentiation. *Nature Genetics*, 36(3), pp.277–282.
Van Der Maaten, L. & Hinton, G., 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9, pp.2579–2605.
Van Dop, W.A. et al., 2010. Loss of Indian hedgehog activates multiple aspects of a wound healing response in the mouse intestine. *Gastroenterology*, 139(5), p.1665–1676.e10.
Vermeulen, L. et al., 2008. Cancer stem cells – old concepts, new insights. *Cell Death and Differentiation*, 15(6), pp.947–958.
Vermeulen, L. et al., 2010. Wnt activity defines colon cancer stem cells and is regulated by the microenvironment. *Nature Cell Biology*, 12(5), pp.468–476.

Vishnubalaji, R. et al., 2016. Bone morphogenetic protein 2 (BMP2) induces growth suppression and enhances chemosensitivity of human colon cancer cells. *Cancer Cell International*, 16(1), p.77.
Visvader, J.E., 2011. Cells of origin in cancer. *Nature*, 469(7330), pp.314–322.
Vogelstein, B. et al., 2013. Cancer Genome Landscapes. *Science*, 339(6127), pp.1546–1558.
Volkenstein, M. V, 1987. Punctualism, non-adaptationism, neutralism and evolution. *Bio Systems*, 20(3), pp.289–304.
Wallace, J.L. & Granger, D.N., 1996. The cellular and molecular basis of gastric mucosal defense. *FASEB J*, 10(7), pp.731–40.
Walther, A. et al., 2009. Genetic prognostic and predictive markers in colorectal cancer. *Nature Reviews Cancer*, 9(7), pp.489–99.
Walton, K.L.W., Holt, L. & Sartor, R.B., 2009. Lipopolysaccharide activates innate immune responses in murine intestinal myofibroblasts through multiple signaling pathways. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, 296(3), pp.G601-11.
Wang, D. et al., 2011. Paneth cell marker expression in intestinal villi and colon crypts characterizes dietary induced risk for mouse sporadic intestinal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25), pp.10272–7.
Waterfield, M.D. et al., 1983. Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus. *Nature*, 304(5921), pp.35–9.
Werner, B. et al., 2016. The Cancer Stem Cell Fraction in Hierarchically Organized Tumors Can Be Estimated Using Mathematical Modeling and Patient-Specific Treatment Trajectories. *Cancer Research*, 76(7).
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.21058–63.
Wong, N.A.C.S. et al., 2006. EpCAM and gpA33 are markers of Barrett's metaplasia. *Journal of Clinical Pathology*, 59(3), pp.260–3.
Wynn, T.A. & Ramalingam, T.R., 2012. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nature Medicine*, 18(7), pp.1028–40.
Xia, Y. et al., 1997. RelB regulation of chemokine expression modulates local inflammation. *American Journal of Pathology*, 151(2), pp.375–387.
Xue, Y. et al., 2012. PDGF-BB modulates hematopoiesis and tumor angiogenesis by inducing erythropoietin production in stromal cells. *Nature Medicine*, 18(1), pp.100–10.
Yagi, Y. et al., 2007. Inflammatory responses induced by interleukin-17 family members in human colonic subepithelial myofibroblasts. *Journal of Gastroenterology*, 42(9), pp.746–753.
Yamauchi, S. et al., 2014. Clinical Significance of Platelet-Derived Growth Factor-C Expression in Colorectal Cancer. *Journal of Cancer Therapy*, 5(1), pp.11–20.
Yeung, T.M. et al., 2013. Myofibroblast activation in colorectal cancer lymph node metastases. *British Journal of Cancer*, 108(10), pp.2106–15.

Yeung, T.M. et al., 2010. Cancer stem cells from colorectal cancer-derived cell lines. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8), pp.3722–7.
Yuan, X. et al., 2015. Notch signaling: An emerging therapeutic target for cancer treatment. *Cancer Letters*, 369(1), pp.20–27.
Yuge, R. et al., 2014. mTOR and PDGF pathway blockade inhibits liver metastasis of colorectal cancer by modulating the tumor microenvironment. *American Journal of Pathology*, 185(2), pp.399–408.
Zanella, E.R. et al., 2015. IGF2 is an actionable target that identifies a distinct subpopulation of colorectal cancer patients with marginal response to anti-EGFR therapies. *Science Translational Medicine*, 7(272), p.272ra12.
Zhang, Y. et al., 2014. Bone morphogenetic protein 2 inhibits the proliferation and growth of human colorectal cancer cells. *Oncology Reports*, 32(3), pp.1013–20.
Zheng, C. et al., 2011. A novel Anti-CEACAM5 Monoclonal antibody, CC4, suppresses colorectal tumor growth and enhances NK cells-mediated tumor immunity J. Zimmer, ed. *PLoS ONE*, 6(6), p.e21146.

Appendix

Table A.1: List of the 500 top probesets upregulated in strict myofibroblast-recruiting cell lines (recruitment score 5) compared to strict non-recruiters (recruitment score 1), sorted by $\log_2$ of the expression ratio between them. P-values uncorrected for multiple comparisons are given.

Gene Symbol	p-value	$\log_2$(Ratio)
DKK1	0.00508866	3.527583347
PHLDB2	0.017618	2.882615662
HOXB8	0.000321707	2.590585254
SLC40A1	0.0231387	2.500468104
PROM1	0.0588223	2.445178439
SERPINA1	0.0280616	2.430510091
SERPINA1	0.00934401	2.349424728
RDX	0.05044	2.34508123
CXCR4	0.0154666	2.312673885
LOC404266	0.00742873	2.296166636
KCTD12	0.0187923	2.279411863
TACSTD2	0.112629	2.274804341
HOXB6	0.0018931	2.169588333
---	0.000737419	2.13349448
PDE4B	0.0245409	2.094296885
MATN2	0.0139223	2.017291259
FRAS1	0.05212	1.983772534
EIF1AY	0.0529422	1.980931957
ZNF655	0.0746304	1.975916276
SMARCA1	0.0299142	1.97368474
C13orf15	0.0537176	1.962282484
AREG	0.0370461	1.960730396
NT5E	0.0273756	1.953663811
ANXA6	0.0068936	1.928377904
AMIGO2	0.0733011	1.92677371
KLK10	0.0644524	1.921790486
HOXB9	0.00260241	1.913251341
IAH1	0.000267095	1.907648292
CRNDE	0.01534	1.894825522
GRB14	0.0222156	1.869125196
HOXA3	0.003055	1.861427404
LYZ	0.184139	1.859019362
TIMP2	0.0979588	1.847820566
HSPA1A	0.201276	1.844052159
SMARCA1	0.044215	1.830847535
DDX3Y	0.0729142	1.796099328
PTGER2	0.00401538	1.784009739
PMEPA1	0.0562797	1.782731458
EDN1	0.0232166	1.774194577
GPR37	0.00223427	1.765318326

ANXA3	0.058729	1.765004246
---	0.00460865	1.762901548
TIMP2	0.0753862	1.757863862
NT5E	0.019374	1.746669633
NPNT	0.0655078	1.745542837
PHLDB2	0.0297581	1.74544818
SEC22B	0.000994878	1.739558671
EPAS1	0.0205447	1.713774985
---	0.00451022	1.694768783
INHBA	0.0122571	1.681687633
IFIH1	0.00381179	1.676226781
CXCR4	0.0162977	1.674275331
SAMD9	0.0255995	1.67257017
KCTD12	0.0173562	1.662023157
PMEPA1	0.0607513	1.644377774
LCN2	0.147426	1.642576818
SRGAP2P1	0.0105884	1.6332917
MTSS1	0.0139322	1.614559089
EGFR	0.00389318	1.614224546
LYZ	0.29896	1.610889008
RHOBTB3	0.0106495	1.608208389
FSTL1	0.0475697	1.60557967
NNMT	0.0543919	1.582368127
TUBA1A	0.121673	1.573486021
SDR16C5	0.0721421	1.565796995
HOXB2	0.00647214	1.565650789
RHOBTB3	0.0149427	1.561844216
ITGB8	0.067045	1.557405188
ABCG1	0.0231568	1.553989548
---	0.0223722	1.541950903
CXCR4	0.0127636	1.526434448
EIF1AY	0.0331013	1.525998679
SEMA3C	0.0388352	1.525758198
CYorf15A	0.0375004	1.520678705
GPX8	0.107446	1.519989689
AMOTL2	0.0100746	1.517255533
QPCT	0.12501	1.51403639
ARMCX3	0.0676178	1.511690676
HOXB3	0.0178374	1.511053034
IGFBP7	0.0839641	1.504574632
ZIC2	0.116529	1.497121804
PDE4B	0.0182036	1.495613305
ANP32E	0.0382738	1.494922449
TOB1	0.0321414	1.491976201
PLK2	0.12662	1.483844147
ETS1	0.0965476	1.471125016

CRNDE	0.0335345	1.466564478
FAM110C	0.0324656	1.46320392
GALNT14	0.0107002	1.461947582
HOXC6	0.0827513	1.460160651
CXCL5	0.153924	1.459667677
UNC5A	0.00223395	1.457357894
NR1D2	0.000525919	1.452853695
SMARCA1	0.039076	1.450775764
HOXA10 /// HOXA9	0.182384	1.450406276
---	0.0364625	1.438990074
GLIS2	0.0212771	1.437627248
LEPREL1	0.00601134	1.432745659
CPNE8	0.0110747	1.429819467
CD44	0.0546802	1.424948954
ZDHHC2	0.0705368	1.416936999
HSPH1	0.00230166	1.401548555
ARMCX3	0.0438224	1.400904022
EPB41L4A	0.0427413	1.40033572
SLC16A7	0.0210576	1.392740551
DMKN	0.079748	1.390777727
LOX	0.00324773	1.390667686
CDC42EP3	0.028275	1.389010551
COCH	0.00530687	1.385883793
TICAM2 /// TMED7-TICAM2	0.00362781	1.385635347
TUSC1	0.0430044	1.379105226
RHOBTB3	0.0311717	1.378966555
SYT13	0.120536	1.375606664
HSPH1	0.00130162	1.372467405
RAB27B	0.0640896	1.370627431
LARP6	0.0700409	1.369042058
FGD6	0.00956671	1.368399598
TFPI2	0.0670642	1.364555624
ARMCX6	0.063189	1.364499595
RUNX2	0.111039	1.363714965
---	0.0361404	1.358244332
RDX	0.12462	1.355991643
BASP1	0.0735861	1.351645289
FXR1	0.00404918	1.347909392
APOOL	0.00342766	1.342668518
ZDHHC2	0.0330819	1.339536494
MDM4	0.022136	1.3373857
CFLAR	0.000728827	1.333910275
ABLIM3	0.0168397	1.332295472
KDM5D	0.0463106	1.331848504
TGFBR2	0.0433581	1.331378467
MAL2	0.0992828	1.33085666

OBSL1	0.149697	1.329869872
LPAR1	0.0613383	1.328882409
CFLAR	0.000239395	1.328635438
OSMR	0.0332925	1.320669516
RBM6	0.0448369	1.317911817
---	0.0280363	1.314748723
EPDR1	0.225109	1.312845209
PDGFC	0.229942	1.312554818
---	0.036217	1.31205521
PDGFA	0.0488173	1.312049399
RHOBTB3	0.041818	1.311770472
EFNA5	0.0427155	1.310427378
NNMT	0.0318365	1.309071387
CAV1	0.330493	1.303055931
CD36	0.107529	1.300709472
EFNA5	0.0355894	1.300645051
INHBA	0.0190611	1.300609911
OGFRL1	0.201075	1.291521049
GLS	0.012458	1.280309034
CD36	0.0735713	1.279857547
CYorf15A	0.0352658	1.276782496
OSMR	0.0632159	1.276615769
HOXB9	0.00440594	1.267385978
CPEB4	0.00570919	1.264362276
CD44	0.172041	1.263671455
LY6K	0.00200328	1.256708768
GPX3	0.204624	1.25651555
LOC100505937	0.0319206	1.255385917
GSN	0.023502	1.255373831
YPEL5	0.00130122	1.254533577
ADAMTS9	0.00430848	1.253571822
SMAD3	0.000330267	1.248923389
DPYD	0.0425757	1.248893037
GJC1	0.247329	1.248856614
CHST15	0.152695	1.248322302
HLA-DQB1 /// LOC100133583	0.0346565	1.246858004
ITGB4	0.0483309	1.246675622
UGCG	0.00487899	1.242590312
WIPI1	0.0199513	1.240247155
C8orf84	0.157644	1.239941776
IRX2	0.146294	1.239892909
ITGB4	0.0269161	1.239165823
ALDH7A1	0.024073	1.238199881
IRF6	0.0361682	1.236706902
EGFR	0.0154168	1.236266055
SGK1	0.0746651	1.235414595

GLS	0.00862063	1.234568763
CLIP4	0.0958474	1.232341487
KRT80	0.115963	1.227124934
KLHL5	0.239107	1.226847585
HIST1H2BG	0.0564015	1.226785944
TCEAL8	0.0450866	1.226261894
TCEAL3	0.034457	1.221969856
SMAD3	0.00021952	1.221215114
DACH1	0.234196	1.219692056
NFIB	0.0576572	1.217615384
CLDN1	0.0827563	1.217547144
ROR1	0.080152	1.215927048
BAMBI	0.188325	1.215833884
LRRC16A	0.0121683	1.215765559
CD36	0.111104	1.214479217
TFAP2C	0.0729897	1.208635867
TGFB2	0.0525402	1.203620902
EEF1A2	0.15065	1.203389117
PDS5B	0.0410797	1.201627589
FKBP1B	0.0918256	1.198952977
CXCL5	0.232944	1.19725523
CDC42EP3	0.0368839	1.196745516
KRT20	0.386913	1.193033652
ITGB4	0.0498975	1.192673932
PARP6	0.000684548	1.18914139
CYorf15B	0.0809321	1.188881954
CLMN	0.0516636	1.183975535
SHANK2	0.0173911	1.183143466
IL1R2	0.182707	1.182374488
PMEPA1	0.063477	1.177669085
---	0.0164445	1.171930556
---	0.0777492	1.170187869
SEMA3C	0.00933775	1.169148943
ARL4C	0.216178	1.163962381
HOXB5	0.00762795	1.160933092
VIPR1	0.0154561	1.1592352
DACH1	0.322953	1.159215821
GLS	0.0101071	1.157910361
ABP1	0.305375	1.157341276
TMEM154	0.0447135	1.156293055
PPP3CA	0.00540104	1.153474613
TGFA	0.0822417	1.146798572
SORL1	0.0727023	1.14530573
CFLAR	0.000427513	1.145233982
FAM20C	0.0177869	1.142393831
DDX26B	0.0247535	1.139548078

LATS2	0.0721046	1.138572034
CSGALNACT1	0.0492936	1.138139481
ARL4C	0.271028	1.137411713
KRTCAP3	0.0592265	1.137228074
BMPR2	0.0131733	1.135153953
MALL	0.126444	1.134602103
IL8	0.208359	1.132464976
CFLAR	0.00241244	1.130548736
NAV1	0.022396	1.130179682
---	0.024804	1.129441293
EDN1	0.0385228	1.128761904
PJA1	0.013483	1.124758232
CYorf15B	0.0850946	1.12291784
SFTA2	0.0745249	1.122275136
SPAG16	0.145201	1.121751485
SHH	0.00766905	1.121492904
ZMYM2	0.0010733	1.120338668
ZNF846	0.00288693	1.120338668
RGL3	0.0218581	1.115865583
CYorf15B	0.072835	1.114746807
DACH1	0.180435	1.114453651
PRRX1	0.0375731	1.112172994
---	0.127969	1.111144853
SLFN11	0.0451795	1.111037992
---	0.0603853	1.110516927
HGD	0.0788655	1.109621324
RIN2	0.0864311	1.10646893
PDGFA	0.0719163	1.106060154
MAFF	0.0496704	1.104705674
TBC1D9	0.206907	1.10467884
KLK6	0.292406	1.1044306
TLR4	0.00986516	1.10435679
FERMT2	0.332214	1.103584921
TMED3	0.0291308	1.099241334
HBEGF	0.0175339	1.098870919
IRX3	0.0145782	1.098540833
YPEL5	0.00540049	1.098331964
IL18	0.163056	1.098170238
COL4A1	0.052417	1.096174128
MACROD2	0.000696158	1.094283371
NT5E	0.0590881	1.09301247
LATS2	0.0294447	1.092187137
TLE1	0.0362821	1.091002434
NME7	0.103877	1.090108184
SMAD7	0.0821609	1.088562265
COL4A5	0.0893409	1.088005871

ATP8B2	0.071778	1.086423727
GRAMD3	0.0217916	1.08627425
IL8	0.133035	1.083533364
TRNP1	0.117466	1.081878151
F3	0.128774	1.08171457
RBPMS	0.176073	1.080944128
AKR1C3	0.419192	1.080889567
CFLAR	0.00219997	1.080377965
FXR1	0.0119805	1.07992078
ALDH3A1	0.147077	1.078507367
FGF9	0.144943	1.075820115
TRIB2	0.221602	1.072634063
LOC554202	0.0248672	1.071796986
BCL9L	0.00112897	1.071776396
ADRB1	0.0495762	1.068877141
MCAM	0.0391756	1.06858827
C3orf14	0.309091	1.068409417
C8orf84	0.158268	1.067005334
GALNT3	0.203854	1.06658522
SMAD3	0.000998455	1.065220728
SCEL	0.225466	1.063516748
ELFN2	0.0666143	1.062750336
LOX	0.0253531	1.061637969
WIPI1	0.0485352	1.059528335
BDNF	0.142759	1.058385771
ZNF512B	0.0581242	1.057976998
PDGFB	0.000699241	1.056902551
RGS20	0.0741857	1.054508162
TUBB2A	0.167622	1.053834241
GAD1	0.0657009	1.052603697
GNAI1	0.234399	1.052360244
ALCAM	0.152224	1.051128441
GJC1	0.0617547	1.050898667
MARCH6	0.0138817	1.050822067
KLF12	0.0765106	1.050508663
---	0.0270422	1.049240421
LOC399959	0.191367	1.049003373
SAMD9	0.024858	1.046435181
FAM149A	0.0544978	1.046358344
CD44	0.114996	1.045254234
GAD1	0.0774025	1.044729835
CFLAR	0.000349254	1.037733627
DCBLD2	0.263391	1.037276784
MYOF	0.214624	1.036974484
KLK7	0.126769	1.035743543
FGD6	0.0310355	1.034462249

PMAIP1	0.132967	1.034377728
FGF9	0.0340531	1.033842313
B2M	0.0136755	1.033095219
SLC2A12	0.113277	1.032961265
BCAR3	0.0126441	1.032390059
CYP26B1	0.055193	1.032079679
KLF4	0.120508	1.032072624
DDX17	0.00494524	1.030547964
ORAI2	0.00412087	1.028243802
PSPC1	0.00153199	1.028045727
MTUS1	0.144578	1.027458415
OAS3	0.0664754	1.027132814
LOC100132999	0.0622474	1.026672601
FREM2	0.150137	1.025376134
FAM91A2 /// FLJ39739 /// LOC100286793 /// LOC728855 /// LOC728875	0.163224	1.022062373
PRKD1	0.0337691	1.021216749
WBP5	0.253847	1.021017707
IL1R1	0.0857042	1.020889736
CFLAR	0.00467018	1.020711981
LOC389831	0.0215931	1.020391966
TGFB2	0.114142	1.019289149
SFXN3	0.00118151	1.019096959
LHFP	0.0977934	1.019047127
MAP2	0.126056	1.018463262
MAP3K2	0.00269172	1.018035894
SERPINB9	0.11586	1.018000274
IL1R2	0.16067	1.01677442
MBP	0.014695	1.016596157
LEPR	0.0755491	1.01628236
LOXL2	0.0360836	1.015868615
HPGD	0.278864	1.014890848
OSBPL6	0.0344238	1.012418497
CFLAR	0.00547606	1.011796169
DAPK1	0.301103	1.009769957
LAT2	0.0139182	1.009289836
SNRPN /// SNURF	0.265995	1.008580111
AMOTL1	0.0910823	1.007934604
ARL4C	0.150738	1.007274453
INPP4B	0.0922416	1.007037586
ANXA1	0.333899	1.006362661
ARHGAP23	0.105235	1.005960428
SH3BGRL	0.226014	1.005529339
N4BP2L2	0.00932809	1.005076558
THBS1	0.136767	1.004731486

RAB25	0.233156	1.004141797
---	0.135272	1.002421694
MCAM	0.0759137	1.001074408
PAQR5	0.149899	0.999675357
MYOF	0.219565	0.998036599
BICC1	0.140265	0.997538106
RFTN1	0.0472359	0.997523654
SERPINB5	0.308663	0.995470047
SERPINB2	0.0585786	0.995397683
HOXA5	0.19678	0.995347027
TUSC3	0.174361	0.994398697
PTGR1	0.0611293	0.994210405
DHRS3	0.0259981	0.994152464
LPCAT2	0.104355	0.993261328
BMPR2	0.00956728	0.992942414
ZNF185	0.0907448	0.992521919
GALNT10	0.015669	0.991963487
CNTNAP3	0.0830705	0.990933084
CYBRD1	0.183237	0.989611349
CLMN	0.0254349	0.989051789
PLEKHB1	0.219613	0.988324763
BMP4	0.180396	0.987037029
LOC100506918	0.0736357	0.982926057
ZBED2	0.316308	0.981947612
CDC42EP3	0.0704139	0.981392378
SPRR1A	0.127709	0.979835123
GPX3	0.210763	0.978605678
ALCAM	0.129268	0.978525142
SLC44A1	0.0542822	0.975204615
---	0.0178438	0.975167922
MYH10	0.304507	0.974521971
CLDN1	0.314676	0.973934491
---	0.0441338	0.973890421
CPNE8	0.00534401	0.973714125
MAGED1	0.0715515	0.97074324
ZNF655	0.168236	0.970264679
GPRC5A	0.221291	0.968769074
THBS1	0.244765	0.967958
B4GALT1	0.164319	0.967544914
C5orf41	0.0125426	0.965078692
NTN4	0.251537	0.963917962
DDX17	0.00895231	0.963126356
FOSL2	0.0129816	0.962349123
C8orf47	0.026139	0.962001083
RDX	0.203207	0.961808515
HNRNPH1	0.0143554	0.961756665

LOC100499466	0.0422251	0.961564064
HBEGF	0.0183171	0.960608084
PLCB4	0.0592147	0.960422744
JUP	0.0280553	0.959488264
HOXA10 /// HOXA9	0.239239	0.959332459
RBMS1	0.25966	0.958924317
ZMYM2	0.00726632	0.958300751
LOC100507025	0.0556488	0.957936879
ST3GAL1	0.199209	0.957520914
TFPI2	0.0435072	0.957015651
GNAT3	0.0127271	0.955372325
---	0.0211996	0.953146018
APOH	0.056659	0.95303424
PRAC	0.311428	0.952840471
HIST1H2AC	0.147792	0.9526914
NAT14	0.00960253	0.951863775
SRGN	0.101115	0.950557997
---	0.00369131	0.950319098
AHNAK2	0.247329	0.949826244
HS3ST1	0.181348	0.948301812
ARHGEF26	0.0873663	0.942473218
FLJ90757	0.0385749	0.941587129
---	0.138755	0.940114117
MMP7	0.327592	0.939241625
ZNF514	0.0104194	0.938473998
NTSR1	0.118911	0.93652299
---	0.0103277	0.935550267
C3	0.0729206	0.935059887
COL27A1	0.134179	0.933459333
ARHGEF37	0.0117196	0.93238627
KRT6A	0.0841778	0.931720873
PPP1CB	0.0704988	0.930003032
GABRP	0.0989057	0.929351685
PRRX1	0.0294627	0.927782702
---	0.012733	0.927661357
PARP6	0.00799054	0.927251742
---	0.0874604	0.926462525
SSTR1	0.077721	0.925847548
SLC7A8	0.104596	0.925680472
IFIT3	0.368924	0.92314918
PRR16	0.0172291	0.923019836
ANXA10	0.257205	0.921352539
SDC4	0.0488749	0.919050179
VNN1	0.0586036	0.918783112
GPX8	0.166102	0.916927536
HOXB6	0.104129	0.915666215

SNCA	0.0560993	0.91539852
UBE2E2	0.25317	0.91500071
FXR1	0.0277061	0.914709934
EFNA1	0.0223008	0.912465924
LGALS8	0.0162108	0.912297291
CAV2	0.355845	0.910110941
DNM1	0.117002	0.909757746
GRAMD1C	0.0447835	0.909657914
SPINK1	0.440146	0.909550396
GJB6	0.0321047	0.909028047
NR1D2	0.00947141	0.908882063
LOC728190	0.0149375	0.908459395
RPS4Y1	0.348325	0.908251857
HGD	0.0795973	0.906990619
---	0.189251	0.90491949
SULT2B1	0.0227096	0.904880965
---	0.00746403	0.904572728
CYP3A5	0.280579	0.904503366
C10orf47	0.127275	0.904372339
LOC645638	0.000901869	0.903362261
MN1	0.0487678	0.902961119
HIST1H2BG	0.0455611	0.902829952
PLAUR	0.095389	0.902822235
PCMTD1	0.0699375	0.900876466
ANTXR1	0.297257	0.900459174
KLK10	0.0159832	0.900443716
---	0.0733566	0.900374155
NAV1	0.0274317	0.899778883
DCBLD2	0.278657	0.898471516
---	0.00385996	0.898022561
FAM129A	0.317635	0.897054509
ODAM	0.151388	0.896318355
ASNS	0.0546539	0.895884236
NR4A2	0.0419772	0.895201784
RABGAP1L	0.0436456	0.894480206
---	0.158318	0.893564135
---	0.0733899	0.89233662
TGFB2	0.0880683	0.891263632
DIS3L2	0.0104093	0.889823952
NHS	0.258585	0.889629291
ABCA5	0.0366997	0.888468563
ITGB6	0.0321646	0.886206747
WFDC2	0.0843379	0.885644643
LAMP3	0.235761	0.885558746
EDIL3	0.293327	0.885176055
ALDH7A1	0.041393	0.884488518

C19orf33	0.217863	0.883745923
HIST2H2BE	0.0983837	0.882071727

Table A.2: List of the 500 top probesets upregulated in strict myofibroblast-recruiting cell lines (recruitment score 5) compared to strict non-recruiters (recruitment score 1), sorted by $\log_2$ of the expression ratio between them. P-values uncorrected for multiple comparisons are given.

Gene Symbol	p-value	$\log_2$(Ratio)
GAL	0.0186307	-2.580180687
MME	0.00919268	-2.548682842
PLIN2	0.00242226	-2.50043828
NETO2	0.00194075	-2.352753655
OLFM4	0.140403	-2.31076168
WT1	0.00021078	-2.25557709
NETO2	0.00406165	-2.162008333
SLC6A20	0.0247771	-2.115791278
RASGEF1A	0.00916585	-2.013142035
H19	0.0902087	-1.950369232
PLBD1	0.0278046	-1.945948958
SMPDL3A	0.00466676	-1.849871897
ABHD12B	0.0860167	-1.803181477
LOC100510557 /// MUC6	0.00258444	-1.789267025
LOC100507192	0.0368176	-1.755185991
MME	0.0119079	-1.751824668
PAPSS2	0.150643	-1.737735237
ELAVL2	0.0274356	-1.681533907
IL33	0.094585	-1.662213564
FAM84A	0.201897	-1.648389761
PAG1	0.0357701	-1.639386259
IFI16	0.0621499	-1.631122161
LOC100505875	0.0118701	-1.630661948
XK	0.0156125	-1.623371768
FAM84A	0.120301	-1.620019795
SRSF6	0.0490637	-1.603638811
DDC	0.183889	-1.599125481
TNFSF9	0.00141463	-1.584362464
PARVB	0.00345774	-1.572346632
BEX5	0.0441149	-1.571583142
PAQR8	0.0188453	-1.561055687
GPA33	0.160504	-1.55714884
C2orf89	0.0528046	-1.538458868
PARVB	0.00404853	-1.537474366
CPVL	0.0591222	-1.524906814
SLC16A3	0.0383909	-1.522990041
GSPT2	0.169357	-1.514746307
H19	0.131109	-1.513238267
---	0.0148832	-1.467469582
MEST	0.0425948	-1.461990342

SLC39A8	0.0592576	-1.45667042
GOLIM4	0.00921947	-1.454545559
SATB2	0.0761521	-1.451500243
TP53	0.0902268	-1.443818801
NRN1	0.125222	-1.440596293
SLC16A3	0.0449265	-1.435389653
UST	0.0594385	-1.421832218
PAPSS2	0.187245	-1.420144049
PIGR	0.208751	-1.417074098
TCEA3	0.0911738	-1.416800588
LOC653602	0.141213	-1.410167495
NUPR1	0.0920439	-1.407375052
SOX2	0.0739605	-1.40312209
TFF2	0.0224795	-1.398957717
AKAP2 /// PALM2-AKAP2	0.123875	-1.387921203
LCP1	0.00265424	-1.379308499
NPTX2	0.0498099	-1.372430514
IFI16	0.0707119	-1.358313407
PARVB	0.0139928	-1.352602593
VIM	0.153365	-1.350559284
CDC42EP5	0.146961	-1.345289512
CEL /// LOC100508206	0.192567	-1.332182338
FAM46A	0.00869696	-1.329290165
PARM1	0.15893	-1.325575507
EGF	0.0129709	-1.325087447
HMGN3	0.115931	-1.310922741
BCL11A	0.0594046	-1.307879943
SLC6A20	0.068566	-1.300359534
BCAT1	0.163301	-1.295347554
IFI16	0.135394	-1.290169447
OXCT1	0.0719145	-1.285498612
PKP2	0.0163073	-1.281982579
BCL11A	0.1123	-1.279532399
NPL	0.00537129	-1.277799806
GRIN2D	0.0270689	-1.277131826
CEACAM5	0.396163	-1.273534951
---	0.00854136	-1.271447294
CEL /// LOC100508206	0.221188	-1.271099064
FZD3	0.00178984	-1.268868905
RAMP1	0.0178845	-1.268073006
SLC26A3	0.284471	-1.258176671
MYB	0.172444	-1.247559673
NFIA	0.00340051	-1.246967179
PAH	0.078579	-1.244148459
GIPC2	0.162005	-1.241898123
DPEP1	0.189017	-1.228979876

TUBAL3	0.0442203	-1.227607561
C13orf18	0.328339	-1.224661159
SAMD5	0.16139	-1.223084119
PCK1	0.290829	-1.222360202
PRTFDC1	0.139934	-1.222070736
---	0.0250542	-1.221808248
AGMAT	0.023756	-1.219605904
LOC654433	0.131473	-1.219575666
VAV3	0.249783	-1.212433804
---	0.0228607	-1.211959157
TNS4	0.0767101	-1.208207595
GATA6	0.213661	-1.206701706
C9orf152	0.26107	-1.204505661
PAG1	0.0197069	-1.203731188
GAP43	0.000180437	-1.202993665
PGAP1	0.0239946	-1.201231117
---	0.108218	-1.196806034
TRBC1 /// TRBC2	0.00299141	-1.196696904
FZD3	0.000552488	-1.19329155
EPHB3	0.0995063	-1.190621778
CNN3	0.168581	-1.183444527
ZNF253	0.0122979	-1.179887224
PAPSS2	0.0508139	-1.178309375
FSCN1	0.241819	-1.16830108
SMOC2	0.295367	-1.167302763
NINJ2	0.0365541	-1.165046131
TP53	0.0778552	-1.164033898
---	0.16854	-1.163490884
RNASE1	0.117384	-1.162880237
RASL11A	0.0103656	-1.16175332
WT1	0.000992862	-1.159930804
DOCK11	0.12653	-1.15803332
GEM	0.0494178	-1.154689005
ID3	0.217795	-1.154332521
VIL1	0.297666	-1.151422926
MYL9	0.0538122	-1.150013547
EIF5A	0.141994	-1.141761134
PSIP1	0.00479053	-1.139604463
NRXN3	0.178538	-1.139582213
AUTS2	0.272318	-1.135763529
STRN3	0.0228719	-1.134926864
C2orf82	0.0806952	-1.132935401
MEX3D	0.00127419	-1.132239511
CPPED1	0.0243435	-1.13215413
PAG1	0.0643206	-1.12917527
IQGAP2	0.247114	-1.128342416

NOS3	0.0887045	-1.127320934
GOLIM4	0.0159581	-1.126980601
SLC17A9	0.0128573	-1.124701178
GTF3C6	0.00172077	-1.123930641
PBK	0.019609	-1.123320793
ITPR2	0.0736366	-1.12256984
EN1	0.0213546	-1.120727107
MTA2	0.0173699	-1.116138364
CDX1	0.105005	-1.110937713
---	0.00499666	-1.110426777
ZNF260	0.0541369	-1.105581882
RNF144A	0.0531402	-1.105044908
ANTXR2	0.0758443	-1.104197951
LOC100216479	0.0628328	-1.101028605
IGF2BP3	0.20216	-1.098820686
FAM171A1	0.123385	-1.097162332
---	0.161811	-1.095687783
---	0.00115797	-1.094744616
MGAT4A	0.0917054	-1.094063822
METTL7A	0.263926	-1.090618579
RTN1	0.00749325	-1.089156838
STOM	0.187669	-1.087107982
STK31	0.0329303	-1.086685078
LOC100287015	7.02E-06	-1.083132066
FAM46A	0.0349021	-1.082948685
---	0.0212838	-1.07388517
UGGT2	0.0258591	-1.073769769
NLRP2	0.253369	-1.069106765
PEG10	0.385282	-1.067920682
AKAP2 /// PALM2-AKAP2	0.134436	-1.067760393
LCN15	0.0373255	-1.066753705
ZNF711	0.0319582	-1.065789998
CALB2	0.0952244	-1.064808826
LOC645984	0.00112099	-1.061637417
CCND2	0.456205	-1.061526002
EPPK1	0.013795	-1.061447715
BCAT1	0.25463	-1.061273091
SLCO3A1	0.112001	-1.061233954
SELENBP1	0.300186	-1.060969055
CALB1	0.431015	-1.059305517
UPP1	0.0393251	-1.058995884
TUBE1	0.0343109	-1.058791502
GSTA4	0.0507852	-1.057346627
STOM	0.123755	-1.057178503
ERLIN2	0.00190268	-1.056785289
TFF3	0.290632	-1.055501336

MT1M	0.0967952	-1.053565568
PLLP	0.0713207	-1.052610628
SHC3	0.02557	-1.051823801
---	0.0644119	-1.050245458
AGR2	0.480321	-1.048633028
CEBPA	0.0759455	-1.048245119
UBXN8	0.0027211	-1.047899073
KLHL13	0.200045	-1.047257914
ASB4	0.161199	-1.044931263
CCDC88C	0.00218295	-1.04005777
PAQR8	0.0487029	-1.040025138
LGR5	0.345054	-1.036993533
GPC4	0.12495	-1.03661466
ERLIN2	0.00179099	-1.033832402
TMEM35	0.00323466	-1.033794002
ZNF681	0.0221341	-1.033188605
ASB4 /// LOC285986	0.188617	-1.029302401
BMP7	0.289882	-1.028030897
IGF2BP3	0.217235	-1.027695548
FZD3	0.00939234	-1.027245596
GSR	0.00116233	-1.024871708
LGALS4	0.521474	-1.023366549
RUNDC3B	0.150121	-1.022616023
SPON1	0.272384	-1.022472412
HNRNPL	0.0283437	-1.020682873
---	0.0494147	-1.020287775
C10orf99	0.406804	-1.016328142
CLRN3	0.448697	-1.015910895
CDH17	0.483503	-1.014581179
ELOVL5	0.406661	-1.014496655
STC1	0.0367274	-1.014458767
CDCA2	0.00290549	-1.012152365
C6orf105	0.0486303	-1.011864323
C14orf139	0.118444	-1.011134295
ASCL2	0.279624	-1.009675345
FADS2	0.0271967	-1.004241983
KCNV1	0.0974622	-1.004210151
ITPR2	0.0900871	-1.00357365
CDC42SE1	0.0195727	-1.000401125
TMEM144	0.0281031	-1.000346288
ACPL2	0.0425667	-0.999890359
ZC3HAV1L	0.0525632	-0.99726148
RNFT2	0.0140532	-0.996705763
NUP210	0.00926721	-0.996069686
MT1X	0.189213	-0.996006382
IQGAP3	0.00514078	-0.995724423

ACSL6	0.142687	-0.995695655
IGFL2	0.0863432	-0.995246945
GHR	0.0187402	-0.99483575
GSR	0.00821384	-0.991587693
AGPAT5	0.00149012	-0.990093919
EIF5A	0.0941524	-0.988561612
STAMBPL1	0.0376712	-0.987537164
ANPEP	0.123701	-0.986179032
KBTBD8	0.00541404	-0.984185609
VAV3	0.301674	-0.983908804
ST6GALNAC1	0.349795	-0.982733688
CYP2J2	0.136065	-0.982323193
KIAA1967	0.0034744	-0.981724764
KRR1	0.00414708	-0.980292393
MAGEA6	0.489266	-0.979154345
MCM4	0.0397151	-0.976844025
SSBP2	0.0422227	-0.976733291
S100A3	0.0842261	-0.97539096
COL6A1	0.135636	-0.973205605
FAM13C	0.00368145	-0.972565648
SPATA18	0.0700319	-0.969810847
CKB	0.170571	-0.967955596
SLC16A1	0.0718374	-0.967334888
RGMB	0.08859	-0.965897
KRR1	0.00871815	-0.96375692
ZNF711	0.131441	-0.963256149
ZEB1	0.0691332	-0.962640271
SLC16A1	0.0801579	-0.961917863
VDR	0.0628825	-0.960131668
THNSL1	0.00393641	-0.959808928
HGF	0.00239133	-0.959783673
ACOX2	0.178419	-0.959643377
EIF5A	0.0791944	-0.959082327
ZNF260	0.0867154	-0.958678506
XIST	0.422825	-0.958347681
RGS2	0.255661	-0.958266389
SRD5A3	0.0136937	-0.956378323
EPPK1	0.187943	-0.953732468
SATB1	0.258999	-0.95268777
FABP5	0.109518	-0.952629133
PTPRG	0.128958	-0.951309032
UTS2	0.0302257	-0.949237823
PEG10	0.298425	-0.948435784
MCM4	0.0184298	-0.947943085
CHAC2	0.020511	-0.947164019
IFITM1	0.280575	-0.947080572

SAMD5	0.220985	-0.946060134
EPB41L2	0.119363	-0.945996207
FAR2	0.202624	-0.944865443
EPPK1	0.215842	-0.94448224
VAV3	0.250139	-0.943921503
BLNK	0.216666	-0.942850602
IRF8	0.175407	-0.941306686
PREP	0.0123626	-0.940841341
UTS2	0.0189445	-0.938409032
SRD5A3	0.0165714	-0.936693112
PDK1	0.0342599	-0.935867655
SMARCA4	0.0160578	-0.935710348
ADH1C	0.196294	-0.934571069
FBXO6	0.152748	-0.933989363
VIL1	0.38068	-0.933622816
CYP1B1	0.278034	-0.933253607
RAD54B	0.00270211	-0.933162697
CFD	0.088963	-0.933052511
INTS10	0.000288783	-0.932818393
SCPEP1	0.0543767	-0.930825796
ITCH	0.0469434	-0.929954212
---	0.0877591	-0.929943217
TPST2	0.00455556	-0.928684896
ZEB1	0.239079	-0.927831072
FAM54A	0.0275334	-0.927710315
CASP2	0.0182785	-0.927183491
PALM2	0.00960354	-0.926160577
EPHX2	0.0572723	-0.925897424
MFAP2	0.0786154	-0.925875496
NTRK1	0.000219724	-0.925456202
MAP7D2	0.312783	-0.923698137
XIST	0.521345	-0.92362151
SMARCA4	0.0142148	-0.92167436
PCDHB14	0.095914	-0.921010415
SUV39H2	0.00546567	-0.919997326
C13orf18	0.376488	-0.919937273
CD163L1	0.069014	-0.918886749
MYH4	0.037278	-0.917180256
---	0.0809308	-0.917035869
KIAA1274	0.0659181	-0.916412175
FAM13A	0.233509	-0.915859519
STAMBPL1	0.0379022	-0.914246322
ZNF567	0.0433014	-0.912822331
RHOU	0.255303	-0.911806831
TFF1	0.251683	-0.911630413
PSMG1	0.0031242	-0.911285782

NUTF2	0.00723814	-0.911141984
TNFRSF10D	0.0358443	-0.909059875
SLC39A8	0.151514	-0.907705949
MCM4	0.0333928	-0.907126856
PROSC	0.00339393	-0.906115355
AGR3	0.500889	-0.906018029
ASB4	0.0908998	-0.90352493
UBE2J1	0.00344639	-0.901103486
EPB41L2	0.11701	-0.899127279
GIPC2	0.238149	-0.899116517
LONRF3	0.0564814	-0.897215575
KIAA1211	0.182305	-0.893868805
ASF1A	0.00424043	-0.893514989
MGST1	0.1388	-0.893169299
EEPD1	0.0130151	-0.893091597
CXorf57	0.15046	-0.892861196
ANTXR2	0.106062	-0.891525056
HS3ST3B1	0.102609	-0.890626067
XPNPEP3	0.00110431	-0.89052978
ASF1A	0.00151812	-0.889727638
NRTN	0.0940421	-0.888955329
ESCO2	0.00207708	-0.886910309
KIAA1324L	0.119759	-0.886203502
PLAGL1	0.349394	-0.88484687
SLC27A5	0.0189284	-0.883355781
PIGR	0.0248299	-0.883233368
KCNRG	0.00105622	-0.882629434
SATB2	0.185754	-0.882227838
HSPC157	0.00647487	-0.879754242
ARG2	0.18388	-0.879698495
ME2	0.0715129	-0.879358753
---	0.0125486	-0.879260562
SLC17A9	0.0530859	-0.879069505
HHEX	0.13477	-0.878353268
C2CD4A	0.26104	-0.876720516
NUP43	0.0219722	-0.876572176
AMFR	0.182445	-0.87603457
BNIP3	0.500058	-0.875356889
MT1G	0.176062	-0.875240445
---	0.107881	-0.874359477
THBS2	0.01472	-0.874015702
PEX3	0.00213486	-0.873539841
NGEF	0.0401059	-0.87298751
TXNIP	0.261707	-0.872826343
HNRNPC	0.00165679	-0.872245235
MLLT3	0.183191	-0.871978532

ZNF438	0.00186198	-0.871566692
SLC16A10	0.137501	-0.867307352
CBX2	0.0283548	-0.866949465
ME2	0.0933221	-0.866244478
MDN1	0.00194294	-0.865542465
LRP12	0.0995483	-0.865463608
SMARCA4	0.0382278	-0.865319049
RAD54B	0.0141395	-0.864974793
SOX2	0.0229031	-0.863897882
HGF	0.0140104	-0.863275742
SLC16A1	0.149851	-0.862950341
CMAS	0.00677445	-0.861961467
SMARCA4	0.0209002	-0.861279879
---	0.156844	-0.861078086
ETS2	0.261493	-0.860609092
PLEKHG4	0.0498915	-0.858807896
SAMD5	0.396043	-0.858104261
FAM84A	0.222858	-0.857900297
STC1	0.0551962	-0.857881994
TNFRSF10D	0.0170696	-0.857348692
BCL11A	0.15478	-0.85691749
---	0.00459061	-0.856880909
---	0.230638	-0.855961464
ETNK1	0.0149728	-0.855663815
PGAP1	0.0341228	-0.854205178
HES6	0.103656	-0.853407335
RGMB	0.0871386	-0.85272196
PTGFRN	0.0937095	-0.852146288
AKAP2 /// PALM2-AKAP2	0.221636	-0.851977017
---	0.0903224	-0.851971808
NPL	0.00717248	-0.851482337
EPHB3	0.0817765	-0.850420649
ZNF506	0.0468566	-0.849661294
---	0.0188218	-0.848499629
FUT10	0.00152174	-0.848382735
CASP2	0.0333929	-0.848374942
CSTF2	0.00219859	-0.848330785
CPPED1	0.0743208	-0.848247669
FKBP5	0.0944411	-0.847450517
PTPRG	0.225263	-0.847338899
FYN	0.124295	-0.845523076
UGT2B15	0.0425913	-0.844520165
IRAK1BP1	0.018081	-0.843958114
---	0.0146387	-0.84331345
PFAS	0.0217753	-0.842079294
GAS2L3	0.0293303	-0.841208001

PCDHB2	0.0595184	-0.839345664
PEX3	0.00786784	-0.838555994
C12orf48	0.00942274	-0.838532775
HNRNPH1	0.0801868	-0.837501196
SIAE	0.190885	-0.834035337
NES	0.143616	-0.833438793
CHML	0.030031	-0.83075996
FANCD2	0.0167104	-0.829267315
FAM40B	0.00832705	-0.829175038
CLEC11A	0.163716	-0.828542068
---	0.131974	-0.82813988
PRLR	0.201106	-0.828114266
KCNS3	0.248383	-0.827535527
SOBP	0.210907	-0.827333279
TXNIP	0.291102	-0.826857212
---	0.0389843	-0.826849535
DLX6	0.125125	-0.826496427
REPS1	0.00127464	-0.824739851
FAM50B	0.0210938	-0.824420471
CASP2	0.031198	-0.824412806
SLC16A9	0.230336	-0.824333611
CLCN5	0.0214243	-0.824142029
MGST1	0.14375	-0.823855979
SLC17A9	0.0325528	-0.822885877
TCP1	0.0012794	-0.822745522
BCKDHB	0.0180154	-0.822546496
MTHFD2L	0.0569563	-0.821388606
---	0.0672309	-0.821202512
MGAT4A	0.130543	-0.820917045
CLEC11A	0.12613	-0.820578126
PKP2	0.0562567	-0.819658604
NR1I2	0.336393	-0.817902699
CHORDC1	0.0273443	-0.816484267
ARID5B	0.182618	-0.816250537
CDC40	0.00983541	-0.815818745
C6orf192	0.0307056	-0.815026617
RHOBTB1	0.173544	-0.814448028
C6orf141	0.175989	-0.813319434
RRAS2	0.219438	-0.811773805
HOXC10	0.0947593	-0.81163959
C6orf203	0.00720498	-0.810667543
ITM2A	0.176029	-0.810533432
---	0.00815125	-0.810285485
C10orf99	0.468552	-0.809984465
---	0.000595394	-0.809582358
C1orf43	0.0412984	-0.808374185

PODXL2	0.00741289	-0.808091245
MGST1	0.133968	-0.808086193
HSF2	0.011612	-0.807884128
STK17A	0.121905	-0.807106439
ITGA9	0.259081	-0.807005472
IFITM1	0.402586	-0.806679899
PPIL4	0.0149486	-0.806601671
CPT1A	0.0735579	-0.806558774
CISD1	0.00795899	-0.806435137
ARHGAP4	0.0755773	-0.806382152
TIMP3	0.234143	-0.805943215
TSPAN12	0.26879	-0.803020371
GALNT5	0.291275	-0.802826564
LEPROTL1	0.00232989	-0.80259252
ATP6V1B2	0.00338776	-0.800942708
TM7SF3	0.0392974	-0.800666247
CLDN10	0.0207376	-0.800477782
SOD3	0.0156039	-0.800105949
NQO1	0.131827	-0.799164233
ZNF738	0.209099	-0.798890622
TPK1	0.0711205	-0.79807512
SLCO3A1	0.0921341	-0.796866516
TRMT11	0.00829212	-0.79624255
ACSL1	0.138904	-0.79594695
HSPA1A /// HSPA1B	0.116361	-0.795401
THAP9	0.00271555	-0.795210717
CCDC25	0.00296859	-0.794802696
ND6	0.0424257	-0.794730115
ZNF253	0.0135814	-0.79459998
TMPRSS2	0.339489	-0.794024521
SYBU	0.197044	-0.793246765
HNRNPL	0.119489	-0.792711833
TRBC1	0.00394913	-0.792401963
LRPAP1	0.0195161	-0.792322007
HMGN5	0.266916	-0.792042197
ATP2C1	0.0235085	-0.79155266
SPON1	0.297066	-0.791387854
RTN4IP1	0.00699388	-0.790945965
SUPT4H1	0.0596079	-0.790476763
NUP210	0.132286	-0.789424109
XIST	0.424243	-0.789037663
---	0.0797505	-0.788584033
PIK3R3	0.0985898	-0.788424549
MARCKSL1	0.0323448	-0.787896383
DMD	0.313654	-0.78782664
NT5DC1	0.0181802	-0.787537739

Table A.3: List of the 500 top probesets upregulated in broad myofibroblast-recruiting cell lines (recruitment scores 4 or 5) compared to broad non-recruiters (recruitment score 1 or 2), sorted by $\log_2$ of the expression ratio between them. P-values uncorrected for multiple comparisons are given.

Gene Symbol	p-value	$\log_2$(Ratio)
LYZ	0.011742	2.724187487
SMARCA1	0.000240127	2.610366985
LYZ	0.0364663	2.496255249
SMARCA1	0.000546532	2.453201482
SERPINA1	0.00579685	2.346032134
TACSTD2	0.0333821	2.338681125
PDGFC	0.00594745	2.333148908
DKK1	0.0166695	2.283776607
SERPINA1	0.00111711	2.270935531
GPR37	8.63E-07	2.262628591
KCTD12	0.00454895	2.11353382
RAB27B	0.000337032	2.07890695
HDGFRP3	0.016242	2.043151993
FSTL1	0.00134241	2.017686783
GSN	7.08E-06	1.982579296
TUBA1A	0.0126819	1.952840471
SMARCA1	0.000389677	1.947404335
ZNF655	0.0267188	1.881186904
TFPI	0.0249969	1.848297437
OSMR	0.000614249	1.837769704
PLK2	0.016514	1.792588856
SLC40A1	0.0324214	1.792222343
DAPK1	0.0183046	1.779011719
HSPA1A	0.115234	1.736700114
HDGFRP3	0.0146611	1.730965952
OBSL1	0.0152174	1.726730966
SLC16A7	0.000314192	1.6954772
MAP2	0.00110762	1.695299015
EFNA5	0.000792198	1.6885698
CXCR4	0.0202297	1.688305708
C3orf14	0.0369826	1.686137427
EFNA5	0.000481059	1.684428282
RDX	0.067233	1.670006377
ITGB8	0.011025	1.662519987
HDGFRP3	0.0148165	1.660253236
AMIGO2	0.0460093	1.639898937
GALNT14	0.000283697	1.615755272
MLPH	0.00135923	1.610907901
EEF1A2	0.0126775	1.609996021
OSMR	0.000851119	1.607981229

SCEL	0.0176216	1.606475393
PROM1	0.10328	1.602784663
ETV1	0.00490275	1.60224782
TFPI	0.0178734	1.59277036
ARMCX3	0.00311115	1.587489811
TUSC3	0.00557371	1.572991502
CPE	0.0234161	1.571147684
IL1R1	0.000730411	1.567457832
HOPX	0.000118547	1.560710064
GJC1	0.0642134	1.532007688
ARMCX3	0.0167764	1.518268228
TIMP2	0.0444778	1.518157421
KLHL5	0.0578584	1.516337952
KCTD12	0.0049976	1.504467853
MTSS1	0.00309626	1.492063394
PDGFC	0.00382315	1.487985341
ZNF512B	0.000641615	1.487095266
CPNE8	0.000650083	1.485040339
STEAP4	0.00409104	1.473651856
ANXA10	0.0192361	1.468604169
FRAS1	0.0593771	1.467018572
MLLT11	0.00694262	1.460485712
TFPI	0.0106463	1.459410634
ABCG1	0.00581149	1.448980218
HDGFRP3	0.0150794	1.442513608
GCNT2	0.00235913	1.438559007
CXCL5	0.0665453	1.437217084
SLITRK6	0.000760874	1.433285313
SLITRK6	0.00429621	1.428340741
SLITRK6	0.000807414	1.42785822
UCHL1	0.0219521	1.420406944
AQP3	0.0367635	1.413388278
MAGEA12	0.0868884	1.408592311
TRIB2	0.0367896	1.408309698
TIMP2	0.09864	1.405600329
CXCL5	0.0683272	1.402252843
TFPI2	0.0147999	1.394673274
---	0.00362584	1.389698961
TBC1D9	0.0403808	1.377851162
KYNU	0.008435	1.374816926
HOXB2	0.00189476	1.37030938
ARMCX6	0.0154776	1.364785319
LUM	0.00335021	1.364040078
CRIP2	0.000337977	1.361088954
CBS	0.0213776	1.359616747
GPX8	0.059585	1.359594259

BASP1	0.0199612	1.348595031
CRNDE	0.0238035	1.343657963
PMEPA1	0.05941	1.341649933
GRB14	0.0319981	1.333692783
MUM1L1	0.0249437	1.333177539
FAM129A	0.0535023	1.33152752
DUSP4	0.0660944	1.330277291
SAMD9	0.020262	1.327422942
CLIP4	0.0198246	1.324136612
AMOTL1	0.00420158	1.321795361
SPP1	0.000401673	1.317963898
SPAG16	0.0268234	1.310043407
TGFB2	0.00637974	1.304581131
CBS	0.0428819	1.301101789
SRGAP2P1	0.00779011	1.29818907
PXDN	0.00448957	1.298018931
GSN	2.19E-05	1.296592493
CAV1	0.205675	1.294147247
NNMT	0.0402154	1.288181352
FERMT2	0.140234	1.284052113
MAGEA3	0.236505	1.282255935
LOX	0.000437189	1.280231816
CDC42EP3	0.00852166	1.277562288
MEIS2	0.0567445	1.263310896
TRNP1	0.0185216	1.251640329
NHS	0.0390203	1.249530292
GPX3	0.0998075	1.246578343
LPAR1	0.0222137	1.244125943
AHNAK2	0.0485855	1.24334005
CPE	0.0264676	1.240454775
SEMA3C	0.0279658	1.239220825
---	0.000786091	1.237288369
SGCE	0.070231	1.233581332
NT5E	0.0669104	1.23146313
TUSC1	0.0183574	1.23024606
KIAA0495	0.000508799	1.228042889
RDX	0.0688816	1.227778036
MAGEA6	0.257554	1.224861438
PHLDB2	0.181224	1.224083505
ANXA1	0.124823	1.223410193
TRIB2	0.0174042	1.222730385
CMTM3	0.00795363	1.221864711
SLFN11	0.00436456	1.219642499
NAP1L5	0.00764547	1.217720839
DPYD	0.0100233	1.217640197
MATN2	0.0500424	1.217485105

TUSC3	0.00517514	1.216535573
DNER	0.00048598	1.210893782
PTGER2	0.0103238	1.206355656
NTS	0.00512526	1.203520676
RAB27B	0.00090524	1.201200988
TLE1	0.00290136	1.198261544
FAM20C	0.00131045	1.196984664
ABLIM3	0.00513062	1.196418197
OBSL1	0.031419	1.196348947
FAM107B	0.0336524	1.19311568
MAGED1	0.00422875	1.191493167
CD109	0.155077	1.188976875
CD36	0.0425257	1.183873934
TGFB2	0.0172484	1.181872207
FN1	0.139722	1.178568054
GULP1	0.0881862	1.176814227
CSAG2 /// CSAG3	0.0166084	1.169937825
MAGEA2 /// MAGEA2B	0.116562	1.168430236
IFIH1	0.00798539	1.16690822
HOXA3	0.0141744	1.165551823
TRPS1	0.0127791	1.161268561
PMEPA1	0.0175773	1.155781609
CRNDE	0.0284588	1.153150301
IAH1	0.00366922	1.142661763
FN1	0.131573	1.14135431
CXCR4	0.0312276	1.140719769
DUSP10	0.0118026	1.138775158
VGLL1	0.123764	1.138637561
CD36	0.0671336	1.130311498
SCEL	0.0371806	1.125360153
LOX	0.00220561	1.121930476
DPYSL3	0.0346778	1.121804522
YPEL5	0.000193526	1.119960355
NT5E	0.049201	1.114766793
SDR16C5	0.0932402	1.112760174
INHBA	0.0291136	1.111839262
OGFRL1	0.150366	1.108838888
MAP1B	0.110596	1.103853444
PCDH7	0.112347	1.102053383
FLRT3	0.0468107	1.099685706
SYT13	0.103988	1.098783352
BDNF	0.0470834	1.098608203
MAGEA2 /// MAGEA2B	0.13038	1.096956717
SRGN	0.0139728	1.096828563
ATP8B2	0.0180099	1.095330346
TLE1	0.00285382	1.094127948

---	0.00611436	1.091869055
CDC42EP3	0.0130073	1.091212366
PXDN	0.00941569	1.089715081
PEG10	0.24706	1.08709618
PCDHA1 /// PCDHA10 /// PCDHA11 /// PCDHA12 /// PCDHA13 /// PCDHA2 /// PCDHA3 /// PCDHA4 /// PCDHA5 /// PCDHA6 /// PCDHA7 /// PCDHA8 /// PCDHA9 /// PCDHAC1 /// PCDHAC2	0.0192223	1.083131656
PMEPA1	0.104323	1.082886491
ARMCX2	0.0135544	1.078555186
ZNF512B	0.00133726	1.07435472
SFTA2	0.0259273	1.073875062
LPHN2	0.0202362	1.073038705
---	0.000798856	1.072249891
SLC2A10	0.128931	1.069269087
LEPR	0.0152379	1.066151203
MUC16	0.00458729	1.064765611
ZNF185	0.0182474	1.064068865
CUEDC1	0.00104975	1.063710013
ALCAM	0.0586553	1.062750336
---	0.0290057	1.062598386
ETS1	0.115402	1.061485902
MAFB	0.00981648	1.059341432
EDN1	0.0739363	1.058711321
ZBED2	0.15771	1.058281857
MYH10	0.144742	1.058261073
FN1	0.149039	1.057907703
GPX3	0.0779521	1.055306733
ANXA1	0.00660966	1.054834588
CXCR4	0.0235437	1.054744308
IRX3	0.00237259	1.051727076
EFEMP1	0.010941	1.050592244
ARMCX1	0.000131677	1.04919162
ZNF655	0.0193538	1.049059152
ABCA5	0.0014649	1.048710495
B3GNT7	0.000531656	1.048354778
VWA5A	0.0112037	1.04416327
ST3GAL6	0.0128438	1.043813427
BMPR2	0.00305426	1.041810695
OSR1	0.00773538	1.041572423
TCEAL3	0.018428	1.04137617
SAMD9L	0.0197889	1.040990594
DST	0.0166022	1.040773224
CDC42EP3	0.0125682	1.040745174
ADAMTS6	0.0146894	1.037895245

VTCN1	0.0126501	1.032707424
NRP1	0.0756638	1.028717657
MAMDC2	0.0166182	1.028625727
IL18	0.0876015	1.027967904
PCDH7	0.0444837	1.02603514
FKBP1B	0.059044	1.025539142
GNB4	0.0806152	1.024653007
SEC22B	0.010026	1.024575001
TCEAL8	0.0285956	1.023368918
NNMT	0.0279635	1.022360711
FN1	0.136216	1.022140515
GPR87	0.0235535	1.021529475
CD36	0.061237	1.021444193
PLAU	0.103789	1.021273613
MAP1B	0.135878	1.01868402
ANXA6	0.0584246	1.018562963
MBOAT2	0.00363628	1.018327943
TFPI	0.0247272	1.014091013
DUSP4	0.0951561	1.012046563
ALCAM	0.0410317	1.010020703
GNAI1	0.136446	1.006886832
RDX	0.0827592	1.003559063
EDIL3	0.119804	1.002572915
DCBLD2	0.120165	1.001009533
MSN	0.246098	0.999971146
C13orf15	0.195322	0.998209948
DST	0.0162586	0.996388746
DUSP4	0.0913974	0.993956896
YPEL3	0.00186492	0.993543943
HAPLN1	0.0125086	0.992021516
---	0.00218263	0.990061746
IDS	0.00923896	0.989487828
PRRX1	0.00269389	0.986745856
GLIS2	0.0377343	0.985813709
NR3C1	0.107907	0.981845348
LIFR	0.0118233	0.981816129
FGF9	0.0824989	0.979352251
HOXC6	0.126238	0.979066842
EFNA5	0.000773697	0.978576392
GABRP	0.0238045	0.976972116
ERAP2	0.0990624	0.976935468
HDAC9	0.0705935	0.976040964
SLC16A7	0.000559068	0.975336702
TFAP2C	0.0580035	0.975145905
YPEL5	0.00132832	0.973104269
CPEB4	0.00531921	0.972133846

CHST15	0.14445	0.97128787
EPHA4	0.0753826	0.971221642
CST6	0.125034	0.971177488
---	0.0818886	0.970058481
MBP	0.00257248	0.966208952
PDGFA	0.0573412	0.964088064
SLC2A12	0.0534128	0.963740443
---	0.000107268	0.963015348
TGFB2	0.0165221	0.962608247
GJC1	0.0258574	0.959606962
ADRB1	0.0213511	0.959013376
OSMR	0.000404417	0.958308176
PLA2G4A	0.220699	0.95689674
FAM107B	0.0781643	0.954211203
GNG11	0.112838	0.953265239
INPP4B	0.0374155	0.952929906
TLE1	0.00743181	0.952847924
KCNMA1	0.0011562	0.951035676
PRRX1	0.0201414	0.9502519
FRY	0.0410841	0.949803838
PPP3CA	0.00276563	0.949116537
GULP1	0.190224	0.948570947
CREB3L1	0.0221124	0.948002715
GLIS3	0.0507745	0.94798028
SERPINA1	3.20E-05	0.946438923
SLC16A6	0.142836	0.946408978
STXBP1	0.0594211	0.942608336
IDS	0.00748541	0.941819968
LOC100131199	0.0273699	0.941669754
---	0.0658936	0.941444403
C10orf10	0.00983902	0.940813234
PLEKHH2	0.00834859	0.940587749
IDS	0.00900419	0.940392301
CUEDC1	0.000511076	0.939625286
RAB26	0.00490044	0.939076093
EGFR	0.026169	0.938601964
NAP1L5	0.0186274	0.938473998
TMED3	0.0150349	0.936945053
DNM1	0.0353218	0.936394839
GPR126	0.108715	0.931123266
CAV1	0.283159	0.930086322
TNFAIP2	0.0590806	0.930025748
GCG	0.0439167	0.92954106
B2M	0.00386508	0.92873036
RGS20	0.039964	0.927486905
FAM129A	0.0489666	0.926971012

SLFN5	0.0685869	0.926781299
NCRNA00261	0.0515958	0.926766121
SH3BGRL	0.144085	0.92653084
RNF144B	0.020268	0.926386616
MCAM	0.0318929	0.926363843
TFPI2	0.010933	0.925946266
NXN	0.209492	0.925300681
CXorf61	0.00350531	0.922418613
PSPC1	0.000223709	0.922327266
BMPR1B	0.0140035	0.918019789
PLAC8	0.292714	0.91735537
COL14A1	0.00868132	0.916874047
FOXC1	0.0670708	0.915826808
GNA15	0.0095106	0.913117276
KLF9	0.0570413	0.912803131
CMTM3	0.0102856	0.910832419
PLAU	0.111473	0.90883596
FGF9	0.015066	0.907506014
HAPLN1	0.0369584	0.905343197
RBMS1	0.163698	0.904865555
SLC38A11	0.00850994	0.904865555
HEPACAM2	0.0850206	0.903045985
MEIS1	0.0159381	0.901733855
KLHL5	0.08469	0.900760564
CPNE8	0.000818	0.900675563
LOC554202	0.0140938	0.896519866
MBOAT2	0.0293651	0.895915249
ERBB4	4.64E-05	0.895287108
TMEM22	0.0174875	0.895124213
TOB1	0.0902179	0.894821643
SYNE1	0.00722131	0.894379311
FLRT3	0.0323752	0.894363789
TGM2	0.171578	0.894161976
PCDHGA1 /// PCDHGA10 /// PCDHGA11 /// PCDHGA12 /// PCDHGA2 /// PCDHGA3 /// PCDHGA4 /// PCDHGA5 /// PCDHGA6 /// PCDHGA7 /// PCDHGA8 /// PCDHGA9 /// PCDHGB1 /// PCDHGB2 /// PCDHGB3 /// PCDHGB4 /// PCDHGB5 /// PCDHGB6 /// PCDHGB7 /// PCDHGC3 /// PCDHGC4 /// PCDHGC5	0.000456332	0.892709647
APLP1	0.000702878	0.89191685
ZNF655	0.0974872	0.891636936
PRKCA	0.0341719	0.890399996
CAPN8	0.126099	0.890026372
CLIC3	0.14642	0.889255467

FAM174B	0.0038011	0.88891271
IKZF2	0.0124788	0.887096298
SRGN	0.0653275	0.886635985
SNCA	0.0158607	0.886581361
PCDHGA1 /// PCDHGA10 /// PCDHGA11 /// PCDHGA12 /// PCDHGA2 /// PCDHGA3 /// PCDHGA4 /// PCDHGA5 /// PCDHGA6 /// PCDHGA7 /// PCDHGA8 /// PCDHGA9 /// PCDHGB1 /// PCDHGB2 /// PCDHGB3 /// PCDHGB4 /// PCDHGB5 /// PCDHGB6 /// PCDHGB7 /// PCDHGC3 /// PCDHGC4 /// PCDHGC5	0.000459287	0.885160433
BCL9L	0.000449485	0.884816702
MCAM	0.0254537	0.884386922
DIP2C	0.0653342	0.883652094
AMOTL2	0.0478611	0.882345677
---	0.0509689	0.880450444
ZNF548	0.00538039	0.87842716
KIAA1370	0.00153493	0.877602918
MALL	0.121047	0.877532247
SRGAP2P1	0.00943153	0.875630675
CGNL1	0.0209208	0.875441953
C8orf84	0.129599	0.874474364
NR3C1	0.174281	0.873860435
KLK6	0.276171	0.870913086
RAB26	0.00384788	0.870416014
ZNF264	0.0178555	0.869776671
DPYSL3	0.0729543	0.868907972
NR3C1	0.158147	0.867667196
FAM84B	0.0715034	0.867129458
RABGAP1L	0.0109757	0.865546714
TP53INP2	0.00623261	0.863954304
KRT4	0.000676804	0.863105886
DUSP10	0.015973	0.862852057
EPB41L4A	0.100082	0.862804459
GULP1	0.094444	0.861034285
F5	0.0991319	0.860660922
TOX	0.0786385	0.85946074
SFXN3	0.000364674	0.859174456
IDS	0.0249517	0.858593762
RAB27B	0.00024787	0.857837713
HDGFRP3	0.00417516	0.857805871
LAT /// SPNS1	0.000582712	0.857041445
PARP6	0.00129142	0.856013609
LY6K	0.00551476	0.855846218
MTSS1	0.00467126	0.855232283

LHFP	0.0688678	0.854929206
PSPH	0.0564737	0.85239842
RAP1GAP	0.0220285	0.8515032
B3GNT7	0.00980958	0.851327288
---	0.00387418	0.85083142
KLF9	0.0254241	0.850479411
SERPINB5	0.254858	0.849310764
FAM171B	0.257118	0.849222676
PRKD1	0.0209644	0.848557846
SNX18	0.0843003	0.848205281
FGFR2	0.0910353	0.848012936
ADAM28	0.00547181	0.847916755
ANP32E	0.12173	0.846489304
LIMCH1	0.172232	0.846144259
LCN2	0.326457	0.845325449
LEPR	0.00569177	0.845261209
COL3A1	0.0225117	0.845221057
---	0.0416021	0.844931934
IRS1	0.0583635	0.84363016
SEMA3C	0.0136901	0.843533685
ELFN2	0.0563739	0.84337288
LTBP3	0.0258021	0.84264099
RBMS1	0.166943	0.842142128
GABARAPL1 /// GABARAPL3	0.173162	0.841932876
COL5A2	0.00927119	0.840048242
FAM91A2 /// FLJ39739 /// LOC100286793 /// LOC728855 /// LOC728875	0.133176	0.839814505
LEPR	0.00862962	0.839693591
KRT80	0.157752	0.839451734
SNAP25	0.0231692	0.838669451
EVL	0.00822979	0.838467763
THBD	0.00121349	0.838241839
RAB34	0.106237	0.838209561
RHBDL2	0.0456282	0.835220741
CALD1	0.404206	0.834282429
EGFR	0.0313741	0.834274337
CNTNAP3	0.0561326	0.833958731
TP63	0.00328745	0.833829232
PCDHGA1 /// PCDHGA10 /// PCDHGA11 /// PCDHGA12 /// PCDHGA2 /// PCDHGA3 /// PCDHGA4 /// PCDHGA5 /// PCDHGA6 /// PCDHGA7 /// PCDHGA8 /// PCDHGA9 /// PCDHGB1 /// PCDHGB2 /// PCDHGB3 /// PCDHGB4 /// PCDHGB5 /// PCDHGB6 /// PCDHGB7 ///	0.00291775	0.83314917

PCDHGC3 /// PCDHGC4 /// PCDHGC5		
IDS	0.0104315	0.832841417
IFIT1	0.292658	0.831496255
C11orf70	0.0676717	0.829013325
PRKACB	0.251485	0.827184912
MMP1	0.149504	0.825606551
---	0.0892098	0.824938889
---	0.10732	0.823733059
---	0.114398	0.82339068
NME7	0.107324	0.822632266
HIST1H2BD	0.0669685	0.82248543
DCBLD2	0.244799	0.822469114
LIMCH1	0.104104	0.822085634
CFLAR	0.0025707	0.820607877
ERAP2	0.130132	0.820419987
CALHM2	0.0285482	0.820093164
SDC2	0.0425247	0.819872517
SLC6A14	0.0888387	0.819382068
ST3GAL1	0.151215	0.816927322
CAV2	0.277788	0.816198276
C17orf28	4.52E-05	0.815181913
LBH	0.126173	0.815026117
MMP7	0.265671	0.81466526
EHD2	0.00329321	0.814378151
SERTAD4	0.015981	0.814156628
---	0.143593	0.81262144
DHRS3	0.0172193	0.812481796
LOC100509683	0.0201745	0.812186035
CAV2	0.380134	0.810574719
---	0.0131965	0.810072871
FAM114A1	0.0372605	0.809825996
---	0.120149	0.809307422
ARHGAP23	0.0877975	0.808714537
DCBLD2	0.201509	0.808508617
BEAN1	0.00643594	0.808203801
PTGS2	0.271281	0.807890681
COL27A1	0.0896761	0.807668159
PPP3CA	0.012232	0.806225056
NRG1	0.182078	0.805919757
LYPD1	0.00403776	0.805333734
FAM149A	0.0527888	0.803425445
TSPAN4	0.0898872	0.803069944
FLJ90757	0.0210922	0.802573749
LATS2	0.0357546	0.802011188
MGLL	0.0761043	0.800993057
RASEF	0.157206	0.800943374

SHISA3	0.0050472	0.800744624
RBMS1	0.153595	0.800289051
---	0.0743819	0.800231059
ANTXR1	0.225209	0.799493525
IL8	0.243648	0.799460369
SPOCK2	0.0453279	0.796921639
PPIC	0.00746211	0.796772163
RAB31	0.139357	0.796589448
LRRN1	0.0819182	0.796240566
TMEM27	0.031713	0.795534236
ZBTB20	0.0725194	0.795318113
EFEMP1	0.0144642	0.792705454
KRT6A	0.0548142	0.791338997
LONRF2	0.000354528	0.790930475
---	0.0767944	0.79074702
VCAN	0.0956809	0.790471793
ZBTB20	0.0885542	0.788134416
RBMS3	0.0308569	0.788126062
INHBA	0.0612704	0.786914138
CST3	0.00853604	0.786738534
PARP6	0.00335759	0.785081783
GNAI1	0.261196	0.785048293
CYR61	0.312005	0.784897582
---	0.0353886	0.784646361
RGNEF	0.0374779	0.784512358
KRT40	0.0933439	0.784512358
RAB34	0.0830994	0.783951088

Table A.4: List of the 500 top probesets downregulated in broad myofibroblast-recruiting cell lines (recruitment scores 4 or 5) compared to broad non-recruiters (recruitment score 1 or 2), sorted by $\log_2$ of the expression ratio between them. P-values uncorrected for multiple comparisons are given.

Gene Symbol	p-value	$\log_2$(Ratio)
CCND2	0.0140733	-2.708651058
GAL	0.00186695	-2.631519939
ABHD12B	0.00509967	-2.27409659
LOC100507192	0.00107858	-2.133679108
NPTX2	0.000180913	-2.056683251
SATB2	0.00188447	-1.975714367
C2orf89	0.0016572	-1.939034012
C13orf18	0.0461358	-1.921013147
LGR5	0.0266706	-1.876874171
C10orf99	0.0466135	-1.875767165
IL33	0.0142572	-1.874703334
PKP2	1.22E-05	-1.850095534
FAM84A	0.0704494	-1.789516372
PLIN2	0.00502345	-1.756296802
CCND2	0.0176625	-1.750459999
IFITM1	0.0100407	-1.749654431
CEACAM5	0.12913	-1.744081172
PLBD1	0.0101304	-1.742521035
BMP7	0.0204129	-1.737826659
FABP1	0.125728	-1.724025122
TDGF1 /// TDGF3	0.0175631	-1.713189806
IFITM1	0.0225392	-1.697593205
PCK1	0.0567012	-1.693056535
GPA33	0.0483648	-1.678930853
CDH17	0.135617	-1.655316916
PIGR	0.0575262	-1.641599237
PAQR8	0.00139893	-1.640829707
C10orf99	0.0587125	-1.624292147
FAR2	0.0031659	-1.621146609
TMPRSS2	0.0119797	-1.61709596
BCL11A	0.00931421	-1.616733109
VIL1	0.0584752	-1.605999591
SLC6A20	0.0256863	-1.602451106
PAH	0.00344977	-1.60177661
CDC42EP5	0.0255708	-1.591688128
FAM84A	0.0473295	-1.583415319
EPB41L2	0.000885956	-1.578038338
ABCB1	0.070356	-1.577978036
DDC	0.0867486	-1.576355141
MYB	0.0267867	-1.557080914

BCL11A	0.00374823	-1.554447024
SPON1	0.0306937	-1.549835527
NETO2	0.00716572	-1.545386099
XK	0.00286156	-1.53976699
CFTR	0.0444783	-1.538882201
HEPH	0.145678	-1.506311681
ASCL2	0.0361145	-1.505623306
PARVB	0.000307426	-1.498777944
CEL /// LOC100508206	0.058614	-1.481686444
MEST	0.0075993	-1.479890589
TBX3	0.000417968	-1.477225092
MSX1	0.0109846	-1.466819428
CEL /// LOC100508206	0.0662969	-1.462248705
WT1	0.00145986	-1.461116227
C2orf82	0.00381723	-1.452392238
AXIN2	0.0229048	-1.451149122
CLRN3	0.158969	-1.445578142
LOC653602	0.0496912	-1.44080777
FABP6	0.012698	-1.435268701
FAR2	0.0127216	-1.427044381
LOC100288092	0.0151377	-1.426129148
VIL1	0.0816204	-1.422814351
ABCB1 /// ABCB4	0.0600254	-1.41761742
PARVB	0.000603015	-1.412915414
NETO2	0.0135047	-1.407953015
---	7.18E-06	-1.402790176
MME	0.0604994	-1.386913482
BCL11A	0.00307928	-1.38535234
SAMD5	0.0403835	-1.373416951
EPB41L2	0.00210328	-1.372079445
AKAP2 /// PALM2-AKAP2	0.0481756	-1.364474044
CDX1	0.0103625	-1.354405314
GPR160	0.0227917	-1.351490379
CFD	0.0015574	-1.35038638
ETS2	0.0228945	-1.344226874
CFTR	0.11925	-1.338009667
DPEP1	0.0637615	-1.328670391
DLX6	0.00156079	-1.328424005
OLFM4	0.26506	-1.327978443
CCND2	0.043531	-1.326707717
LEF1	0.0705868	-1.312828202
NFE2L3	0.00241536	-1.312068581
PARVB	0.00204991	-1.306469776
ELAVL2	0.0254046	-1.299638411
---	0.0322299	-1.288156246
C13orf18	0.106163	-1.287779303

TPD52L1	0.00357682	-1.284697004
ASB4 /// LOC285986	0.0330133	-1.28262824
FAM171A1	0.0192489	-1.281919432
ASB4	0.0255488	-1.280534402
NFE2L3	0.00248654	-1.28004032
SLC16A1	0.00584473	-1.279227732
SAMD5	0.0990634	-1.278475094
FABP5	0.00550448	-1.27834561
ID3	0.0750323	-1.277054905
C10orf116	0.0678092	-1.273960521
SMOC2	0.136861	-1.26894539
NRARP	0.0031641	-1.26600711
CYP2J2	0.0129114	-1.263195997
NRXN3	0.0527224	-1.25783508
PTPRO	0.136053	-1.249558139
CALB1	0.225301	-1.246371505
EPHX2	0.00104412	-1.242348592
LY75	0.109491	-1.236112563
TPD52L1	0.00287644	-1.234784379
---	0.0034669	-1.233423495
SLC16A1	0.00383273	-1.230363652
PARM1	0.0871389	-1.230356882
SMPDL3A	0.0132027	-1.228520038
NTRK2	0.0587161	-1.226280429
ACOX2	0.0260527	-1.22069825
IRF8	0.0224387	-1.220583935
CEBPA	0.00750159	-1.217732356
SPON1	0.0370523	-1.217628342
LGALS4	0.320429	-1.210973595
DUSP27	0.195056	-1.209604947
AGMAT	0.00361319	-1.208164263
MFAP3L	0.0268907	-1.206675067
SYK	0.0211617	-1.182098462
PKP2	0.000437673	-1.180230464
SAMD5	0.0469169	-1.179031117
CD200	0.024328	-1.175970279
CPT1A	0.000869501	-1.170419931
SPON1	0.0160137	-1.169640819
GPX2	0.270834	-1.158902825
ICAM2	0.0186046	-1.157573017
P2RX5	0.00640699	-1.150564322
AUTS2	0.146052	-1.149376575
IQGAP2	0.123089	-1.149251774
SATB2	0.0252128	-1.148116263
ACPL2	0.00269478	-1.143758451
SLCO1B3	0.159662	-1.142933068

ITPR2	0.0125002	-1.138251033
PPP1R1B	0.185415	-1.136778319
RHOU	0.0647971	-1.135120142
ITPR2	0.0184632	-1.134701935
TNFRSF1B	0.0103993	-1.128515888
LCN15	0.00431688	-1.126876626
ASB4	0.00641092	-1.125185726
GOLIM4	0.00839172	-1.122927987
TMPRSS2	0.0230709	-1.122196081
LGR5	0.0284857	-1.120783578
LPAR5	0.00488245	-1.120175059
STAMBPL1	0.00258589	-1.106383059
CNN3	0.093207	-1.102824642
LRP4	0.0993233	-1.102589172
RNF157	0.0189661	-1.102307278
SLC16A1	0.00778333	-1.102226747
VAV3	0.079626	-1.102183386
KBTBD8	6.65E-05	-1.101762233
STXBP6	0.0343449	-1.100539726
TUBAL3	0.018614	-1.099117353
UTS2	0.000414046	-1.098280045
C14orf139	0.0271656	-1.098190473
KIAA1274	0.00441027	-1.09635392
METTL7A	0.141887	-1.095838871
GUCY2C	0.152773	-1.093331019
GIPC2	0.108092	-1.089709423
TBX3	0.00120485	-1.082044348
SLCO3A1	0.0352928	-1.07729977
FZD3	5.57E-05	-1.074933319
SLC16A9	0.0422602	-1.072127813
GAS6	0.0319745	-1.067418705
KCNV1	0.0219957	-1.06552728
UGT8	0.033522	-1.061715714
PAQR8	0.00886688	-1.06157418
ME2	0.00774413	-1.060803518
SNHG5	0.00325602	-1.056671247
PPIL4 /// ZC3H12D	0.00792417	-1.051067296
MLLT3	0.0368712	-1.050045299
UTS2	0.00195929	-1.04852858
TNFSF9	0.00533808	-1.046745192
PTPRG	0.0295383	-1.046688566
FZD3	0.000768386	-1.046539559
BMP7	0.0440761	-1.045675621
NUPR1	0.10123	-1.043152313
ANKRD43	0.041734	-1.040253581
CADM1	0.12912	-1.040072603

SLC26A3	0.246265	-1.039500152
ACSL6	0.0472984	-1.03231787
CA12	0.119175	-1.029758885
CKMT1A /// CKMT1B	0.0640208	-1.02862236
C2CD4A	0.0870088	-1.022440175
SLC16A1	0.0188745	-1.020211694
SULT1C2	0.0914114	-1.019611969
AGPAT5	2.78E-05	-1.017770469
MGAT4A	0.0408589	-1.015108836
FZD7	0.0335683	-1.013555569
HUNK	0.0514132	-1.011026712
C9orf152	0.216726	-1.009910653
---	0.0413377	-1.009669535
PTPRG	0.0595546	-1.009495258
---	0.0172992	-1.007257668
BCL11A	0.00994714	-1.006909717
DEPDC7	0.0741803	-1.000378036
SLC6A20	0.0667624	-0.997851984
ETS2	0.0341709	-0.997013828
VAV3	0.171803	-0.994479284
SLC38A4	0.0438573	-0.988510086
GOLIM4	0.00580693	-0.988441388
HSPD1	6.15E-06	-0.985219103
LDLRAD3	0.000531208	-0.983877417
NRTN	0.0160087	-0.983184226
SLC16A1	0.00591464	-0.982830627
EPB41L4B	0.0153357	-0.980733628
OBSCN	0.0185256	-0.979959421
PAPSS2	0.286779	-0.979088934
VAV3	0.22504	-0.974659305
CA12	0.166732	-0.972588297
AKAP2 /// PALM2-AKAP2	0.0746537	-0.971883529
SEL1L3	0.070286	-0.96982215
UGGT2	0.00886828	-0.96651427
CENPV	0.00542178	-0.966398688
EPHB2	0.0558445	-0.966269021
CALB1	0.317888	-0.965668761
EPB41L4B	0.0172992	-0.962524999
PSIP1	0.00189558	-0.961232332
SLC39A8	0.101622	-0.961201434
GDA	0.148007	-0.959273059
EPHB3	0.0822987	-0.95810662
NLRP2	0.180071	-0.957775926
ALDH2	0.055157	-0.953587171
GSPT2	0.256029	-0.952643094
CA12	0.141836	-0.95244486

---	0.0221071	-0.951847532
PAPSS2	0.247304	-0.948555504
SPON1	0.0269167	-0.948193589
TNS4	0.0688326	-0.947144547
NOX1	0.301588	-0.94431011
---	0.000265309	-0.944212948
---	0.017816	-0.942520614
NTRK2	0.0596041	-0.942046562
PLEKHH1	0.0248033	-0.940234956
KCNK1	0.0297842	-0.938702131
---	0.0219196	-0.937690363
---	0.0327226	-0.936342442
VDR	0.0186639	-0.931043086
FUT4	0.0187405	-0.930611288
BIK	0.0591157	-0.928737076
---	0.0260714	-0.928684896
GGT6	0.086404	-0.925439761
STAMBPL1	0.00634441	-0.924609763
KISS1R	0.105853	-0.923840455
NOS3	0.0678261	-0.922992227
DDC	0.0901263	-0.917676185
CKMT2	0.010286	-0.917117596
TSPAN6	0.00620638	-0.913504258
MATR3	0.00189804	-0.912376944
FAM3B	0.174683	-0.912254758
TMEM158	0.0913388	-0.912181451
RASL11A	0.00836549	-0.911964268
ME2	0.0151671	-0.910086998
JHDM1D	0.0212115	-0.909154697
GATA6	0.219844	-0.907405549
FAM84A	0.0920192	-0.907370371
UNC93A	0.0603674	-0.907075453
SLC3A1	0.194779	-0.90701864
FZD7	0.0285728	-0.904583228
MME	0.0855849	-0.904016186
MRPL16	0.000163604	-0.903770537
MPP7	0.0648332	-0.903662573
---	0.112571	-0.901017273
IL32	0.165898	-0.900378921
CDHR1	0.126923	-0.899713935
SERPINI1	0.147775	-0.895309085
NUP43	0.00245621	-0.893453349
ATP10B	0.234874	-0.893096956
PGAP1	0.0276561	-0.892654937
CA12	0.261011	-0.891166463
FZD3	0.00332777	-0.889802486

UGT1A1 /// UGT1A10 /// UGT1A3 /// UGT1A4 /// UGT1A5 /// UGT1A6 /// UGT1A7 /// UGT1A8 /// UGT1A9	0.333815	-0.887892411
PROX1	0.164827	-0.887710885
EPDR1	0.281821	-0.887046375
ACSL5	0.171445	-0.885369116
TRIM36	0.0285708	-0.885307822
RASGEF1A	0.128905	-0.882703914
PTGFRN	0.0239512	-0.880768674
TSPAN6	0.00790361	-0.88029058
TMEM45B	0.183873	-0.879146455
ABHD6	0.000103921	-0.87843018
FERMT1	0.0347417	-0.875965728
TMEM20	0.0110001	-0.874309228
HIP1	0.0156606	-0.873859708
ABHD6	0.000711079	-0.872683686
---	0.0143799	-0.871732998
CENPV	0.0154172	-0.871004561
HAUS6	0.000942537	-0.868887323
NT5DC1	0.00078946	-0.867252084
TNIK	0.0760985	-0.866475929
HMGCS2	0.383042	-0.865742254
---	0.0316398	-0.864557066
SEL1L3	0.0734792	-0.863034308
DENND4C	0.00362808	-0.862840141
---	0.0904775	-0.86237582
NEK4	2.58E-05	-0.860100969
HOOK1	0.00864274	-0.85954067
GIPC2	0.14028	-0.858737256
GLS2	0.0396808	-0.858444265
---	0.0464787	-0.857011559
TBC1D4	0.0144654	-0.85488343
ADH1C	0.122324	-0.853555919
NFIA	0.00825893	-0.853162335
EPPK1	0.123719	-0.851466718
DYNC1I1	0.0130886	-0.849435125
PREP	0.00328715	-0.848520411
GEM	0.0583555	-0.847253245
LCP1	0.0146494	-0.846404762
GRIN2D	0.054947	-0.84231466
ARG2	0.0960816	-0.842125847
GNAL	0.0118608	-0.841381184
SRSF6	0.173053	-0.841352749
UPP1	0.0320612	-0.840520645
PRSS3	0.138313	-0.840169342
RPL31 /// TBC1D8	0.0162681	-0.83832898

EPCAM	0.151372	-0.837359411
GRTP1	0.0938445	-0.837302701
ETNK1	0.000694077	-0.837245994
CDX2	0.112135	-0.836689346
XPNPEP3	7.67E-05	-0.834879148
PIGR	0.00575381	-0.834704171
MAP7	0.0451948	-0.834580671
TRAF5	0.0582895	-0.834171651
WT1	0.00184334	-0.833536486
---	0.00962512	-0.832878473
LOC100287015	5.95E-06	-0.832806521
TRBC1 /// TRBC2	0.00654755	-0.832051241
KIAA1199	0.353617	-0.831922831
---	0.0026388	-0.830344327
BCAT1	0.240408	-0.83014425
ZNF703	0.0470962	-0.828803423
HAUS6	0.00342805	-0.826527129
C12orf75	0.0666655	-0.825266344
---	0.000889822	-0.824341275
TBX3	0.00210311	-0.823659353
KCNE3	0.0688001	-0.823501051
---	0.161824	-0.823383611
BCL11A	0.0289927	-0.822707245
CELP	0.109848	-0.821472738
KIF26A	0.0160405	-0.820218908
---	0.134661	-0.816115906
NPL	0.0190983	-0.81462564
---	0.0441497	-0.814364304
GNAL	0.00239842	-0.813864599
ZNF91	0.006216	-0.813844311
PRTFDC1	0.196567	-0.813798661
PLEKHH1	0.024381	-0.813661723
SYK	0.0508764	-0.811295249
STX3	0.0176436	-0.811269933
LOC339290	0.00941967	-0.810672604
CDC25A	0.00971179	-0.810191884
---	0.122981	-0.810017346
STC2	0.165893	-0.81000217
HBA1 /// HBA2	0.195496	-0.809605116
C12orf48	0.00114604	-0.809536844
CPT1A	0.0585616	-0.808007889
TNFRSF10D	0.0151353	-0.805494324
NRXN3	0.125111	-0.804032617
LOC100506691	0.0302879	-0.80075672
CYP2S1	0.0964172	-0.800198898
ABCB1	0.0396671	-0.798760111

CDH1	0.289795	-0.797984816
ACSL6	0.113138	-0.79704699
KANK1	0.0264414	-0.796127309
ICAM2	0.0129012	-0.794157106
RUNDC3B	0.142771	-0.793981996
FABP1	0.0565361	-0.793731875
---	0.147749	-0.793579323
LPHN1	0.0107044	-0.79343929
MYBBP1A	0.00139543	-0.79293678
HOXA13	0.0789408	-0.792669347
CDH3	0.107269	-0.791894819
UGT1A1 /// UGT1A10 /// UGT1A4 /// UGT1A6 /// UGT1A8 /// UGT1A9	0.292347	-0.790396914
JPH1	0.053536	-0.790234734
SMARCA4	0.0135651	-0.789760775
---	0.0896828	-0.788743535
PHLDA1	0.174818	-0.78863886
---	0.072451	-0.78735596
SCARB1	0.00405841	-0.787054707
PALM2	0.00410431	-0.785827935
IL17RB	0.0737813	-0.785815498
LONRF3	0.0291547	-0.784704102
NR1I2	0.227287	-0.784214564
ITGA6	0.0420989	-0.784149967
ACN9	0.00159326	-0.78353396
ISX	0.186883	-0.781134535
SCARB1	0.00377977	-0.780785004
CCDC88C	0.00252819	-0.780735432
STRN3	0.0395596	-0.78038352
DUSP16	0.0168424	-0.779947466
C12orf11	0.000413431	-0.779838473
NDUFAF4	0.00408754	-0.779410012
ZBTB38	0.0431148	-0.778875234
GAS6	0.041374	-0.777801325
SUV39H2	0.00216827	-0.777729595
KIF26A	0.0104631	-0.777576252
NOTCH1	0.00762629	-0.776678782
MLLT3	0.111292	-0.774710261
FAM105A	0.161972	-0.774673238
---	0.00361887	-0.773881174
TRIM31	0.359919	-0.773350907
ATP5S	0.00869778	-0.771889366
UGT8	0.051047	-0.770825566
RPL22L1	0.028904	-0.769248547
TTC39C	0.00210811	-0.76915757
LHX6	0.00707643	-0.768267778

CMAS	0.0016912	-0.76744484
DUSP2	0.00396014	-0.767393269
EEPD1	0.00537702	-0.766538923
IL17RB	0.130662	-0.76587887
PHF14	0.0021987	-0.764868519
---	0.0112265	-0.764302346
HAS2	0.135662	-0.763028655
PCSK9	0.138616	-0.76214508
SYBU	0.105713	-0.760340452
ARHGAP8 /// PRR5-ARHGAP8	0.0545141	-0.760059446
MST4	0.00606983	-0.759480502
SH2D2A	0.00128687	-0.759055602
FOS	0.142888	-0.75832087
ESRP1	0.192505	-0.758064657
PDSS1	0.00283486	-0.757925588
ASCL2	0.0874141	-0.757859718
SH3BP5	0.0692614	-0.757323113
FAM84A	0.150926	-0.757057323
STX3	0.00990994	-0.756677012
CCL14-CCL15 /// CCL15	0.325892	-0.75640403
ABHD6	0.000525318	-0.755809499
CFTR	0.0488494	-0.755083722
TCAM1P	0.072333	-0.754923029
HBE1	0.108337	-0.754285303
TOX2	0.212509	-0.753484893
HNF4A	0.0900357	-0.753013127
C17orf76	0.0324422	-0.752920737
RGMB	0.0818192	-0.752675203
C12orf48	0.00366454	-0.752575543
LRP12	0.0735	-0.751863539
RAD54B	0.00157532	-0.751564747
STXBP6	0.0731587	-0.751530742
HNF4G	0.127559	-0.750860524
---	0.0145422	-0.749729631
SLCO3A1	0.0384662	-0.749584087
APOBEC3B	0.234204	-0.748839615
TUBB6	0.384242	-0.748539025
FAM54A	0.0198714	-0.748437227
TMEM52	0.00310634	-0.747843546
PDK1	0.0268971	-0.746606082
SERPINE2	0.270991	-0.746482636
IQGAP3	0.00589273	-0.746414866
---	0.00176029	-0.746402765
LOC339290	0.00383849	-0.746320478
SMARCA4	0.0119074	-0.746158338
IRAK1BP1	0.00640845	-0.745822016

GALNTL4	0.0702067	-0.743462712
ZNF704	0.134911	-0.74334678
C2CD2	0.00486469	-0.742793816
LOC158402	0.0303869	-0.740150249
LRRC19	0.176752	-0.740039402
---	0.00219207	-0.739911698
CPA6 /// CPB1	0.0500863	-0.739499748
ALDH1B1	0.00598204	-0.73818278
GALNT6	0.132184	-0.738023959
SLC46A3	0.148411	-0.737930118
SMARCC1	0.000485558	-0.737528355
FERMT1	0.0718072	-0.737254162
BMP7	0.046822	-0.737177205
KCNK1	0.0573986	-0.737107466
CTNNB1	0.0625392	-0.736907888
LGALS3	0.173921	-0.736705932
MCM4	0.0418456	-0.736568908
EPPK1	0.205545	-0.736508813
PARM1	0.132679	-0.735112921
COTL1	0.127845	-0.734378278
AIFM2	0.010111	-0.733982302
LOC100505644	0.010746	-0.7316782
PSMG1	0.00190702	-0.731587166
CCNC	0.00114579	-0.731486557
MGAT4A	0.0779856	-0.731144059
GNAL	0.0392092	-0.73050718
HIP1	0.0410878	-0.730478455
LARS2	0.000125811	-0.730399465
OPLAH	0.0419846	-0.730385104
HPDL	0.0820241	-0.730303725
IL17RB	0.145162	-0.72916491
BCCIP	0.0028813	-0.727551532
CHORDC1	0.0103014	-0.727195637
CBX2	0.0162109	-0.726703739
TGS1	0.000165706	-0.726527078
HES6	0.0702541	-0.724919031
GSTM4	0.104942	-0.724172876
KCNE3	0.336455	-0.724003674
TTC39C	0.0171443	-0.723057939
GZMB	0.149861	-0.722869819
IFI30	0.0353444	-0.72248175
ME2	0.0276181	-0.721746371
QRSL1	0.016416	-0.721679754
PPARGC1B	0.0229793	-0.721658341
PRSS3	0.177919	-0.721594107
NES	0.0974619	-0.721173084

BLNK	0.21513	-0.721113627
SMARCC1	0.000321652	-0.720338537
KHDRBS3	0.0504264	-0.718989067

Code A.1: Gaussian Mixture Modeling (GMM) performed on the cell line microarray data. Microarray expression data from 78 colorectal cancer cell lines gathered in-house was first preprocessed in Partek Genomics Suite, and subsequently loaded into R. GMM was implemented in the shiny R package.

```
library("XML")
library("plyr")
library("ggplot2")
library("gridExtra")
library("dplyr")
library(preprocessCore)

## --- DATA INPUT ---
setwd("/Users/pkalugin/Downloads/bodmer94")

data78 <- read.csv("20170505clean78.csv", header=FALSE)
listProbesets <-
read.csv("/Users/pkalugin/Downloads/probesetids.csv")

# cleaning up
data78T <- t(data78)
d78 <- as.matrix(data78T)
colnames(d78) <- d78[5,]
d78 <- d78[-c(1:172),]
rownames(d78) <- d78[,1]
d78[,1] <- 0
e78 <- matrix(as.numeric(d78), ncol=79)
dd78 <- data.frame(e78)
dd78[,1] <- rownames(d78)
colnames(dd78) <- colnames(d78)

# appending probeset information
lP <- listProbesets[,c(2,4,5)]
lP <-
data.frame(matrix(unlist(lP),nrow=nrow(lP)),stringsAsFactors=FALSE)
colnames(lP) <- c("ID","Gene Symbol","Gene Title")
all78 <- merge(lP,dd78,by.y="ID")
## ------------------

## --- MCLUST RUN 1 ---
library(mclust)

# cleaning up
ma78 <- all78[,-c(1:3)]
ma78 <- as.matrix(ma78)
rownames(ma78) <- all78[,2]

# G=1:2 for all94, log2 data using modelNames="V" saving means, sds,
ns for validation
mma78 <-
function(x){Mclust(ma78[x,],G=c(1:2),modelNames="V",verbose=FALSE)}
mma78p <- function(x){mma78(x)$parameters}
mma78g <- function(x){mma78(x)$G}
mma78b <- function(x){mma78(x)$BIC}
mma78c <- function(x){mma78(x)$classification}
para78a <- sapply(c(1:54675), mma78p)
gaus78a <- sapply(c(1:54675), mma78g)
bic78a <- sapply(c(1:54675), mma78b)
clas78a <- sapply(c(1:54675), mma78c)
para78 <- t(para78a)
clas78 <- t(clas78a)
names(para78) <- rownames(ma78)
rownames(clas78) <- rownames(ma78)
colnames(clas78) <- colnames(ma78)
## --------------------
```

```
## --- TAIL ANALYSIS ---
# initialize validated table as original
clas78val <- clas78
para78val <- para78

# identify probesets with tail problems
plval78 <- function(x){
  a <- rbind(ma78[x,],clas78[x,])
  b <- a[,order(a[1,])]
  #plot(b[1,],type="p",col=b[2,],main=rownames(ma78)[x])
  #axis(1,at=1:78,labels=colnames(b))
  return(all(b[2,] == cummax(b[2,])) | all(b[2,] == cummin(b[2,])))
}
plv <- sapply(c(1:54675),plval78)

# identify tail ends in probesets with tail problems
plvtest <- function(x){
  if (!plv[x]) {
    a <- rbind(ma78[x,],clas78[x,])
    b <- a[,order(a[1,])]
    return(c(1,1+which(diff(b[2,])!=0)))
  } else {return(1)}
}
plvt <- sapply(c(1:54675),plvtest)

# trim shortest tail, rerun Mclust and test if the remaining data
splits without tail problems
plvtestcut <- function(x){
  if (!plv[x]) {
    a <- data.frame(rbind(ma78[x,],clas78[x,]))
    b <- a[,order(a[1,])]
    ## calculate the shorter tail, then cut it
    t1 <- plvt[[x]][2]-1
    t2 <- 81-plvt[[x]][3]
    if (t1<t2) {
      tail <- c(1:t1)
    } else {
      tail <- c(plvt[[x]][3]:78)
    }
    arg <- b[1,-tail]
    m <-
Mclust(arg,G=c(1:2),modelNames="V",verbose=FALSE)$classification
    p <-
Mclust(arg,G=c(1:2),modelNames="V",verbose=FALSE)$parameters
    if (all(m == cummax(m)) | all(m == cummin(m))) {
      # successful tail trimming - retain the NEW cutoff in the
validated dataset
      if (t1<t2) {
        n <- m[1]
        b[2,] <- c(rep(n,t1),m)
        b <- b[colnames(a)]
      } else {
        n <- m[length(m)]
        b[2,] <- c(m,rep(n,t2))
        b <- b[colnames(a)]
      }
    } else {
      # uncategorizable - deem unimodal
      b[2,] <- rep(1,78)
      p <- Mclust(b[1,],G=1,modelNames="V",verbose=FALSE)$parameters
    }
    clas78val[x,] <- as.numeric(b[2,])
    para78val[[x]] <- p
    assign('clas78val',clas78val,envir=.GlobalEnv)
    assign('para78val',para78val,envir=.GlobalEnv)
    return(all(m == cummax(m)) | all(m == cummin(m)))
  } else {return(TRUE)}
}
plvtc <- sapply(c(1:54675),plvtestcut)
```

```
### at this point, clas78val is the final table of 1s and 2s
write.csv(clas78val,"/Users/pkalugin/Desktop/phenoSubsets/clas78valC
lean.csv",quote=FALSE)
##--------------

##---- EXTENDED DATA -----
library("XML")
library("plyr")
library("ggplot2")
library("gridExtra")
library("dplyr")
library(preprocessCore)

## --- DATA INPUT ---
setwd("/Users/pkalugin/Documents")

data78ext <- read.csv("20170705extendedanalysisdata78.csv")

# cleaning up
data78extT <- t(data78ext)
d78ext <- as.matrix(data78extT)
colnames(d78ext) <- d78ext[2,]
d78ext <- d78ext[-c(1:3),]
e78ext <- matrix(as.numeric(d78ext),ncol=78)
dd78ext <- data.frame(e78ext)
rownames(dd78ext) <- rownames(d78ext)
colnames(dd78ext) <- colnames(d78ext)

# GMM/binarization
ma78ext <- as.matrix(dd78ext)

lumen1v3 <- numeric()
for (i in 1:length(ma78ext[17,])) {
  if (!is.na(ma78ext[17,i]) & ma78ext[17,i]==1) {
    lumen1v3[i] <- 1
  } else if (!is.na(ma78ext[17,i]) & ma78ext[17,i]==3) {
    lumen1v3[i] <- 2
  } else {
    lumen1v3[i] <- NA
  }
}
names(lumen1v3) <- names(ma78ext[17,])

lumen12v3 <- numeric()
for (i in 1:length(ma78ext[17,])) {
  if (!is.na(ma78ext[17,i]) & ma78ext[17,i]<3) {
    lumen12v3[i] <- 1
  } else if (!is.na(ma78ext[17,i]) & ma78ext[17,i]==3) {
    lumen12v3[i] <- 2
  } else {
    lumen12v3[i] <- NA
  }
}
names(lumen12v3) <- names(ma78ext[17,])

lumen1v23 <- numeric()
for (i in 1:length(ma78ext[17,])) {
  if (!is.na(ma78ext[17,i]) & ma78ext[17,i]==1) {
    lumen1v23[i] <- 1
  } else if (!is.na(ma78ext[17,i]) & ma78ext[17,i]>1) {
    lumen1v23[i] <- 2
  } else {
    lumen1v23[i] <- NA
  }
}
names(lumen1v23) <- names(ma78ext[17,])

library(mclust)
```

```
goblethot <- ma78ext[3,][!is.na(ma78ext[3,]) & ma78ext[3,]<3]
gobletall <- numeric()
for (i in 1:length(ma78ext[3,])) {
  if (!is.na(ma78ext[3,i]) & ma78ext[3,i]>2) {
    gobletall[i] <- 1
  } else {
    gobletall[i] <- ma78ext[3,i]
  }
}
names(gobletall) <- names(ma78ext[3,])

mf1v5 <- numeric()
for (i in 1:length(ma78ext[4,])) {
  if (!is.na(ma78ext[4,i]) & ma78ext[4,i]==1) {
    mf1v5[i] <- 1
  } else if (!is.na(ma78ext[4,i]) & ma78ext[4,i]==5) {
    mf1v5[i] <- 2
  } else {
    mf1v5[i] <- NA
  }
}
names(mf1v5) <- names(ma78ext[4,])

mf12v45 <- numeric()
for (i in 1:length(ma78ext[4,])) {
  if (!is.na(ma78ext[4,i]) & ma78ext[4,i]<3) {
    mf12v45[i] <- 1
  } else if (!is.na(ma78ext[4,i]) & ma78ext[4,i]>3) {
    mf12v45[i] <- 2
  } else {
    mf12v45[i] <- NA
  }
}
names(mf12v45) <- names(ma78ext[4,])

cealysis <-
Mclust(ma78ext[5,][!is.na(ma78ext[5,])],G=c(1:2),prior=priorControl(
),modelNames="V")$classification
ceabinding <-
Mclust(ma78ext[6,][!is.na(ma78ext[6,])],G=c(1:2),prior=priorControl(
),modelNames="V")$classification
wntphilips <-
Mclust(ma78ext[7,][!is.na(ma78ext[7,])],G=c(1:2),modelNames="E")$cla
ssification

apchot <- ma78ext[8,][!is.na(ma78ext[8,]) & ma78ext[8,]<3]
apcall <- numeric()
for (i in 1:length(ma78ext[8,])) {
  if (!is.na(ma78ext[8,i]) & ma78ext[8,i]>2) {
    apcall[i] <- 2
  } else {
    apcall[i] <- ma78ext[8,i]
  }
}
names(apcall) <- names(ma78ext[8,])

p53 <- ma78ext[9,]

ctnnb1hot <- ma78ext[10,][!is.na(ma78ext[10,]) & ma78ext[10,]<3]
ctnnb1all <- numeric()
for (i in 1:length(ma78ext[10,])) {
  if (!is.na(ma78ext[10,i]) & ma78ext[10,i]>2) {
    ctnnb1all[i] <- 2
  } else {
    ctnnb1all[i] <- ma78ext[10,i]
  }
}
names(ctnnb1all) <- names(ma78ext[10,])
```

```
krashot <- ma78ext[11,][!is.na(ma78ext[11,]) & ma78ext[11,]<3]
krasall <- numeric()
for (i in 1:length(ma78ext[11,])) {
  if (!is.na(ma78ext[11,i]) & ma78ext[11,i]>2) {
    krasall[i] <- 2
  } else {
    krasall[i] <- ma78ext[11,i]
  }
}
names(krasall) <- names(ma78ext[11,])

brafhot <- ma78ext[12,][!is.na(ma78ext[12,]) & ma78ext[12,]<3]
brafall <- numeric()
for (i in 1:length(ma78ext[12,])) {
  if (!is.na(ma78ext[12,i]) & ma78ext[12,i]>2) {
    brafall[i] <- 2
  } else {
    brafall[i] <- ma78ext[12,i]
  }
}
names(brafall) <- names(ma78ext[12,])

pik3cahot <- ma78ext[13,][!is.na(ma78ext[13,]) & ma78ext[13,]<3]
pik3caall <- numeric()
for (i in 1:length(ma78ext[13,])) {
  if (!is.na(ma78ext[13,i]) & ma78ext[13,i]>2) {
    pik3caall[i] <- 2
  } else {
    pik3caall[i] <- ma78ext[13,i]
  }
}
names(pik3caall) <- names(ma78ext[13,])

fbxw7hot <- ma78ext[14,][!is.na(ma78ext[14,]) & ma78ext[14,]<3]
fbxw7all <- numeric()
for (i in 1:length(ma78ext[14,])) {
  if (!is.na(ma78ext[14,i]) & ma78ext[14,i]>2) {
    fbxw7all[i] <- 2
  } else {
    fbxw7all[i] <- ma78ext[14,i]
  }
}
names(fbxw7all) <- names(ma78ext[14,])

tgfbr2hot <- ma78ext[15,][!is.na(ma78ext[15,]) & ma78ext[15,]<3]
tgfbr2all <- numeric()
for (i in 1:length(ma78ext[15,])) {
  if (!is.na(ma78ext[15,i]) & ma78ext[15,i]>2) {
    tgfbr2all[i] <- 2
  } else {
    tgfbr2all[i] <- ma78ext[15,i]
  }
}
names(tgfbr2all) <- names(ma78ext[15,])

smad4 <- ma78ext[16,]

library(gtools)
ext78 <-
smartbind(ma78ext[1,],lumen1v3,lumen12v3,lumen1v23,goblethot,gobleta
ll,mf1v5,mf12v45,cealysis,ceabinding,wntphilips,apchot,apcall,p53,ct
nnb1hot,ctnnb1all,krashot,krasall,brafhot,brafall,pik3cahot,pik3caal
l,fbxw7hot,fbxw7all,tgfbr2hot,tgfbr2all,smad4)
rownames(ext78) <- c("RER status","Lumen Formation 1 (no) vs 3
(yes)","Lumen Formation 1+2 vs 3","Lumen Formation 1 vs 2+3","Goblet
Cells Confirmed","Goblet Cells Broad (+/- as negative)","MF
Recruiters Firm (1 vs 5)","MF Recruiters Broad (1 and 2 vs 4 and
5)","CEA lysis","CEA binding","WNT Philips", "APC hotspot
mutations", "APC all mutations", "p53 all mutations", "CTNNB1
```

```
hotspot mutations", "CTNNB1 all mutations", "KRAS hotspot
mutations", "KRAS all mutations", "BRAF hotspot mutations", "BRAF
all mutations", "PIK3CA hotspot mutations", "PIK3CA all mutations",
"FBXW7 hotspot mutations", "FBXW7 all mutations", "TGFBR2 hotspot
mutations", "TGFBR2 all mutations", "SMAD4 all mutations")

cealpara <-
Mclust(ma78ext[5,][!is.na(ma78ext[5,])],G=c(1:2),prior=priorControl(
),modelNames="V")$parameters
ceabpara <-
Mclust(ma78ext[6,][!is.na(ma78ext[6,])],G=c(1:2),prior=priorControl(
),modelNames="V")$parameters
wntppara <-
Mclust(ma78ext[7,][!is.na(ma78ext[7,])],G=c(1:2),modelNames="E")$par
ameters

cealmeans <- cealpara$mean
cealssq <- cealpara$variance$sigmasq
ceabmeans <- ceabpara$mean
ceabssq <- ceabpara$variance$sigmasq
wntpmeans <- wntppara$mean
wntpssq <- wntppara$variance$sigmasq

meanext78 <-
cbind(rbind(rep(NA,5),rep(NA,5)),data.frame(cealmeans),data.frame(ce
abmeans),data.frame(wntpmeans),rbind(rep(NA,16),rep(NA,16)))
ssqext78 <-
cbind(rbind(rep(NA,5),rep(NA,5)),data.frame(cealssq),data.frame(ceab
ssq),data.frame(wntpssq),rbind(rep(NA,16),rep(NA,16)))

write.csv(meanext78,"/Users/pkalugin/Desktop/phenoSubsets/meanext78.
csv",quote=FALSE)
write.csv(ssqext78,"/Users/pkalugin/Desktop/phenoSubsets/ssqext78.cs
v",quote=FALSE)
##-----------

# adding extended data
#write.csv(ext78,"/Users/pkalugin/Desktop/phenoSubsets/ext7820170705
.csv",quote=FALSE) ##USE THIS ONE

clas78valext <- as.matrix(smartbind(clas78val,ext78))
nm <- c(rownames(clas78val),rownames(ext78))
rownames(clas78valext) <- nm

#write.csv(clas78valext,"/Users/pkalugin/Desktop/phenoSubsets/clas78
valExt.csv",quote=FALSE)

#n <- c(54676:54702)
#ext78$N <- n

#write.csv(ext78,"/Users/pkalugin/Desktop/phenoSubsets/ext78.csv",qu
ote=FALSE)
## ---------------------

##--- MEANS/STDEV OUTPUT---
para78valmeansf <- function(x){
  if (length(para78val[[x]]$mean)==1){
    return(c(para78val[[x]]$mean,0))
  } else {return(para78val[[x]]$mean)}
}
para78valmeans <- sapply(c(1:54675),para78valmeansf)
colnames(para78valmeans) <- names(para78val)

para78valsigmasqf <- function(x){
  if (length(para78val[[x]]$variance$sigmasq)==1){
    return(c(para78val[[x]]$variance$sigmasq,0))
  } else {return(para78val[[x]]$variance$sigmasq)}
}
para78valsigmasq <- sapply(c(1:54675),para78valsigmasqf)
```

```
colnames(para78valsigmasq) <- names(para78val)

para78valmeans <- cbind(para78valmeans,meanext78)
para78valsigmasq <- cbind(para78valsigmasq,ssqext78)

write.csv(para78valmeans,"/Users/pkalugin/Desktop/phenoSubsets/para7
8valmeansExt.csv",quote=FALSE)
write.csv(para78valsigmasq,"/Users/pkalugin/Desktop/phenoSubsets/par
a78valsigmasqExt.csv",quote=FALSE)
####
```

Code A.2: Gaussian Mixture Modeling (GMM) performed on the TCGA data. RNA-Seq expression data from 622 primary colorectal cancers was queried from the Genomic Data Commons and loaded into R. GMM was implemented in the shiny R package.

```
library(TCGAbiolinks)
library(biomaRt)
library(SummarizedExperiment)
library(GEOquery)
library(preprocessCore)
library(plyr)
library(gtools)

# Querying data
qa <- GDCquery(project=c("TCGA-COAD","TCGA-
READ"),data.category="Transcriptome Profiling",data.type="Gene
Expression Quantification",workflow.type="HTSeq - FPKM")
#qb <- GDCquery(project=c("TCGA-COAD","TCGA-
READ"),data.category="Transcriptome Profiling",data.type="Gene
Expression Quantification",workflow.type="HTSeq - FPKM-UQ")
#qc <- GDCquery(project=c("TCGA-COAD","TCGA-
READ"),data.category="Transcriptome Profiling",data.type="Gene
Expression Quantification",workflow.type="HTSeq - Counts")
View(getResults(qa))
#View(getResults(qb))
#View(getResults(qc))
dim(getResults(qa))
#dim(getResults(qb))
#dim(getResults(qc))

a <- getResults(qa)
ab <- a[which(a[,26]=="Primary solid Tumor"),]
ac <- ab$cases
ac <- ac[order(ac)]
ad <- strtrim(ac,12)
ae <- duplicated(ad)
af <- cbind(ac,ae)
ag <- af[which(af[,2]=="FALSE"),]
qqa <- ag[,1]

#b <- getResults(qb)
#bb <- b[which(b[,26]=="Primary solid Tumor"),]
#bc <- bb$cases
#bc <- bc[order(bc)]
#bd <- strtrim(bc,12)
#be <- duplicated(bd)
#bf <- cbind(bc,be)
#bg <- bf[which(bf[,2]=="FALSE"),]
#qqb <- bg[,1]

#c <- getResults(qc)
#cb <- c[which(c[,26]=="Primary solid Tumor"),]
#cc <- cb$cases
#cc <- cc[order(cc)]
#cd <- strtrim(cc,12)
#ce <- duplicated(cd)
#cf <- cbind(cc,ce)
#cg <- cf[which(cf[,2]=="FALSE"),]
#qqc <- cg[,1]
#qqa==qqb==qqc!!

qat <- GDCquery(project=c("TCGA-COAD","TCGA-
READ"),data.category="Transcriptome Profiling",data.type="Gene
Expression Quantification",workflow.type="HTSeq - FPKM",barcode=qqa)
#qbt <- GDCquery(project=c("TCGA-COAD","TCGA-
READ"),data.category="Transcriptome Profiling",data.type="Gene
```

```
Expression Quantification",workflow.type="HTSeq - FPKM-
UQ",barcode=qqa)
#qct <- GDCquery(project=c("TCGA-COAD","TCGA-
READ"),data.category="Transcriptome Profiling",data.type="Gene
Expression Quantification",workflow.type="HTSeq -
Counts",barcode=qqa)

#GDCdownload(qat,method="client")
#GDCdownload(qbt,method="client")
#GDCdownload(qct,method="client")

tdata <- GDCprepare(qat)
#tdatauq <- GDCprepare(qbt)
#tdatacts <- GDCprepare(qct)

td <- assay(tdata)
#tduq <- assay(tdatauq)
#tdcts <- assay(tdatacts)

#tdctsnorm <-
TCGAanalyze_Normalization(tabDF=tdcts,geneInfo=geneInfoHT)

# Cleaning up
tdnorm <- normalize.quantiles(td)
rownames(tdnorm) <- rownames(td)
colnames(tdnorm) <- colnames(td)

tdnormlog <- log2(tdnorm)
tdnl <- as.data.frame(tdnormlog)
tdnl$ID <- rownames(tdnl)

tdnlI <- do.call(data.frame,lapply(tdnl, function(x) replace(x,
is.infinite(x),NA)))

ensembl = useMart("ensembl",dataset="hsapiens_gene_ensembl")
listGenes <-
getBM(attributes=c("ensembl_gene_id","external_gene_name"),mart=ense
mbl)
colnames(listGenes)[1] <- c("ID")

tdnlnamed <- merge(listGenes,tdnlI,by.y="ID")

tdnl622 <- as.matrix(tdnlnamed[,-c(1:2)])
rownames(tdnl622) <- tdnlnamed[,2]
hist(rowSums(is.na(tdnl622)),breaks=100)
remove(td,tdnorm,tdnormlog,tdnl,tdnlI,tdnlnamed)
tdnl622 <- tdnl622[rowSums(is.na(tdnl622))<621, ] #at least 2
nonzero RNAseq counts to get an estimate of bimodality

# GMM
library(mclust)
mtdnl622 <-
function(x){Mclust(na.omit(tdnl622[x,]),G=c(1:2),modelNames="V",verb
ose=FALSE)}
mtdnl622p <- function(x){mtdnl622(x)$parameters}
mtdnl622g <- function(x){mtdnl622(x)$G}
mtdnl622b <- function(x){mtdnl622(x)$BIC}
mtdnl622c <- function(x){
  mm <- mtdnl622(x)$classification
  names(mm) <- rownames(mtdnl622(x)$data)
  print(x)
  return(mm)
}
para622a <- sapply(c(1:53393), mtdnl622p)
gaus622a <- sapply(c(1:53393), mtdnl622g)
bic622a <- sapply(c(1:53393), mtdnl622b)
clas622a <- sapply(c(1:53393), mtdnl622c)
##
clas622aC <- do.call(smartbind,clas622a)
```

```
clas622 <- as.matrix(clas622aC)
##

names(para622a) <- rownames(tdnl622)
rownames(clas622) <- rownames(tdnl622)
colnames(clas622) <- colnames(tdnl622)

write.csv(clas622,"/Users/pkalugin/Desktop/RWorkspace/clas622.csv")
write.csv(tdnl622,"/Users/pkalugin/Desktop/RWorkspace/tdnl622.csv")

## --- TAIL ANALYSIS ---
# initialize validated table as original
clas622val <- clas622
para622val <- para622a

# identify genes with tail problems
plval622 <- function(x){
  a <- rbind(tdnl622[x,],clas622[x,])
  b <- a[,order(a[1,])]
  #plot(b[1,],type="p",col=b[2,],main=rownames(ma622)[x])
  #axis(1,at=1:622,labels=colnames(b))
  return(all(b[2,] == cummax(b[2,]),na.rm=TRUE) | all(b[2,] ==
cummin(b[2,]),na.rm=TRUE))
}
plv622 <- sapply(c(1:53393),plval622)

# identify tail ends in genes with tail problems
plv622test <- function(x){
  if (!plv622[x]) {
    a <- rbind(tdnl622[x,],clas622[x,])
    b <- a[,order(a[1,])]
    return(c(1,1+which(diff(b[2,])!=0)))
  } else {return(1)}
}
plv622t <- sapply(c(1:53393),plv622test)

# trim shortest tail, rerun Mclust and test if the remaining data
splits without tail problems
plv622testcut <- function(x){
  if (!plv622[x]) {
    a <- data.frame(rbind(tdnl622[x,],clas622[x,]))
    b <- a[,order(a[1,])]
    ## calculate the shorter tail, then cut it
    t1 <- plv622t[[x]][2]-1
    t2 <- length(na.omit(as.numeric(b[2,])))+1-plv622t[[x]][3]
    if (t1<t2) {
      tail <- c(1:t1)
    } else {
      tail <- c(plv622t[[x]][3]:length(na.omit(as.numeric(b[2,]))))
    }
    arg <- na.omit(as.numeric(b[1,]))[-tail]
    m <-
Mclust(arg,G=c(1:2),modelNames="V",verbose=FALSE)$classification
    p <-
Mclust(arg,G=c(1:2),modelNames="V",verbose=FALSE)$parameters
    if (all(m == cummax(m)) | all(m == cummin(m))) {
      # successful tail trimming - retain the NEW cutoff in the
validated dataset
      if (t1<t2) {
        n <- m[1]
        b[2,] <- c(rep(n,t1),m,rep(NA,622-
length(na.omit(as.numeric(b[2,])))))
        b <- b[colnames(a)]
      } else {
        n <- m[length(m)]
        b[2,] <- c(m,rep(n,t2),rep(NA,622-
length(na.omit(as.numeric(b[2,])))))
        b <- b[colnames(a)]
      }
```

```
    } else {
      # uncategorizable - deem unimodal
      b[2,] <-
c(rep(1,length(na.omit(as.numeric(b[2,])))),rep(NA,622-
length(na.omit(as.numeric(b[2,]))))
      b <- b[colnames(a)]
      p <-
Mclust(na.omit(as.numeric(b[1,])),G=1,modelNames="V",verbose=FALSE)$
parameters
    }
    clas622val[x,] <- as.numeric(b[2,])
    para622val[[x]] <- p
    assign('clas622val',clas622val,envir=.GlobalEnv)
    assign('para622val',para622val,envir=.GlobalEnv)
    return(all(m == cummax(m)) | all(m == cummin(m)))
  } else {return(TRUE)}
}
plv622tc <- sapply(c(1:53393),plv622testcut)

write.csv(clas622val,"/Users/pkalugin/Desktop/RWorkspace/clas622val.
csv",quote=FALSE)

# Save mean/sd parameters
para622valmeansf <- function(x){
  if (length(para622val[[x]]$mean)==1){
    return(c(para622val[[x]]$mean,0))
  } else {return(para622val[[x]]$mean)}
}
para622valmeans <- sapply(c(1:53393),para622valmeansf)
colnames(para622valmeans) <- names(para622val)

para622valsigmasqf <- function(x){
  if (length(para622val[[x]]$variance$sigmasq)==1){
    return(c(para622val[[x]]$variance$sigmasq,0))
  } else {return(para622val[[x]]$variance$sigmasq)}
}
para622valsigmasq <- sapply(c(1:53393),para622valsigmasqf)
colnames(para622valsigmasq) <- names(para622val)

write.csv(para622valmeans,"/Users/pkalugin/Desktop/RWorkspace/para62
2valmeans.csv",quote=FALSE)
write.csv(para622valsigmasq,"/Users/pkalugin/Desktop/RWorkspace/para
622valsigmasq.csv",quote=FALSE)
```

Code A.3: Calculation of the *r* and p-values for 2x2 comparisons between genes using the categorized data, shown here for the cell line microarray data and analogous for TCGA.

```
setwd("/t1-home/pharm/pkalugin/")
#setwd("/Users/pkalugin/Desktop/phenoSubsets")

aaa <- read.csv("clas78valExt.csv")
aa <- as.matrix(aaa[,-1])
rownames(aa) <- aaa[,1]

bimindf <- function(x){
  if (max(aa[x,],na.rm=TRUE)==2){
    return(x)
  }
}
biminds <- unlist(sapply(c(1:54702),bimindf))

# chisq iterative functions
chis1 <- function(x,y){

tryCatch({chisq.test(aa[x,],aa[y,])$p.value},error=function(e){NA})
}
chis2cut <- function(x){
  print(x)
  return(mapply(chis1,x,biminds))
}

#chis <- sapply(c(1:1),chis2)

library(parallel)
chispvext <-
parallel::mcmapply(chis2cut,biminds,mc.cores=4,mc.preschedule=FALSE)
rownames(chispvext) <- biminds
colnames(chispvext) <- biminds
write.csv(chispvext,"chispvextS0705.csv",quote=FALSE)

###
setwd("/t1-home/pharm/pkalugin/")
#setwd("/Users/pkalugin/Desktop/phenoSubsets")
library(psych)

aaa <- read.csv("clas78valExt.csv")
aa <- as.matrix(aaa[,-1])
rownames(aa) <- aaa[,1]

bimindf <- function(x){
  if (max(aa[x,],na.rm=TRUE)==2){
    return(x)
  }
}
biminds <- unlist(sapply(c(1:54702),bimindf))

# chisq iterative functions
phi1 <- function(x,y){

tryCatch({phi(table(aa[x,],aa[y,]),digits=10)},error=function(e){NA}
)
}
phi2cut <- function(x){
  print(x)
  return(mapply(phi1,x,biminds))
}

#chis <- sapply(c(1:1),chis2)

library(parallel)
```

```
phiext <-
parallel::mcmapply(phi2cut,biminds,mc.cores=4,mc.preschedule=FALSE)
rownames(phiext) <- biminds
colnames(phiext) <- biminds
write.csv(phiext,"phiextS0705.csv",quote=FALSE)
```

Code A.4: Definition of the user interface for the interactive program to explore the 2x2 comparisons, shown here for the cell line microarray data and analogous for TCGA.

```
library(shiny)
library(shinythemes)

#setwd("/t1-data/public/shinyApp/pkalugin/caa2/csv")
ext78 <- read.csv("/Users/pkalugin/Desktop/phenoSubsets/ext78.csv")
#ext78 <- read.csv("ext78.csv")
ne <- ext78[,1]
ext78 <- as.matrix(ext78[,-1])
rownames(ext78) <- ne
remove(ne)

phenos <- list()
phenos <- ext78[,79]
names(phenos) <- rownames(ext78)

shinyUI(fluidPage( # layout examples:
https://rstudio.github.io/shinydashboard/structure.html

  titlePanel("Microarray Categorical Association Analysis"),

  sidebarLayout(
  sidebarPanel(
    # query input
    tabsetPanel(id="pan1",
      tabPanel("Gene Query",value=1,
        textInput("query",h3("Enter Query"),value="Enter gene name
or probeset ID...")
      ),
      tabPanel("Phenotype Query",value=2,
        selectInput("queryp",label=h3("Enter
Query"),choices=phenos))
    ),
    # direct vs total associations
    checkboxInput("directcomp","Check to enter second query for one-
to-one comparison"),
    # second query for direct associations
    conditionalPanel(
      condition="input.directcomp",
      tabsetPanel(id="pan2",
        tabPanel("Gene Query",value=1,
          textInput("query2",h3("Enter Second Query"),value="Enter
gene name or probeset ID...")
        ),
        tabPanel("Phenotype Query",value=2,
          selectInput("query2p",label=h3("Enter Second
Query"),choices=phenos))
      )
    ),
    # submit query
    actionButton("submitq","Submit Query",class="btn-primary"),
    p("Loading..."),
    textOutput("load")
  ),

  mainPanel(
    uiOutput("errm"),
    fluidRow(
      column(6,
    uiOutput("title1"),
    uiOutput("details1")
      ),
    column(6,
    uiOutput("title2"),
    uiOutput("details2")
```

```
    )
  ),
  fluidRow(
    column(6,
           uiOutput("querypick1")
    ),
    column(6,
           uiOutput("querypick2")
    )
  ),
  uiOutput("subsetting"),
  uiOutput("subanbutton"),
  uiOutput("queryout"),
  uiOutput("assoc2"),
  uiOutput("assoc1")
))))
```

Code A.5: Definition of the server-side script for the interactive program to explore the 2x2 comparisons, shown here for the cell line microarray data and analogous for TCGA.

```
#library("XML")
library("plyr")
#library("ggplot2")
#library("gridExtra")
library("dplyr")
library(preprocessCore)
#library(mclust)
library(shiny)
library(DT)

setwd("/Users/pkalugin/Desktop/RWorkspace/ShinyAlgo")
#setwd("/t1-data/public/shinyApp/pkalugin/caa2/csv")

# file input/cleanup
chispvext <- read.csv("/Volumes/Migration3/chispvextS0705.csv")
#chispvext <- read.csv("/Volumes/MIGRATION2/chispvextS.csv")
#chispvext <- read.csv("chispvextS.csv")
names <- chispvext[,1]
chispvext <- as.matrix(chispvext[,-1])
rownames(chispvext) <- names
colnames(chispvext) <- names
remove(names)

phiext <- read.csv("/Volumes/Migration3/phiextS0705.csv")
#phiext <- read.csv("/Volumes/MIGRATION2/phiextS.csv")
#phiext <- read.csv("phiextS.csv")
np <- phiext[,1]
phiext <- as.matrix(phiext[,-1])
rownames(phiext) <- np
colnames(phiext) <- np
remove(np)

aaa <-
read.csv("/Users/pkalugin/Desktop/phenoSubsets/clas78valExt.csv")
#aaa <- read.csv("clas78valExt.csv")
clas78val <- as.matrix(aaa[,-1])
rownames(clas78val) <- aaa[,1]
remove(aaa)

#bbb <- read.csv("/Users/pkalugin/Desktop/phenoSubsets/paramsS.csv")
#params <- as.matrix(bbb[,-1])
#rownames(params) <- bbb[,1]
#remove(bbb)

#cct <-
read.csv("/Users/pkalugin/Desktop/phenoSubsets/chistest.csv")
#chistest <- as.numeric(cct[,2])
#names(chistest) <- rep("X-squared",54675)

ddd <-
read.csv("/Users/pkalugin/Desktop/phenoSubsets/para78valmeansExt.csv
")
#ddd <- read.csv("para78valmeansExt.csv")
para78valmeans <- as.matrix(ddd[,-1])
rownames(para78valmeans) <- ddd[,1]
remove(ddd)

#meandiffs <- sapply(c(1:54675),function(x){para78valmeans[2,x]-
para78valmeans[1,x]})
#names(meandiffs) <- c(1:54675)

eee <-
read.csv("/Users/pkalugin/Desktop/phenoSubsets/para78valsigmasqExt.c
sv")
```

```
#eee <- read.csv("para78valsigmasqExt.csv")
para78valsigmasq <- as.matrix(eee[,-1])
rownames(para78valsigmasq) <- eee[,1]
remove(eee)

# Welch's t-slider
welcht <- sapply(c(1:54675),
                 function(x){
                   if (max(clas78val[x,],na.rm=TRUE)==2){
                   return(abs(para78valmeans[2,x]-
para78valmeans[1,x])/sqrt((para78valsigmasq[2,x]/table(clas78val[x,]
)[2])+(para78valsigmasq[1,x]/table(clas78val[x,])[1])))
                   } else {return(NA)}})
names(welcht) <- c(1:54675)

data78 <-
read.csv("/Users/pkalugin/Downloads/bodmer94/20170505clean78.csv",
header=FALSE)
listProbesets <-
read.csv("/Users/pkalugin/Downloads/probesetids.csv")
#data78 <- read.csv("20170505clean78.csv", header=FALSE)
#listProbesets <- read.csv("probesetids.csv")

# cleaning up
data78T <- t(data78)
d78 <- as.matrix(data78T)
colnames(d78) <- d78[5,]
d78 <- d78[-c(1:172),]
rownames(d78) <- d78[,1]
d78[,1] <- 0
e78 <- matrix(as.numeric(d78), ncol=79)
dd78 <- data.frame(e78)
dd78[,1] <- rownames(d78)
colnames(dd78) <- colnames(d78)

# appending probeset information
lP <- listProbesets[,c(2,4,5)]
lP <-
data.frame(matrix(unlist(lP),nrow=nrow(lP)),stringsAsFactors=FALSE)
colnames(lP) <- c("ID","Gene Symbol","Gene Title")
all78 <- merge(lP,dd78,by.y="ID")
remove(data78,data78T,d78,e78,dd78,lP)

ma78 <- all78[,-c(1:3)]
ma78 <- as.matrix(ma78)
rownames(ma78) <- all78[,2]

details <-
as.data.frame(cbind(c(1:54675),all78[,1],all78[,2],all78[,3]),string
sAsFactors=FALSE)
details[,1] <- as.numeric(details[,1])
remove(all78)

data78ext <-
read.csv("/Users/pkalugin/Documents/20170705extendedanalysisdata78.c
sv")
#data78ext <- read.csv("20170509extendedanalysisdata78.csv")

# cleaning up
data78extT <- t(data78ext)
d78ext <- as.matrix(data78extT)
colnames(d78ext) <- d78ext[2,]
d78ext <- d78ext[-c(1:3),]
e78ext <- matrix(as.numeric(d78ext),ncol=78)
dd78ext <- data.frame(e78ext)
rownames(dd78ext) <- rownames(d78ext)
colnames(dd78ext) <- colnames(d78ext)

ma78ext <- as.matrix(dd78ext)
```

```
ext78 <- read.csv("/Users/pkalugin/Desktop/phenoSubsets/ext78.csv")
#ext78 <- read.csv("ext78.csv")
ne <- ext78[,1]
ext78 <- as.matrix(ext78[,-1])
rownames(ext78) <- ne
remove(ne)

ext78vals <- ext78
ext78vals[6,] <- c(ma78ext[5,],ext78vals[6,79]) # CEA lysis
ext78vals[7,] <- c(ma78ext[6,],ext78vals[7,79]) # CEA binding
ext78vals[8,] <- c(ma78ext[7,],ext78vals[8,79]) # WNT Philips
remove(ma78ext)
ext78vals <- ext78vals[,-79]

ma78 <- ma78[,order(colnames(ma78))]
ext78vals <- ext78vals[,order(colnames(ext78vals))]
clas78val <- clas78val[,order(colnames(clas78val))]
colnames(ma78) <- colnames(ext78vals)
ma78 <- rbind(ma78,ext78vals)

detext <- data.frame(cbind(ext78[,79],"",rownames(ext78),""))
rownames(detext) <- detext[,1]
colnames(detext) <- colnames(details)
details <- rbind(details,detext)

maxes <- sapply(c(1:54675),function(x){
  max(ma78[x,],na.rm=TRUE)
})
names(maxes) <- c(1:54675)

# plotting
plot78v <- function(x){
  if (x>54675 & x!=54681 & x!=54682 & x!=54683) {
    return()
    } else {
  a <- rbind(ma78[x,],clas78val[x,])
  b <- a[,order(a[1,])]
  tails <- c(1,1+which(diff(b[2,])!=0))
  plot(b[1,],type="p",col=b[2,],main=rownames(ma78)[x],xlab="Cell
line rank order",ylab="Microarray gene expression level (log2)")
  if (length(tails)==2) {
    abline(v=mean(c(tails[2]-1,tails[2])),col="blue",lwd=2)
  }}
  #axis(1,at=1:78,labels=colnames(b))
}

plot78riv <- function(x){
  if (x>54675 & x!=54681 & x!=54682 & x!=54683) {
    return()
  } else {
  a <- rbind(ma78[x,],clas78val[x,])
  b <- a[,order(a[1,])]
  tails <- c(1,1+which(diff(b[2,])!=0))
  if (length(tails)==2) {
    d <- c(qnorm(ppoints(tails[2]-
1),mean=para78valmeans[1,x],sd=sqrt(para78valsigmasq[1,x])),qnorm(pp
oints(79-
tails[2]),mean=para78valmeans[2,x],sd=sqrt(para78valsigmasq[2,x])))
  } else {
    d <-
qnorm(ppoints(78),mean=para78valmeans[1,x],sd=sqrt(para78valsigmasq[
1,x]))
  }

plot(x=b[1,],y=d,type="p",col=b[2,],main=rownames(ma78)[x],xlab="Act
ual expression value (log2)",ylab="Theoretical expression value
(log2)")
  abline(a=0,b=1)
```

```
  if (length(tails)==2) {
    abline(v=mean(c(b[1,tails[2]-
1],b[1,tails[2]])),col="blue",lwd=2)
  }}
  #axis(1,at=1:78,labels=colnames(b))
}

plot78vQ <- function(q,x){
  if (x>54675 & x!=54681 & x!=54682 & x!=54683) {
    return()
  } else {
  a <- rbind(ma78[q,],clas78val[q,],ma78[x,],clas78val[x,])
  b <- a[,order(a[3,])]
  tails <- c(1,1+which(diff(b[4,])!=0))
  plot(b[3,],type="p",col=b[2,],main=rownames(ma78)[x],xlab="Cell
line rank order",ylab="Microarray gene expression level (log2)")
  if (length(tails)==2) {
    abline(v=mean(c(tails[2]-1,tails[2])),col="blue",lwd=2)
  }}
  #axis(1,at=1:78,labels=colnames(b))
}

plot78rivQ <- function(q,x){
  if (x>54675 & x!=54681 & x!=54682 & x!=54683) {
    return()
  } else {
  a <- rbind(ma78[q,],clas78val[q,],ma78[x,],clas78val[x,])
  b <- a[,order(a[3,])]
  tails <- c(1,1+which(diff(b[4,])!=0))
  if (length(tails)==2) {
    d <- c(qnorm(ppoints(tails[2]-
1),mean=para78valmeans[1,x],sd=sqrt(para78valsigmasq[1,x])),qnorm(pp
oints(79-
tails[2]),mean=para78valmeans[2,x],sd=sqrt(para78valsigmasq[2,x])))
  } else {
    d <-
qnorm(ppoints(78),mean=para78valmeans[1,x],sd=sqrt(para78valsigmasq[
1,x]))
  }

plot(x=b[3,],y=d,type="p",col=b[2,],main=rownames(ma78)[x],xlab="Act
ual expression value (log2)",ylab="Theoretical expression value
(log2)")
  abline(a=0,b=1)
  if (length(tails)==2) {
    abline(v=mean(c(b[3,tails[2]-
1],b[3,tails[2]])),col="blue",lwd=2)
  }}
  #axis(1,at=1:78,labels=colnames(b))
}

plot78hist <- function(q){
  if (q>54675 & q!=54681 & q!=54682 & q!=54683) {
    return()
  } else {
  a <- rbind(ma78[q,],clas78val[q,])
  b <- a[,order(a[1,])]
  tails <- c(1,1+which(diff(b[2,])!=0))
  hist(a[1,],breaks=20,main=rownames(ma78)[q],xlab="Microarray gene
expression level (log2)",freq=FALSE)
  if (length(tails)==2) {
    curve(((tails[2]-
1)/78)*dnorm(x,mean=para78valmeans[1,q],sd=sqrt(para78valsigmasq[1,q
]))+((79-
tails[2])/78)*dnorm(x,mean=para78valmeans[2,q],sd=sqrt(para78valsigm
asq[2,q])),col="darkblue",lwd=2,add=TRUE)
    curve(((tails[2]-
1)/78)*dnorm(x,mean=para78valmeans[1,q],sd=sqrt(para78valsigmasq[1,q
])),lty="dashed",lwd=2,col="red",add=TRUE)
```

```
    curve(((79-
tails[2])/78)*dnorm(x,mean=para78valmeans[2,q],sd=sqrt(para78valsigm
asq[2,q])),lty="dashed",lwd=2,col="green",add=TRUE)
    abline(v=mean(c(b[1,tails[2]-
1],b[1,tails[2]])),col="blue",lwd=2)
  } else {

curve(dnorm(x,mean=para78valmeans[1,q],sd=sqrt(para78valsigmasq[1,q]
)),col="darkblue",lwd=2,add=TRUE)
  }}
}

shinyServer(function(input,output,session){

  # loaded message
  output$load <- renderText("Loaded")

  # query processing
  qout <- eventReactive(input$submitq, {
    if (input$pan1==1) {
      q1 <-
c(which(details[,2]==tolower(input$query)),which(details[,3]==touppe
r(input$query)))
    } else {q1 <- as.integer(input$queryp)}
    lq <- length(q1)
    if (input$directcomp) {
      if (input$pan2==1) {
        q2 <-
c(which(details[,2]==tolower(input$query2)),which(details[,3]==toupp
er(input$query2)))
      } else {q2 <- as.integer(input$query2p)}
      if (length(q2) == 0) {
        lq <- 0
      }
    }
    if (lq > 0 & input$directcomp) {
      if (length(q1) == length(q2)) {
        qq <- rbind(q1,q2)
      } else if (length(q1) > length(q2)) {
        q2 <- c(q2,rep(0,length(q1)-length(q2)))
        qq <- rbind(q1,q2)
      } else {
        q1 <- c(q1,rep(0,length(q2)-length(q1)))
        qq <- rbind(q1,q2)
      }
      return(qq)
    } else if (lq > 0) {
      return(q1)
    } else {return(0)}
  })

  output$title1 <- renderUI({
    if (qout()!=0){
      if (class(qout())=="integer") {
        h2(details[qout()[1],3])
      } else {
        h2(details[qout()[1,1],3])
      }
    } else {return()}
  })

  output$details1 <- renderUI({
    if (qout()!=0){
      if (class(qout())=="integer") {
        h3(details[qout()[1],4])
      } else {
        h3(details[qout()[1,1],4])
      }
    } else {return()}
```

```
  })

  output$title2 <- renderUI({
    if (qout()!=0){
      if (class(qout())!="integer") {
        h2(details[qout()[2,1],3])
      }
    } else {return()}
  })

  output$details2 <- renderUI({
    if (qout()!=0){
      if (class(qout())!="integer") {
        h3(details[qout()[2,1],4])
      }
    } else {return()}
  })

  output$errm <- renderUI({
    if (qout()==0) {
      h2("ERROR: Invalid query, please check spelling and try
again")
    } else {return()}
  })

  output$querypick1 <- renderUI({
    qt <- 0
    if (qout()==0) {
      return()
    } else { # This outputs query1 info
      if (class(qout())=="integer") { # In case of only one query
        qt <- qout()
      } else { # In case of two queries
        qt <- qout()[1,] ### Replacement has length 0??
        qt <- qt[qt>0]
      }
      if (length(qt)==1 & qt>54675) {
        return()
      } else {
      nm <- list()
     for (i in 1:length(qt)) {
        # output name, description of queried gene and a dropdown of
probeset options
        # probeset options must include probeset ID and if it's UNI
or BImodal
        # mean spread slider, option to show top N correlates (or N
per page?)
        # include a submit button!
        # https://shiny.rstudio.com/articles/layout-guide.html
        # https://gist.github.com/wch/5436415/
       id <- as.integer(details[qt[i],1])
       probeid <- details[qt[i],2]
       if (max(clas78val[id,],na.rm=TRUE)==1){
         bim <- "UNImodal"
       } else {
         bim <- "BImodal"
       }
       nm[i] <- id
       names(nm)[i] <- paste0(probeid," - ",bim)
     }
     selectInput(
       "pick1",
       label = h4("Select Probeset"),
       choices = nm
     )
      }
    }
  })
```

```
  output$querypick2 <- renderUI({
    qt <- 0
    if (qout()==0) {
      return()
    } else {
      if (class(qout())=="integer") { # In case of only one query
        return()
      } else { # In case of two queries
        qt <- qout()[2,]
        qt <- qt[qt>0]
      }
      if (length(qt)==1 & qt>54675) {
        return()
    } else {
    nm <- list()
    for (i in 1:length(qt)) {
      # output name, description of queried gene and a dropdown of
probeset options
      # probeset options must include probeset ID and if it's UNI or
BImodal
      # mean spread slider, option to show top N correlates (or N
per page?)
      # include a submit button!
      # https://shiny.rstudio.com/articles/layout-guide.html
      # https://gist.github.com/wch/5436415/
      id <- as.integer(details[qt[i],1])
      probeid <- details[qt[i],2]
      if (max(clas78val[id,],na.rm=TRUE)==1){
        bim <- "UNImodal"
      } else {
        bim <- "BImodal"
      }
      nm[i] <- id
      names(nm)[i] <- paste0(probeid," - ",bim)
    }
    nm[length(qt)+1] <- 0
    names(nm)[length(qt)+1] <- "Output all (bimodal) probesets"
    selectInput(
      "pick2",
      label = h4("Select Probeset"),
      choices = nm
    )
    }
    }
  })

  # mean-spread slider, etc
  output$subsetting <- renderUI({
    if (class(qout())=="integer") {
      fluidRow(
        column(6,
               sliderInput("welchts","Range of Welch's t values to
be
included:",min=floor(min(welcht,na.rm=TRUE)/0.01)*0.01,max=ceiling(m
ax(welcht,na.rm=TRUE)/0.01)*0.01,value=c(floor(min(welcht,na.rm=TRUE
)/0.01)*0.01,ceiling(max(welcht,na.rm=TRUE)/0.01)*0.01)),
               sliderInput("maxcutoff","Exclude probesets with
maximal value (log2)
below:",min=floor(min(maxes,na.rm=TRUE)/0.01)*0.01,max=ceiling(max(m
axes,na.rm=TRUE)/0.01)*0.01,value=6)
               #textOutput("meanspnum")
               ),
        column(6,
               numericInput("topn","Top N associated probesets to be
displayed",value=20)
               )
      )
    } else {return()}
  })
```

```
  output$subanbutton <- renderUI({
    if (qout()!=0) {
      actionButton("submitcor","Submit for Analysis",class="btn-
primary")
    } else {return()}
  })

  aout <- eventReactive(input$submitcor, { # should return row
details from pick1/2 or corrs
    if (qout()==0) { # error query
      return(0)
    } else if (class(qout())=="integer") { # single query
      if (length(qout())==1 & qout()>54675) {
        return(as.integer(qout()))
      } else {return(as.integer(input$pick1))}
    } else { # dual query
      q1 <- as.numeric(qout()[1,])
      q1 <- q1[q1>0]
      q2 <- as.numeric(qout()[2,])
      q2 <- q2[q2>0]
      a <- integer()
      b <- integer()
      if (length(q1)==1 & q1>54675) {
        a <- as.integer(q1)
      } else {a <- as.integer(input$pick1)}
      if (length(q2)==1 & q2>54675) {
        b <- as.integer(q2)
      } else {b <- as.integer(input$pick2)}
      return(rbind(a,b))
    }
  })

  q1name <- reactive({
    if (aout()!=0){
      if (class(aout())=="integer") {
        details[aout()[1],3]
      } else {
        details[aout()[1,1],3]
      }
    } else {return()}
  })

  q1probeid <- reactive({
    if (aout()!=0){
      if (class(aout())=="integer") {
        details[aout()[1],2]
      } else {
        details[aout()[1,1],2]
      }
    } else {return()}
  })

  q1descr <- reactive({
    if (aout()!=0){
      if (class(aout())=="integer") {
        details[aout()[1],4]
      } else {
        details[aout()[1,1],4]
      }
    } else {return()}
  })

  # plots/tables describing the query 1 gene
  output$queryout <- renderUI({
    if (aout()!=0) {
        wellPanel(
          h4(paste0("Query - ",q1name()," - ",q1probeid())),
          h4(q1descr()),
```

```
          fluidRow(
          column(6,
                 tabsetPanel(
                 tabPanel("Rank Plot", plotOutput("queryrank")),
                 tabPanel("Rankit Plot", plotOutput("queryrankit")),
                 tabPanel("Histogram", plotOutput("queryhist"))
               )),
          column(6,
               tabsetPanel(
               tabPanel("Data Table",
DT::dataTableOutput("querytable"))
               ))
      ))
    } else {return()}
    })

  output$queryrank <- renderPlot( # should call this on aout()
output rather than pick1/2
    if (class(qout())=="integer") {
      return(plot78v(as.numeric(aout()[1])))
    } else {
      return(plot78v(as.numeric(aout()[1,1])))
    }
  )

  output$queryrankit <- renderPlot(
    if (class(qout())=="integer") {
      return(plot78riv(as.numeric(aout()[1])))
    } else {
      return(plot78riv(as.numeric(aout()[1,1])))
    }
  )

  output$queryhist <- renderPlot(
    if (class(qout())=="integer") {
      return(plot78hist(as.numeric(aout()[1])))
    } else {
      return(plot78hist(as.numeric(aout()[1,1])))
    }
  )

  dt <- function(x) {
    dat <- t(rbind(ma78[x,],clas78val[x,]))
    rownames(dat) <- colnames(clas78val)
    colnames(dat) <- c("Query Expression Value","Query Subgroup")
    return(dat)
  }

  output$querytable <- DT::renderDataTable({ # figure out how to
make scrollY reactive to window size (width/3?)
    if (class(qout())=="integer") {

return(datatable(dt(as.numeric(aout()[1])),options=list(scrollY =
'300px',paging=FALSE)))
    } else {

return(datatable(dt(as.numeric(aout()[1,1])),options=list(scrollY =
'300px',paging=FALSE)))
    }
  })

  q2name <- reactive({
    if (aout()[2,1]!=0) {
        details[aout()[2,1],3]
      } else {return()}
  })

  q2probeid <- reactive({
    if (aout()[2,1]!=0) {
```

```
      details[aout()[2,1],2]
    } else {return()}
  })

  q2descr <- reactive({
    if (aout()[2,1]!=0) {
      details[aout()[2,1],4]
    } else {return()}
  })

  # output if pairwise query specified, 1 probeset
  output$assoc2 <- renderUI({
    if (tryCatch({class(qout())!="integer" &
aout()[2,1]!=0},error=function(e){FALSE})) {
      wellPanel(style = "background-color: #e8efff;",
        h4(paste0("Output - ",q2name()," - ",q2probeid())),
        h4(q2descr()),
        fluidRow(
            column(6,
                  tabsetPanel(
                    tabPanel("Rank Plot", plotOutput("assoc2rank")),
                    tabPanel("Rankit Plot",
plotOutput("assoc2rankit")),
                    tabPanel("Histogram", plotOutput("assoc2hist"))
                  )),
           column(6,
                  tabsetPanel(
                    tabPanel("Chi-squared",
uiOutput("assoc2chisq")),
                    tabPanel("Data Table",
DT::dataTableOutput("assoc2table"))
                  ))
               )
      )
    } else {return()}
  })

  output$assoc2rank <- renderPlot( # should call this on aout()
output rather than pick1/2

return(plot78vQ(as.numeric(aout()[1,1]),as.numeric(aout()[2,1])))
  )

  output$assoc2rankit <- renderPlot(

return(plot78rivQ(as.numeric(aout()[1,1]),as.numeric(aout()[2,1])))
  )

  output$assoc2hist <- renderPlot(
      return(plot78hist(as.numeric(aout()[2,1])))
  )

  output$assoc2chisq <- renderUI({
    if (max(clas78val[aout()[2,1],],na.rm=TRUE)!=1 &
max(clas78val[aout()[1,1],],na.rm=TRUE)!=1) {
    fluidRow(
      br(),
      br(),
      tableOutput("assoc2contingency"),
      textOutput("assoc2r"),
      textOutput("assoc2p"),
      textOutput("assoc2n")
      )
    } else {
      fluidRow(
        uiOutput("assoc2chiserr")
      )
    }
  })
```

```
  output$assoc2chiserr <- renderUI({
    h4("Unimodal Query or Comparison - Cannot Calculate
Association")
  })

  output$assoc2contingency <- renderTable({
    ct <-
as.table(xtabs(data=data.frame(t(rbind(clas78val[as.numeric(aout()[1
,1]),],clas78val[as.numeric(aout()[2,1]),])))))
    names(attributes(ct)$dimnames) <-
c(details[aout()[1,1],3],details[aout()[2,1],3])

return(data.frame(format(ftable(ct),method="compact",quote=FALSE)))
  },rownames=FALSE,colnames=FALSE)

  output$assoc2r <- renderText({
    id <-
c(which(rownames(phiext)==as.numeric(aout()[1,1])),which(rownames(ph
iext)==as.numeric(aout()[2,1])))
    phi <- signif(phiext[id[1],id[2]],digits=3)
    #chis <- 0
    return(paste0("r value: ",phi))
  })

  output$assoc2p <- renderText({
    id <-
c(which(rownames(chispvext)==as.numeric(aout()[1,1])),which(rownames
(chispvext)==as.numeric(aout()[2,1])))
    pv <- signif(chispvext[id[1],id[2]],digits=3)
    #chis <- 0
    return(paste0("p value: ",pv))
  })

  output$assoc2n <- renderText({
    n <-
sum(xtabs(data=data.frame(t(rbind(clas78val[as.numeric(aout()[1,1]),
],clas78val[as.numeric(aout()[2,1]),])))))
    return(paste0("n = ",n))
  })

  dta2 <- function(x,q) {
    dat <- t(rbind(ma78[x,],clas78val[x,],clas78val[q,]))
    rownames(dat) <- colnames(clas78val)
    colnames(dat) <- c("Output Expression Value","Output
Subgroup","Query Subgroup")
    return(dat)
  }

  output$assoc2table <- DT::renderDataTable({ # figure out how to
make scrollY reactive to window size (width/3?)

return(datatable(dta2(as.numeric(aout()[2,1]),as.numeric(aout()[1,1]
)),options=list(scrollY = '300px',paging=FALSE)))
  })

  # vector of output ids, ordered by |r|
  cout <- eventReactive(input$submitcor,{
    if (class(aout())=="integer") {
      qt <- as.numeric(aout()[1])
      qr <- as.numeric(abs(phiext[which(rownames(phiext)==qt),]))
      names(qr) <- rownames(phiext)
      mdd <- welcht[welcht>=input$welchts[1] &
welcht<=input$welchts[2]]
      mx <- maxes[maxes>=input$maxcutoff]
      md <- numeric()
      k <- 1
      for (i in 1:length(mx)) {
        if (length(which(names(mdd)==names(mx)[i]))!=0) {
```

```
          md[k] <- mdd[which(names(mdd)==names(mx)[i])]
          names(md)[k] <- names(mx)[i]
          k <- k+1
        }
      }
      qc <- numeric()
      j <- 1
      for (i in 1:length(md)) {
        if (length(which(names(qr)==names(md)[i]))!=0) {
          qc[j] <- qr[which(names(qr)==names(md)[i])]
          names(qc)[j] <- names(md)[i]
          j <- j+1
        }
      }
      qext <- qr[names(qr)>54675]
      for (i in 1:length(qext)) {
        qc[j] <- qr[which(names(qr)==names(qext)[i])]
        names(qc)[j] <- names(qext)[i]
        j <- j+1
      }
      qc <- qc[order(-qc)]
      qc <- qc[names(qc)!=qt]
      nm <- as.numeric(names(qc)[c(1:input$topn)])
      return(nm)
    } else if
(tryCatch({as.numeric(aout()[2,1])==0},error=function(e){FALSE})) {
      qt <- as.numeric(qout()[2,])
      qt <- qt[qt>0]
      nm <- numeric()
      j <- 1
      for (i in 1:length(qt)) {
        id <- as.numeric(details[qt[i],1])
        if (max(clas78val[id,],na.rm=TRUE)==2){
          rows <-
c(which(rownames(phiext)==as.numeric(aout()[1,1])),which(rownames(ph
iext)==as.numeric(id)))
          nm[j] <- abs(phiext[rows[1],rows[2]])
          names(nm)[j] <- id
          j <- j+1
        }
      }
      nm <- nm[order(-nm)]
      nm <- nm[names(nm)!=aout()[1,1]]
      nmn <- as.numeric(names(nm))
      return(nmn)
    } else {return(0)}
  })

  lout <- reactive({
    if
(tryCatch({class(cout())=="numeric"},error=function(e){FALSE})) {
      return(length(cout()))
    } else {return(0)}
  })

  # output for single query/pairwise query, >1 probeset
  output$assoc1 <- renderUI({
    if ((class(aout())=="integer" &
max(clas78val[aout()[1],],na.rm=TRUE)!=1) | tryCatch({aout()[2,1]==0
&
max(clas78val[aout()[1,1],],na.rm=TRUE)!=1},error=function(e){FALSE}
)) {
      lapply(1:lout(), function(i) {
        uiOutput(paste0("qo",i))
      })
    } else if (class(aout())=="integer" |
tryCatch({aout()[2,1]==0},error=function(e){FALSE})) {
      fluidRow(
        uiOutput("associchiserr")
```

```
      )
    } else {return()}
  })

  output$associchiserr <- renderUI({
    h4("Unimodal Query - Cannot Calculate Association")
  })

  observe(lapply(1:lout(), function(i) {
    output[[paste0("qo",i)]] <- renderUI({
      wellPanel(style = "background-color: #e8efff;",
                h4(paste0("Output ",i," - ",details[cout()[i],3]," -
",details[cout()[i],2])),
                h4(details[cout()[i],4]),
                fluidRow(
                  column(6,
                         tabsetPanel(
                           tabPanel("Rank Plot",
plotOutput(paste0("assocrank",i))),
                           tabPanel("Rankit Plot",
plotOutput(paste0("assocrankit",i))),
                           tabPanel("Histogram",
plotOutput(paste0("assochist",i)))
                         )),
                  column(6,
                         tabsetPanel(
                           tabPanel("Chi-squared",
uiOutput(paste0("assocchisq",i))),
                           tabPanel("Data Table",
DT::dataTableOutput(paste0("assoctable",i)))
                         ))
                )
      )
    })
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assocrank",i)]] <- renderPlot({
      if (class(aout())=="integer") {

return(plot78vQ(as.numeric(aout()[1]),as.numeric(cout()[i])))
      } else {

return(plot78vQ(as.numeric(aout()[1,1]),as.numeric(cout()[i])))
      }
    })
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assocrankit",i)]] <- renderPlot({
      if (class(aout())=="integer") {

return(plot78rivQ(as.numeric(aout()[1]),as.numeric(cout()[i])))
      } else {

return(plot78rivQ(as.numeric(aout()[1,1]),as.numeric(cout()[i])))
      }
    })
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assochist",i)]] <- renderPlot(
      return(plot78hist(as.numeric(cout()[i])))
    )
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assocchisq",i)]] <- renderUI({
        fluidRow(
```

```
          br(),
          br(),
          tableOutput(paste0("assoccontingency",i)),
          textOutput(paste0("assocr",i)),
          textOutput(paste0("assocp",i)),
          textOutput(paste0("assocn",i))
        )
    })
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assoccontingency",i)]] <- renderTable({
      if (class(aout())=="integer") {
        ct <-
as.table(xtabs(data=data.frame(t(rbind(clas78val[as.numeric(aout()[1
]),],clas78val[as.numeric(cout()[i]),])))))
        names(attributes(ct)$dimnames) <-
c(details[aout()[1],3],details[cout()[i],3])

return(data.frame(format(ftable(ct),method="compact",quote=FALSE)))
      } else {
        ct <-
as.table(xtabs(data=data.frame(t(rbind(clas78val[as.numeric(aout()[1
,1]),],clas78val[as.numeric(cout()[i]),])))))
        names(attributes(ct)$dimnames) <-
c(details[aout()[1,1],3],details[cout()[i],3])

return(data.frame(format(ftable(ct),method="compact",quote=FALSE)))
      }
    },rownames=FALSE,colnames=FALSE)
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assocr",i)]] <- renderText({
      if (class(aout())=="integer") {
        id <-
c(which(rownames(phiext)==as.numeric(aout()[1])),which(rownames(phie
xt)==as.numeric(cout()[i])))
        phi <- signif(phiext[id[1],id[2]],digits=3)
        #chis <- 0
        return(paste0("r value: ",phi))
      } else {
        id <-
c(which(rownames(phiext)==as.numeric(aout()[1,1])),which(rownames(ph
iext)==as.numeric(cout()[i])))
        phi <- signif(phiext[id[1],id[2]],digits=3)
        #chis <- 0
        return(paste0("r value: ",phi))
      }
    })
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assocp",i)]] <- renderText({
      if (class(aout())=="integer") {
        id <-
c(which(rownames(chispvext)==as.numeric(aout()[1])),which(rownames(c
hispvext)==as.numeric(cout()[i])))
        pv <- signif(chispvext[id[1],id[2]],digits=3)
        #chis <- 0
        return(paste0("p value: ",pv))
      } else {
        id <-
c(which(rownames(chispvext)==as.numeric(aout()[1,1])),which(rownames
(chispvext)==as.numeric(cout()[i])))
        pv <- signif(chispvext[id[1],id[2]],digits=3)
        #chis <- 0
        return(paste0("p value: ",pv))
      }
```

```
    })
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assocn",i)]] <- renderText({
      if (class(aout())=="integer") {
        n <-
sum(xtabs(data=data.frame(t(rbind(clas78val[as.numeric(aout()[1]),],
clas78val[as.numeric(cout()[i]),])))))
        return(paste0("n = ",n))
      } else {
        n <-
sum(xtabs(data=data.frame(t(rbind(clas78val[as.numeric(aout()[1,1]),
],clas78val[as.numeric(cout()[i]),])))))
        return(paste0("n = ",n))
      }
    })
  }))

  observe(lapply(1:lout(), function(i) {
    output[[paste0("assoctable",i)]] <- DT::renderDataTable({
      if (class(aout())=="integer") {

return(datatable(dta2(as.numeric(cout()[i]),as.numeric(aout()[1])),o
ptions=list(scrollY = '300px',paging=FALSE)))
      } else {

return(datatable(dta2(as.numeric(cout()[i]),as.numeric(aout()[1,1]))
,options=list(scrollY = '300px',paging=FALSE)))
      }
    })
  }))

})
```

Table A.5: List of the 200 top probesets most associated with the lumen formation phenotype in the cell lines, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Phenotype	*r*	p-value
Lumen Formation 1+2 vs 3	1	6.77E-10
Lumen Formation 1 vs 2+3	1	6.77E-10
TINAG	0.619	0.000214
PRKCQ	-0.573	0.000669
PRKCDBP	-0.548	0.00151
Goblet Cells Confirmed	0.527	0.2
FOXA1	-0.523	0.00208
NOV	-0.516	0.00275
NPNT	0.509	0.00369
TST	0.483	0.00476
IGFBP6	-0.476	0.00556
C16orf5	0.476	0.00556
AIFM3	0.476	0.00556
PRKCQ	-0.474	0.0067
LOC100506119 /// LOC100508342	-0.474	0.0067
DPEP1	0.474	0.00571
CDK14	-0.469	0.00866
CDK14	-0.469	0.00866
TFAP2A	-0.46	0.00862
LARS	0.46	0.00862
CRNDE	-0.46	0.00862
CEA lysis	0.455	0.0105
BANK1	-0.452	0.00962
NKX3-1	-0.452	0.00962
IGDCC4	-0.452	0.00962
AK4	-0.44	0.011
FABP1	0.44	0.011
GSPT2	0.438	0.012
CDX1	0.438	0.012
PMP22	-0.438	0.012
RASSF5	-0.438	0.012
DACH1	0.438	0.012
C6orf168	-0.438	0.012
C13orf18	0.432	0.0128
---	0.432	0.0128
MOGAT3	0.432	0.0128
ABR	0.432	0.0155
MMP24	-0.432	0.0155
SAV1	-0.432	0.0155
LOXL3	-0.432	0.0155
FAM46B	-0.432	0.0155

GJA3	-0.432	0.0155
ITPR2	0.43	0.0135
CLDN3	0.43	0.0135
RARRES1	-0.43	0.0135
ATP10B	0.43	0.0135
VAV3	0.43	0.0135
VAV3	0.43	0.0135
IKBIP	-0.43	0.0135
---	-0.428	0.0196
PHLDB2	-0.428	0.0196
HLA-DOB	-0.428	0.0196
RGS7	-0.428	0.0196
SAA1 /// SAA2	-0.428	0.0196
DIRAS3	-0.428	0.0196
PLEKHO1	-0.428	0.0196
HNRNPH2	-0.428	0.0196
UBASH3B	-0.427	0.0139
KSR2	-0.427	0.0142
SLC39A5	0.427	0.0142
NMU	-0.426	0.0188
---	0.426	0.0188
---	0.426	0.0188
SLC6A4	0.426	0.0188
C17orf77	0.421	0.018
ALDH5A1	-0.421	0.018
CAB39L	0.421	0.018
CAB39L	0.421	0.018
C17orf77	0.421	0.018
MME	0.412	0.0188
MEIS2	-0.412	0.0188
---	0.412	0.0188
RASSF5	-0.408	0.0219
ZNF321	-0.408	0.0219
PLAT	-0.408	0.0219
QPRT	0.408	0.0219
CDX2	0.408	0.0219
CHN1	-0.408	0.0219
MT1E	-0.408	0.0219
LIX1L	-0.408	0.0219
MCC	-0.408	0.0219
CEA binding	0.4	0.0269
CAP2	-0.398	0.0236
LOXL1	-0.393	0.027
USH1C	0.393	0.027
DDC	0.393	0.027
GLI2	-0.393	0.027

DDC	0.393	0.027
SGK2	0.393	0.027
CALHM2	-0.393	0.027
SLC7A2	-0.393	0.027
TMPRSS2	0.393	0.027
SLC27A1	-0.393	0.027
TPBG	-0.393	0.027
MME	0.392	0.0279
SLC6A4	0.392	0.0279
NRXN3	0.392	0.0279
---	-0.392	0.0279
SHH	0.392	0.0279
LOC339260	0.389	0.0374
TG	0.389	0.0374
APOA2	0.389	0.0374
APOA2	0.389	0.0374
---	0.389	0.0374
---	0.389	0.0374
HPS4	0.389	0.0374
LOC339260	0.389	0.0374
ALDH1A3	-0.388	0.0277
SLC22A18	0.388	0.0277
IRX5	-0.388	0.0277
NOX1	0.388	0.0277
LRRC2	0.388	0.0277
RGMB	0.388	0.0277
CR2	-0.387	0.0343
ANKRD1	-0.387	0.0343
MAPT	-0.387	0.0343
CNGA1	-0.387	0.0343
HIST2H4A /// HIST2H4B	-0.387	0.0343
SH3BGRL2	0.387	0.0343
LOXL4	-0.387	0.0343
TUSC1	0.387	0.0343
---	-0.387	0.0343
LOC100506831	-0.387	0.0343
---	-0.387	0.0343
KLF12	-0.387	0.0343
LOC283404	-0.386	0.0432
MAPT	-0.386	0.0432
SMARCD3	-0.386	0.0432
DCLK1	-0.386	0.0432
HOXD9	-0.386	0.0432
SAA4	-0.386	0.0432
---	-0.386	0.0432
IL1RN	-0.386	0.0432

HHAT	-0.386	0.0432
TSPAN14	-0.386	0.0432
KRTAP1-5	-0.386	0.0432
SAMD9L	-0.386	0.0432
---	-0.386	0.0432
CEACAM5	0.384	0.0304
CYP2B6 /// CYP2B7P1	0.384	0.0304
CYP2B6	0.384	0.0304
VIL1	0.384	0.0304
PHGR1	0.384	0.0304
---	0.384	0.0304
GULP1	-0.382	0.0359
HOXA10 /// HOXA9	-0.382	0.0359
HOXA10 /// HOXA9	-0.382	0.0359
SLC19A3	0.382	0.0359
KIF13A	-0.382	0.0359
---	0.382	0.0359
LOC145837	0.382	0.0359
ACSL6	0.381	0.0306
IDS	-0.381	0.0306
AASS	-0.381	0.0306
UGCG	-0.381	0.0306
C9orf125	-0.381	0.0306
AGXT2L2	0.381	0.0306
PPP1R9A	-0.381	0.0306
TBC1D1	-0.381	0.0306
CEACAM6	0.379	0.0321
ASCL2	0.379	0.0321
B3GALNT1	-0.379	0.0321
ST3GAL1	-0.379	0.0321
C20orf112	0.379	0.0321
SHISA9	0.379	0.0321
RPS6KA6	0.379	0.0321
PRAP1	0.379	0.0321
ADM2	0.371	0.0379
ZNF300	0.371	0.0379
---	0.371	0.0379
TRIB2	-0.362	0.047
ARNT2	-0.362	0.047
KLK7	-0.362	0.047
TUBB2A /// TUBB2B	-0.362	0.047
APOL1	-0.362	0.047
MCAM	-0.362	0.047
MCAM	-0.362	0.047
IDS	-0.362	0.047
SLC43A3	-0.362	0.047

CDKN1C	-0.362	0.047
PDLIM5	-0.362	0.047
TOX3	0.362	0.047
TOX3	0.362	0.047
TOX3	0.362	0.047
EPB41L4B	0.362	0.047
KATNAL1	-0.362	0.047
MRAS	-0.362	0.047
---	-0.362	0.047
SPHK2	0.362	0.047
NRXN3	0.355	0.0472
PTPRD	0.355	0.0472
C9orf125	-0.355	0.0472
CYP39A1	0.355	0.0472
USP7	0.355	0.0472
Goblet Cells Broad (+/- as negative)	0.354	0.185
ETV5	-0.351	0.0543
EPB41	0.351	0.0543
MAGEA5	0.351	0.0543
CEP55	-0.351	0.0543
TMPRSS13	0.351	0.0543

Table A.6: List of the 200 top probesets most associated with the *CDX1* gene in the cell lines, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Phenotype	*r*	p-value
ATP10B	0.867	1.14E-13
TMPRSS2	0.867	1.22E-13
GPA33	0.842	5.80E-13
HEPH	0.841	6.82E-13
CDX2	0.796	1.21E-11
GUCY2C	0.795	1.14E-11
CEA lysis	0.791	2.73E-11
CDH17	0.79	1.76E-11
CCL14-CCL15 /// CCL15	0.79	1.76E-11
CYP2B6	0.789	1.62E-11
CYP2B6 /// CYP2B7P1	0.786	2.03E-11
CLRN3	0.786	2.13E-11
VAV3	0.772	4.49E-11
VAV3	0.772	4.49E-11
PPP1R14D	0.76	9.26E-11
ASCL2	0.76	1.03E-10
DDC	0.758	1.07E-10
DDC	0.758	1.07E-10
PIP5K1B	0.758	1.12E-10
PHGR1	0.758	1.12E-10
C10orf99	0.752	1.47E-10
SLC39A5	0.75	1.64E-10
AIFM3	0.75	1.64E-10
TMPRSS2	0.739	3.58E-10
CEA binding	0.734	7.13E-10
USH1C	0.733	4.69E-10
PPP1R1B	0.733	4.69E-10
DPEP1	0.733	4.42E-10
CDHR1	0.733	4.42E-10
C10orf99	0.733	4.42E-10
CFTR	0.732	4.62E-10
SLC3A1	0.732	4.62E-10
FOXA3	0.731	5.29E-10
CEACAM5	0.73	5.25E-10
SGK2	0.73	5.25E-10
PPARA	0.73	5.25E-10
ATP8B1	0.73	5.25E-10
PPARA	0.73	5.25E-10
LRRC19	0.714	1.26E-09
CLDN3	0.714	1.50E-09
MYO1A	0.713	1.36E-09

TDGF1 /// TDGF3	0.707	1.83E-09
NR1I2	0.707	1.79E-09
ADD3	0.704	2.21E-09
FABP1	0.704	2.21E-09
C2orf89	0.704	2.21E-09
LGALS4	0.704	2.35E-09
USH1C	0.704	2.35E-09
CALML4	0.704	2.35E-09
MLLT3	0.703	2.41E-09
WWC2	-0.698	3.05E-09
ALDH2	0.688	5.95E-09
FUT2	0.688	5.95E-09
TMEM45B	0.688	5.95E-09
ASCL2	0.683	6.63E-09
VAV3	0.68	7.77E-09
ABCB1 /// ABCB4	0.679	8.49E-09
DCBLD2	-0.679	8.49E-09
VIL1	0.677	9.84E-09
C9orf152	0.677	9.84E-09
ETS2	0.676	9.88E-09
ADD3	0.676	9.88E-09
PLEKHH1	0.676	9.88E-09
VIL1	0.676	9.88E-09
NOX1	0.676	9.53E-09
NOX1	0.676	9.53E-09
NOX1	0.676	9.53E-09
TRIM31	0.676	1.04E-08
CALML4	0.676	1.04E-08
LIX1L	-0.676	1.04E-08
SH3BGRL2	0.674	1.33E-08
MYB	0.667	1.52E-08
CCDC88A	-0.667	1.52E-08
MYO5A	-0.667	1.52E-08
IL33	0.667	1.56E-08
FRMD6	-0.667	1.56E-08
ERBB3	0.663	2.26E-08
HNMT	0.663	2.26E-08
IFITM1	0.663	2.26E-08
MUC13	0.663	2.26E-08
PRR15L	0.663	2.26E-08
TGFB2	-0.663	2.26E-08
---	0.663	2.26E-08
ATP10B	0.661	2.08E-08
---	0.66	2.23E-08
C13orf18	0.66	2.23E-08
NAT2	0.66	2.29E-08

ACSL6	0.66	2.29E-08
IFITM1	0.655	3.19E-08
TSPAN8	0.655	3.19E-08
EPS8L3	0.655	3.19E-08
BCL2L14	0.655	3.19E-08
ERBB3	0.655	3.19E-08
LOC100505806 /// SNORD123	0.655	2.81E-08
SLC39A5	0.655	2.81E-08
IHH	0.654	2.99E-08
MGAT4A	0.651	3.90E-08
TMEM45B	0.651	3.90E-08
VDR	0.65	3.67E-08
AKR1C3	0.65	3.67E-08
CFTR	0.65	3.67E-08
WNK4	0.65	3.67E-08
GATA6	0.65	4.74E-08
LGR5	0.65	4.74E-08
AUTS2	0.648	4.23E-08
GPX2	0.648	4.13E-08
LOC100507192	0.648	4.13E-08
SGK2	0.648	4.13E-08
C6orf168	-0.648	4.13E-08
C13orf18	0.647	4.25E-08
ITGA9	0.647	4.25E-08
MYO15B	0.645	4.62E-08
NR1I2	0.642	5.33E-08
PTPRO	0.642	5.33E-08
EPB41L4B	0.638	8.20E-08
MUC13	0.638	8.20E-08
KATNAL1	-0.638	8.20E-08
LOC399959	-0.638	8.20E-08
PRR15	0.638	8.20E-08
FAM171B	-0.638	8.20E-08
CKMT1A /// CKMT1B	0.637	6.98E-08
DACH1	0.637	6.98E-08
STARD10	0.637	6.98E-08
IDS	-0.634	7.84E-08
FAM84A	0.634	7.84E-08
IKBIP	-0.634	7.84E-08
FAM84A	0.634	7.84E-08
GGT6	0.634	7.84E-08
ACOX2	0.63	9.73E-08
LTBP2	-0.626	1.62E-07
AKT3	-0.626	1.62E-07
PPFIBP2	0.626	1.62E-07
ETS2	0.625	1.25E-07

WWTR1	-0.625	1.25E-07
CEACAM6	0.625	1.25E-07
MECOM	0.625	1.25E-07
FABP1	0.625	1.30E-07
CYR61	-0.624	1.47E-07
TACC2	0.624	1.47E-07
ABP1	0.624	1.47E-07
CYR61	-0.624	1.47E-07
CLDN3	0.624	1.32E-07
CCDC88A	-0.624	1.32E-07
CDC42EP5	0.624	1.32E-07
FAM84A	0.624	1.32E-07
Goblet Cells Confirmed	0.623	0.00244
FOXA2	0.622	1.49E-07
CEACAM6	0.622	1.49E-07
SMURF2	-0.621	1.63E-07
DACH1	0.621	1.63E-07
ABCB1	0.621	1.63E-07
USH1C	0.621	1.63E-07
DACH1	0.614	2.05E-07
CDHR5	0.614	2.05E-07
MEP1A	0.614	2.08E-07
C17orf73	0.614	2.08E-07
UBASH3B	-0.614	2.08E-07
NOSTRIN	0.612	2.86E-07
---	-0.612	2.86E-07
SLC26A3	0.608	2.95E-07
C20orf112	0.607	2.95E-07
SOAT1	-0.607	2.95E-07
FLJ22763	0.606	3.12E-07
GPR35	0.603	4.17E-07
TMEM136	-0.603	4.17E-07
ADD3	0.603	3.58E-07
PCK1	0.603	3.58E-07
FRMD6	-0.603	3.58E-07
CADPS	0.601	3.94E-07
CAV1	-0.601	3.94E-07
SAMD5	0.601	3.94E-07
CADPS	0.601	3.94E-07
HNMT	0.601	5.37E-07
ZNF703	0.601	5.37E-07
HMGA2	-0.597	5.26E-07
TRIB2	-0.597	5.26E-07
KIAA1199	0.597	5.26E-07
GIPC2	0.597	5.26E-07
VDR	0.596	4.90E-07

TSPAN4	-0.596	4.90E-07
MECOM	0.596	4.90E-07
UBASH3B	-0.596	4.90E-07
KCNQ1	0.594	5.27E-07
FAM84A	0.594	5.27E-07
NKX3-1	-0.594	5.92E-07
ZEB1	-0.594	5.92E-07
CD109	-0.594	5.92E-07
FLJ32255	-0.594	5.92E-07
CALU	-0.594	5.65E-07
CSRP2	-0.594	5.65E-07
LOC554202	-0.593	6.04E-07
CD109	-0.593	6.04E-07
CFTR	0.593	5.68E-07
HNF4G	0.593	5.68E-07
ISX	0.591	6.50E-07
IRF8	0.589	6.64E-07
ABCB1	0.589	6.64E-07
DPYD	-0.587	9.57E-07
AGR2	0.587	9.57E-07
ATP8B2	-0.587	9.57E-07

Table A.7: List of the 200 top genes most associated with the *CDX1* gene in the TCGA primary tumors, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Transcript	*r*	p-value
AL139275.1	1	6.81E-06
HEPH	0.533	2.30E-37
CRCT1	-0.53	4.62E-06
CDX2	0.488	2.12E-32
AC025284.1	0.471	0.208
Metazoa_SRP	0.471	0.208
C6orf222	0.471	5.22E-29
HNF4A	0.471	9.55E-29
CEBPA-AS1	0.468	2.95E-29
AC005833.2	0.467	2.09E-29
TRABD2A	0.467	1.02E-29
FAM84A	0.462	3.45E-29
GUCY2C	0.45	1.08E-27
LRRC75A	0.449	2.08E-27
RBMY2TP	0.435	4.73E-07
CYP2J2	0.43	5.33E-25
GPA33	0.429	4.20E-25
AP001043.1	0.424	1.55E-24
CRB2	0.424	1.26E-24
ANKS4B	0.424	2.80E-24
ALDH2	0.421	5.97E-23
CDH17	0.42	1.84E-23
ACOX2	0.417	5.81E-24
ACSS2	0.416	5.41E-24
AC124067.4	0.413	1.50E-23
PRR15L	0.41	1.67E-22
MISP	0.404	1.38E-20
IHH	0.404	1.83E-22
CD109	-0.403	1.64E-22
GGT6	0.402	3.20E-22
B3GNT8	0.401	2.21E-22
ATP10B	0.4	4.52E-22
MYB	0.399	9.89E-22
AL109613.1	0.398	2.04E-21
NAT2	0.397	1.22E-21
AC112251.1	-0.397	0.000553
OVOL1	0.394	3.63E-21
MUC20	0.393	8.69E-21
EPS8L3	0.39	2.70E-20
AUTS2	0.388	1.26E-20
SHD	0.386	2.88E-20

ITGA10	-0.383	1.34E-18
PPFIBP2	0.382	2.88E-20
AC106876.1	0.38	1.24E-19
Z83839.2	0.379	0.234
SH3RF2	0.378	7.31E-20
PIP5K1B	0.378	5.14E-19
CLRN3	0.378	5.14E-19
METTL7B	0.378	2.72E-19
MAB21L3	0.376	1.86E-19
AL359313.1	-0.375	0.00202
LINC00675	0.372	4.91E-19
PLEKHJ1	0.372	8.14E-19
LRRC19	0.371	2.81E-19
LCE3D	-0.37	0.115
MYO7B	0.367	8.63E-18
CFTR	0.365	1.10E-18
SLC22A5	0.364	6.83E-18
AC021218.1	0.364	2.98E-18
U3	-0.362	2.40E-08
DACH1	0.36	2.75E-18
MAP2	-0.358	1.45E-17
LINC00686	-0.355	0.187
MAOA	0.35	1.64E-16
VAV3	0.35	2.16E-17
SNHG18	0.349	8.10E-17
AL024508.2	0.349	1.01E-16
LINC01940	-0.348	9.50E-06
LINC01637	0.348	1.29E-15
DDC	0.346	5.95E-17
KRT41P	0.346	0.0117
MIR3162	-0.345	0.186
DAGLA	0.344	1.00E-16
WNK4	0.343	2.92E-16
LCK	-0.343	2.92E-16
GATA6	0.342	8.19E-15
LGALS4	0.342	8.37E-16
LINC00973	-0.342	2.07E-06
NR1I2	0.338	2.80E-16
AC007207.1	0.336	1.53E-14
CDHR1	0.334	5.64E-16
CAV2	-0.334	1.25E-15
PLEKHG6	0.333	1.11E-15
WNK2	0.333	1.97E-14
PHGR1	0.333	1.97E-14
FHIT	0.33	2.64E-15
LRRC2	0.33	1.88E-15

PPP1R14D	0.328	2.16E-15
BCAR3	-0.328	1.32E-12
PRR5L	0.328	7.65E-15
PCK1	0.327	2.01E-14
A2ML1	-0.327	1.68E-11
PDZD3	0.325	6.85E-15
GLOD5	0.324	2.08E-14
MIR194-2HG	0.324	1.29E-14
USH1C	0.323	2.34E-14
TRAP1	0.323	2.34E-14
GPX2	0.323	2.34E-14
SHH	0.323	8.34E-15
MARS2	0.323	1.38E-14
VIL1	0.322	2.06E-14
CALML4	0.322	2.06E-14
IL23R	0.321	4.06E-13
KRT18P28	0.321	4.23E-11
CEBPA	0.32	1.05E-14
CLDN3	0.319	2.13E-14
LINC00114	0.319	4.94E-14
GPR35	0.319	1.20E-12
CSMD2-AS1	0.319	0.333
GAL3ST2	0.319	1.13E-14
ZNF703	0.319	2.90E-14
CEACAM5	0.318	1.60E-14
FCAMR	0.318	5.54E-14
EPHB2	0.317	5.31E-14
TEX45	0.317	5.31E-14
HNF1B	0.317	6.92E-14
HNF1A-AS1	0.312	4.66E-13
AC124798.1	0.312	4.66E-13
MYO1A	0.311	1.47E-13
PLA2G4F	0.311	1.56E-13
VIPR1	0.31	7.82E-14
IGHD2-8	-0.309	0.0354
ELK3	-0.309	1.26E-13
AOC1	0.308	1.66E-12
TJP3	0.308	1.66E-12
PPARA	0.307	2.32E-13
BDH1	0.307	3.10E-13
TMEM253	0.306	3.48E-13
GPR160	0.306	1.71E-13
STXBP6	0.303	1.42E-12
PLK2	-0.303	6.97E-13
CDHR5	0.303	2.25E-12
AC120498.1	-0.303	0.29

TSPAN12	0.303	6.36E-13
TDGF1	0.302	2.86E-13
RUBCNL	0.302	2.76E-13
PDSS1	0.3	1.72E-12
TFF3	0.3	1.72E-12
C9orf152	0.3	1.72E-12
EIF4EBP3	0.3	8.25E-13
GLYATL1	0.299	1.30E-12
VIPR1-AS1	0.298	1.07E-12
FAR2	0.298	9.87E-13
TTC39A	0.298	9.87E-13
CYP2B6	0.297	8.30E-13
FAAH	0.297	1.51E-12
CX3CL1	-0.297	1.03E-11
DCBLD1	-0.296	1.65E-10
PPP1R1B	0.295	2.46E-12
EPHB6	-0.295	2.59E-12
RBBP8NL	0.294	9.41E-12
TKT	0.294	3.22E-12
NIPAL1	0.294	1.32E-12
AC120498.3	-0.294	0.206
CXCL14	0.294	2.28E-12
FRAT2	0.292	2.29E-12
AC009014.1	0.292	5.40E-12
AIFM3	0.291	1.92E-12
SLC3A1	0.29	2.59E-12
LINC01224	0.29	1.09E-11
KCTD14	0.289	1.46E-11
C10orf99	0.288	3.39E-12
AP001790.1	-0.288	0.00368
MUC13	0.286	5.25E-11
RAPGEFL1	0.286	3.47E-11
RFX5	-0.286	3.47E-11
ZSWIM5	0.286	1.31E-11
LPAR5	0.286	1.31E-11
JHDM1D-AS1	0.285	7.74E-12
CD24	0.283	2.93E-11
WFDC10B	0.283	5.64E-07
TPRN	0.283	8.15E-12
FAM171A1	0.283	4.13E-11
TMEM139	0.283	4.13E-11
MUC20P1	0.283	4.33E-11
CASP6	0.282	9.68E-12
PARM1	0.282	2.52E-11
TOX3	0.281	2.17E-10
MYC	0.281	2.17E-10

BCL11A	0.281	2.32E-11
RGS17	-0.28	2.75E-10
AC254629.1	0.28	1.48E-11
CAPN13	0.279	2.31E-11
AC010641.1	-0.279	0.401
AC009065.2	0.279	1.15E-10
ITGB6	0.278	6.78E-11
INSC	0.277	4.20E-11
PAPSS2	0.277	2.25E-10
MNX1	0.277	1.10E-10
HDHD3	0.277	3.95E-11
DSG4	0.277	5.14E-10
EMID1	0.277	3.24E-11
ISX	0.276	3.52E-11
AC023347.1	0.275	0.138
CCAT1	0.275	6.17E-11
LOXL3	-0.274	4.99E-11
KCNQ1	0.274	8.20E-11
AP000424.2	-0.273	0.0574
AC087491.1	0.272	6.13E-11
AC124248.1	0.271	1.00E-10

Table A.8: List of the 200 top probesets most associated with the goblet cell formation phenotype in the cell lines, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Phenotype	*r*	p-value
Goblet Cells Broad (+/- as negative)	1	3.69E-07
FCGBP	0.793	9.14E-05
MUC2	0.793	9.14E-05
ZDHHC8P1	0.791	8.01E-05
AMN	0.783	0.000103
RSPH1	0.783	0.000103
HNF4G	0.731	0.000296
IDS	-0.713	0.000445
REG4	0.713	0.000445
SMURF2	-0.712	0.000434
C2orf89	0.712	0.000434
AGR3	0.712	0.000434
SMURF2	-0.712	0.000434
MYO5A	-0.675	0.000913
MYO5A	-0.675	0.000913
FOXA2	0.675	0.000913
CEA lysis	0.675	0.000913
HEPH	0.665	0.00111
CFTR	0.665	0.00111
DDC	0.665	0.00111
AKR1B10	0.665	0.00111
CYP2B6 /// CYP2B7P1	0.665	0.00111
CA12	0.665	0.00111
DDC	0.665	0.00111
CFTR	0.665	0.00111
ATP8B1	0.665	0.00111
ENTPD5	0.657	0.00177
HSPA8	-0.657	0.00177
UGT1A10 /// UGT1A7 /// UGT1A8	0.657	0.00177
IDS	-0.649	0.00152
UGT1A1 /// UGT1A10 /// UGT1A4 /// UGT1A6 /// UGT1A8 /// UGT1A9	0.649	0.00152
NR3C2	0.649	0.00152
UGT1A1 /// UGT1A10 /// UGT1A3 /// UGT1A4 /// UGT1A5 /// UGT1A6 /// UGT1A7 /// UGT1A8 /// UGT1A9	0.649	0.00152
UGT1A1 /// UGT1A10 /// UGT1A4 /// UGT1A6 /// UGT1A8 /// UGT1A9	0.649	0.00152
UGT1A1 /// UGT1A10 /// UGT1A3 /// UGT1A4 /// UGT1A5 /// UGT1A6 /// UGT1A7 /// UGT1A8 /// UGT1A9	0.649	0.00152

UGT1A1 /// UGT1A10 /// UGT1A3 /// UGT1A4 /// UGT1A5 /// UGT1A6 /// UGT1A7 /// UGT1A8 /// UGT1A9	0.649	0.00152
CDC42EP5	0.649	0.00152
ANXA6	-0.636	0.00225
REG4	0.636	0.00206
WWTR1	-0.636	0.00206
RTN4RL1	0.636	0.00206
ETS1	-0.623	0.00244
SAMD5	0.623	0.00244
ETS2	0.623	0.00244
CHST15	-0.623	0.00244
CYP3A5	0.623	0.00244
CDX1	0.623	0.00244
CYP2B6	0.623	0.00244
IL33	0.623	0.00244
CYP3A5	0.623	0.00244
CFTR	0.623	0.00244
SMPD1	-0.623	0.00244
MECOM	0.623	0.00244
MECOM	0.623	0.00244
AIFM3	0.623	0.00244
SAMD5	0.623	0.00244
CEACAM6	0.621	0.00256
CLDN3	0.621	0.00256
CA12	0.621	0.00256
FABP1	0.621	0.00256
XK	0.621	0.00256
CCL14-CCL15 /// CCL15	0.621	0.00256
CEACAM6	0.621	0.00256
PPP1R1B	0.621	0.00256
MOBKL2B	0.621	0.00256
CLRN3	0.621	0.00256
PCK1	0.591	0.00431
AKAP12	-0.591	0.00431
ATP10B	0.591	0.00431
CEA binding	0.591	0.00431
C11orf53	0.588	0.00671
ALOX5	0.588	0.00671
HSPA8	-0.588	0.00671
SPRR1A	0.588	0.00671
ALOX5	0.588	0.00671
SPDEF	0.588	0.00671
HSPA8	-0.588	0.00671
---	0.588	0.00671
---	0.588	0.00671

MLLT3	0.579	0.00547
AUTS2	0.579	0.00547
BCL2L14	0.579	0.00547
LOC399959	-0.579	0.00547
VASN	-0.579	0.00547
TMEM136	-0.579	0.00547
SGK2	0.579	0.00547
SLC39A5	0.573	0.00579
ETS2	0.573	0.00579
WWTR1	-0.573	0.00579
WWTR1	-0.573	0.00579
GJB1	0.573	0.00579
NOX1	0.573	0.00579
NOX1	0.573	0.00579
LGR5	0.573	0.00579
ATP10B	0.573	0.00579
CA12	0.573	0.00579
CLDN2	0.573	0.00579
FAM108C1	0.573	0.00579
AKAP12	-0.573	0.00579
ELF3	0.573	0.00579
SYTL5	0.573	0.00579
PEX11A	0.569	0.00642
PF4	0.569	0.00642
ALDH1A1	0.569	0.00642
OLFM4	0.569	0.00642
RUNDC3B	0.569	0.00642
AHCYL2	0.569	0.00642
TIMP2	-0.569	0.00642
---	0.562	0.0086
UGT2B17	0.562	0.0086
CYP4X1	0.562	0.0086
---	0.562	0.0086
C11orf92	0.562	0.0086
C11orf41	0.562	0.0086
---	0.562	0.0086
CDHR1	0.558	0.00815
DAPK2	0.558	0.00815
CYP39A1	0.558	0.00815
PIGR	0.558	0.00815
LOC100128893	0.558	0.00815
CCDC75	-0.538	0.0109
PDLIM7	-0.538	0.0109
GPNMB	-0.538	0.0109
TSPAN8	0.538	0.0109
HNMT	0.538	0.0109

SLC16A2	-0.538	0.0109
DOCK4	-0.538	0.0109
USH1C	0.538	0.0109
SLC44A4	0.538	0.0109
LIMS1	-0.538	0.0109
AGR2	0.538	0.0109
GATA6	0.538	0.0109
USH1C	0.538	0.0109
MAP1B	-0.538	0.0109
PRR15L	0.538	0.0109
HOOK1	0.538	0.0109
ERBB3	0.538	0.0109
TMEM45B	0.538	0.0109
---	-0.538	0.0109
TMEM136	-0.538	0.0109
PDLIM7	-0.538	0.0109
CCND2	0.536	0.0106
CKMT1A /// CKMT1B	0.536	0.0106
SLC12A2	0.536	0.0106
MYB	0.536	0.0106
GPA33	0.536	0.0106
ATP10D	-0.536	0.0106
HHLA2	0.536	0.0106
UGT1A8 /// UGT1A9	0.536	0.0106
SLC12A2	0.536	0.0106
IKBIP	-0.536	0.0106
---	0.536	0.0106
Lumen Formation 1 (no) vs 3 (yes)	0.527	0.2
CALU	-0.526	0.0125
CYR61	-0.526	0.0125
CEACAM5	0.526	0.0125
TRIM31	0.526	0.0125
AZGP1	0.526	0.0125
ABCB1 /// ABCB4	0.526	0.0125
FOXA2	0.526	0.0125
CA12	0.526	0.0125
CYR61	-0.526	0.0125
AZGP1	0.526	0.0125
SGK2	0.526	0.0125
LRRC31	0.526	0.0125
LOC100507192	0.526	0.0125
RARRES1	0.526	0.0125
CALML4	0.526	0.0125
PPARA	0.526	0.0125
SCARA3	-0.526	0.0125
TMPRSS2	0.526	0.0125

PPARA	0.526	0.0125
PHGR1	0.526	0.0125
PRUNE2	0.515	0.0234
C9orf71	0.515	0.0234
FLJ35024	0.515	0.0234
LRRC66	0.515	0.0234
PIK3C2A	-0.515	0.0234
---	0.515	0.0234
CAPRIN1	-0.515	0.0234
CELSR1	0.515	0.0234
CYP2C18	0.515	0.0234
MYH4	0.515	0.0234
CABP1	0.515	0.0234
BAT3	-0.515	0.0234
NAT8	0.515	0.0234
---	0.515	0.0234
HSPA8	-0.515	0.0234
SLC23A1	0.515	0.0234
GPT2	0.515	0.0234
---	0.515	0.0234
DQX1	0.515	0.0234
L1TD1	0.515	0.0234
NCRNA00238	0.515	0.0234
C11orf93	0.515	0.0234
SMPDL3A	0.515	0.0234
LRRC66	0.515	0.0234
PIK3C2A	-0.515	0.0234
---	0.515	0.0234
WWC2	-0.508	0.0165

Table A.9: List of the 200 top probesets most associated with the *MUC2* gene in the cell lines, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Phenotype	*r*	p-value
Goblet Cells Confirmed	0.793	9.14E-05
Goblet Cells Broad (+/- as negative)	0.614	3.98E-05
REG4	0.584	1.17E-06
OLFM4	0.567	2.48E-06
FCGBP	0.552	4.86E-06
REG4	0.534	9.61E-06
LPCAT1	-0.478	0.000176
LIN7A	0.477	0.000114
IGK@ /// IGKC	0.476	0.000375
MMP28	0.47	0.000309
NCRNA00238	0.47	0.000309
SPINK4	0.469	0.000129
WWTR1	-0.468	0.000124
KLHL5	-0.462	0.00015
PRLR	0.45	0.000319
MYO5A	-0.449	0.000239
FAM3D	0.437	0.000404
ENTPD5	0.428	0.000591
HEPACAM2	0.428	0.000591
ABCC13	0.422	0.000767
RDX	-0.417	0.000714
SULT1B1	0.416	0.00152
PRLR	0.416	0.00152
---	0.416	0.00152
---	0.41	0.00407
IL17F	0.41	0.00407
---	0.41	0.00407
---	0.405	0.00118
MMP28	0.404	0.00291
DAB1	0.404	0.00291
RDX	-0.401	0.00115
---	0.401	0.00115
CREB3L1	0.398	0.00123
FOXA2	0.398	0.00123
NR3C2	0.397	0.00135
TUBB3	-0.397	0.00211
TFF1	0.397	0.00131
SIDT1	0.397	0.00131
CFTR	0.396	0.00134
B3GALNT1	-0.396	0.00134
C17orf57	0.393	0.00205

IDS	-0.387	0.00179
WWC2	-0.387	0.00179
CHST15	-0.387	0.00178
MYO5A	-0.387	0.00178
VDR	0.386	0.00186
DDC	0.386	0.00186
CYP3A5	0.386	0.00186
FABP1	0.386	0.00186
AKR1C3	0.386	0.00186
FN1	-0.386	0.00186
DDC	0.386	0.00186
CFTR	0.386	0.00186
MOBKL2B	0.386	0.00186
C2orf89	0.386	0.00186
CALHM2	-0.386	0.00186
ATOH1	0.384	0.0022
SEC16B	0.383	0.00195
RFC1	-0.381	0.00219
SEMA6A	0.38	0.00236
CEACAM5	0.375	0.00254
MAP1B	-0.375	0.00254
FN1	-0.375	0.00254
PPARA	0.375	0.00254
ATP8B1	0.375	0.00254
PPARA	0.375	0.00254
AGR3	0.375	0.00254
PAPSS2	0.375	0.00252
PAPSS2	0.375	0.00252
LRRC31	0.375	0.00252
B3GALNT1	-0.375	0.00252
---	0.373	0.00324
PRLR	0.373	0.00324
RETNLB	0.373	0.00324
---	0.373	0.00324
LGALS9	0.372	0.00273
NAT8	0.371	0.00504
CYP2C9	0.371	0.00504
CHST5	0.371	0.00504
CHST5	0.371	0.00504
IDS	-0.37	0.00284
LYZ	0.37	0.00284
CALU	-0.365	0.00344
PIP5K1B	0.365	0.00344
FN1	-0.365	0.00344
CYP3A5	0.365	0.00344
CALHM2	-0.365	0.00344

SGK2	0.365	0.00344
SMURF2	-0.365	0.00344
---	0.365	0.00436
RDX	-0.365	0.00343
EGLN3	0.364	0.00386
KIAA1324	0.364	0.00386
KIAA1324	0.364	0.00386
C10orf114	-0.363	0.00351
GJB1	0.363	0.00351
MYO1A	0.363	0.00351
FAM55A	0.363	0.00521
ADH6	0.363	0.00521
CYP2C9	0.363	0.00521
CYP2C9	0.363	0.00521
AHCYL2	0.362	0.00387
RETNLB	0.362	0.00387
ATP10B	0.358	0.00405
NR1I2	0.358	0.00407
HEPH	0.355	0.0046
FN1	-0.355	0.0046
CLRN3	0.355	0.0046
CAV1	-0.352	0.00483
GPA33	0.352	0.00483
TAOK3	0.352	0.00483
LCMT2	0.35	0.00589
CYTIP	0.35	0.00589
TM4SF4	0.35	0.00589
---	-0.35	0.00589
LYZ	0.35	0.00521
RFC2	-0.347	0.0114
B3GNT6	0.347	0.0114
OAT	-0.347	0.0114
LIPA	-0.347	0.0114
RFC2	-0.347	0.0114
OTC	0.347	0.0114
NEUROG3	0.347	0.0114
FBXO5	-0.347	0.0114
ERCC6L	-0.347	0.0114
SKA2	-0.347	0.0114
NEK6	0.347	0.0114
---	0.347	0.0114
SUGT1P3	0.347	0.0114
PHF2P1	0.347	0.0114
---	0.347	0.0114
NOX1	0.345	0.00572
NOX1	0.345	0.00572

HTR1D	0.345	0.00572
A1CF	0.345	0.00572
CDC42EP5	0.345	0.00572
RSPH1	0.345	0.00572
LOC374443	-0.345	0.00572
USH1C	0.345	0.0061
AUTS2	0.345	0.0061
TMEM67	-0.345	0.0061
PPARA	0.345	0.0061
HNRNPH2	-0.345	0.00642
CASP1	0.345	0.00642
---	-0.345	0.00642
APOBEC1	0.345	0.00587
TDGF1 /// TDGF3	0.341	0.00656
AZGP1	0.341	0.00656
CAV1	-0.341	0.00656
AZGP1	0.341	0.00656
FAM108C1	0.341	0.00656
IHH	0.341	0.00656
SMCR7	-0.341	0.00656
NID2	0.34	0.0084
PTGER4	0.34	0.0084
DNA2	-0.34	0.0084
PIGR	0.34	0.0084
---	0.34	0.0084
TRIM2	0.335	0.008
CDH17	0.335	0.008
CCL14-CCL15 /// CCL15	0.335	0.008
USH1C	0.335	0.008
PPP1R1B	0.335	0.008
SMURF2	-0.335	0.008
ZNF512B	-0.335	0.008
CELF2	0.335	0.00771
ANXA13	0.335	0.00771
NOX1	0.335	0.00771
ACSL6	0.335	0.00771
KLHL5	-0.335	0.00771
WWC2	-0.333	0.00792
---	-0.333	0.00792
CCND2	0.333	0.00792
FMO5	0.333	0.00792
AKR1B10	0.333	0.00792
CFTR	0.333	0.00792
HNF4G	0.333	0.00792
ALG5	0.333	0.011
CHEK1	-0.333	0.0128

ESR1	0.333	0.0128
AP3M1	-0.333	0.0128
LOC645166 /// LOC654342	-0.333	0.0128
---	0.333	0.0128
---	0.333	0.0128
---	0.333	0.0128
MGC34034	0.332	0.0426
LOC440602	0.332	0.0426
FOXP2	0.332	0.0426
---	0.332	0.0426
SEPT2	-0.332	0.0426
ARHGDIA	0.332	0.0426
STK25	-0.332	0.0426
ATXN2L	0.332	0.0426
SLC13A3	0.332	0.0426
CD1D	0.332	0.0426
RAB1A	-0.332	0.0426
SLC11A1	0.332	0.0426
ILF3	-0.332	0.0426
MCM3AP	0.332	0.0426
IFT57	-0.332	0.0426

Table A.10: List of the 200 top genes most associated with the *MUC2* gene in the TCGA primary tumors, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Transcript	*r*	p-value
FCGBP	0.695	1.33E-66
B3GNT6	0.671	5.43E-59
SPINK4	0.653	7.36E-59
REG4	0.632	4.55E-55
MUC4	0.586	8.14E-48
ITLN1	0.571	6.09E-45
FER1L6	0.564	2.58E-43
BEST2	0.56	2.56E-40
TPSG1	0.549	6.29E-42
HEPACAM2	0.543	8.72E-41
SPDEF	0.532	1.06E-39
CLCA1	0.513	2.25E-36
FAM177B	0.503	1.56E-34
AC139749.1	0.499	1.71E-34
SLC4A4	0.493	2.86E-34
FFAR4	0.492	3.96E-34
VSIG2	0.491	4.25E-34
CTSE	0.482	9.99E-33
ATOH1	0.481	3.49E-32
ZG16	0.473	5.60E-31
ANO7	0.462	2.46E-30
ZBTB7C	0.461	3.77E-30
B4GALNT2	0.452	2.87E-27
MUC5B	0.446	2.79E-28
TSPAN1	0.444	4.72E-28
DHRS9	0.443	7.05E-28
CA2	0.436	3.91E-27
SI	0.434	2.71E-26
PLA2G2A	0.432	1.14E-26
CHST5	0.431	2.72E-26
PYY	0.417	1.15E-22
PIGR	0.416	9.32E-25
OR6F1	0.414	0.023
PLAC8	0.411	3.14E-24
IGHA2	0.41	3.91E-24
MB	0.408	1.20E-23
AC139491.2	0.405	7.24E-22
ITM2C	0.402	2.68E-23
IGFALS	0.402	1.25E-22
GALNT8	0.397	9.93E-23
CHGA	0.396	6.54E-22

ERN2	0.392	4.17E-22
NR3C2	0.386	1.48E-21
NXPE1	0.383	5.45E-21
GCNT3	0.382	4.17E-21
ALDH1L1	0.382	4.61E-21
MUC1	0.38	6.81E-21
RNA5SP75	0.377	0.0629
CA4	0.374	9.24E-20
CBFA2T3	0.372	3.94E-20
ST6GALNAC6	0.37	6.14E-20
CA1	0.37	5.91E-19
GUCA2B	0.368	5.04E-17
SNORD77	0.366	0.00803
Metazoa_SRP	-0.364	0.0227
NPY2R	0.362	2.68E-08
CLCA4	0.356	3.99E-18
AL133373.2	0.355	2.70E-16
BTNL8	0.354	2.27E-18
ST6GALNAC1	0.353	4.83E-18
B3GALT1	0.351	4.64E-17
POU2AF1	0.35	5.77E-18
TMIGD1	0.349	2.11E-14
VWA3B	0.343	6.84E-17
CES3	0.342	3.48E-17
MIR185	0.341	0.0816
BMP2	0.34	4.81E-17
MS4A12	0.337	9.39E-16
UGT2B17	0.337	8.25E-16
LYPD8	0.337	1.12E-16
PLA2G10	0.337	9.53E-17
MIR3165	0.335	0.00223
CDC42EP5	0.334	2.02E-16
DUOX2	0.334	1.82E-16
CA12	0.333	2.06E-16
AC005550.1	0.332	0.0668
CCL28	0.331	3.06E-16
BARX2	0.328	6.49E-16
HSPH1	-0.327	6.92E-16
IGHA1	0.327	6.79E-16
DMBT1	0.327	8.57E-16
FABP2	0.326	1.22E-15
RAE1	-0.326	8.71E-16
DUOXA2	0.323	2.42E-15
AKR1B10	0.321	3.55E-15
SIDT1	0.32	4.39E-15
Y_RNA	0.318	0.0251

C11orf86	0.318	1.60E-14
CAPN5	0.316	7.11E-15
RNVU1-18	0.314	0.212
AL133297.2	0.314	2.62E-12
GCG	0.314	1.02E-10
FFAR2	0.313	1.40E-14
NKX2-2	0.313	1.01E-09
MRAP2	0.313	1.39E-14
DUSP4	0.31	1.94E-14
MIR550A1	0.309	0.0914
IGLV5-45	0.309	3.00E-13
AC079328.2	0.309	3.13E-14
JCHAIN	0.307	3.38E-14
LINC00261	0.307	4.62E-14
MIR4322	0.306	0.065
FA2H	0.305	5.18E-14
SCARNA3	0.304	0.00863
DDX53	-0.303	0.0155
HYAL1	0.303	7.54E-14
MOGAT2	0.302	1.06E-13
PIGU	-0.302	9.82E-14
KIF19	0.301	1.54E-13
AKT1	0.3	1.53E-13
GNE	0.299	1.67E-13
PP14571	0.298	1.08E-11
VILL	0.298	2.26E-13
ABCC13	0.298	9.26E-13
AQP3	0.297	2.69E-13
AC009163.2	0.297	3.72E-09
SLITRK6	0.297	3.67E-13
TRIM7	0.294	4.54E-13
RNA5SP191	-0.294	0.157
NXPE2	0.293	1.09E-12
TRPA1	0.291	8.48E-13
L1TD1	0.29	1.14E-12
ABCB11	0.29	8.99E-10
AL133243.2	-0.29	1.37E-12
LGALS9B	0.289	2.48E-12
AC098934.2	-0.289	1.57E-12
LDHD	0.289	1.16E-12
HSPD1P21	0.288	0.175
AL662899.2	-0.287	1.47E-11
NUGGC	0.285	5.15E-12
AC002511.1	0.284	3.39E-10
RTFDC1	-0.283	2.95E-12
RHBDL3	0.283	5.81E-12

AGR3	0.282	3.65E-12
CACNA2D2	0.282	3.67E-12
DNASE1L3	0.281	5.87E-12
C2orf15	-0.281	4.50E-12
CDKN2B-AS	0.28	0.000548
NUFIP1	-0.28	4.77E-12
STYK1	0.28	5.38E-12
IGLV3-19	0.28	5.72E-12
LINC01913	0.28	3.97E-06
FAM30A	0.28	1.53E-11
STXBP5-AS1	0.278	7.95E-12
AC107953.1	0.278	0.267
MT1E	0.277	8.18E-12
RPL39	-0.277	8.68E-12
UGT1A8	0.277	8.04E-11
AC008870.5	0.277	1.27E-11
CEACAM7	0.277	9.27E-12
SLC30A10	0.276	3.32E-09
AL139246.3	0.276	1.31E-11
NANOGP7	0.276	1.93E-07
MS4A8	0.275	1.41E-11
AC009945.1	0.275	1.75E-05
AC008443.4	0.275	1.48E-06
VPS37B	0.274	1.38E-11
INSL5	0.273	7.22E-07
AC026402.1	0.272	0.026
CWH43	0.272	2.22E-10
KLK15	0.272	4.11E-11
COL4A6	0.272	5.35E-11
IGHV1-46	0.272	2.40E-11
PTPRH	0.272	2.82E-11
WDR92	-0.271	2.98E-11
PLLP	0.271	2.37E-11
IGHM	0.271	2.61E-11
RCC2P6	0.27	7.34E-11
PIPOX	-0.27	3.03E-11
IGLV4-69	0.27	5.23E-11
SNORD126	0.27	0.041
KRTAP10-1	0.269	0.0999
IGHV3-48	0.269	6.39E-11
DSCR8	-0.269	0.0143
AL117355.1	-0.268	0.00431
LY9	0.268	5.02E-11
SIDT2	0.268	4.53E-11
KIAA1324	0.266	6.74E-11
ADH1C	0.266	6.22E-11

SDCBP2	0.266	5.92E-11
ZCRB1	-0.266	6.06E-11
TBC1D2	0.265	7.82E-11
DUOXA1	0.265	9.26E-11
AC012236.1	0.264	9.45E-11
ANKRD36BP2	0.264	1.42E-10
POFUT1	-0.263	9.51E-11
PLA2G12B	-0.262	8.44E-10
IL1R2	0.262	1.27E-10
AL117382.1	-0.262	1.23E-10
TTLL6	0.262	1.70E-10
IGKV3-15	0.261	1.27E-10
IGHV3-73	0.261	2.43E-10
IGLV2-8	0.261	1.70E-10
OLFM4	0.261	1.39E-10
C16orf54	0.261	1.41E-10
IQGAP2	0.26	1.44E-10
MIR4507	0.26	0.0234
AC073288.1	0.26	0.0788
IGHV4-55	0.259	4.99E-10
NDRG3	-0.259	1.71E-10

Table A.11: List of the 200 top probesets most associated with the *CDH1* gene in the cell lines, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Phenotype	*r*	p-value
CDH1	0.942	6.83E-15
PKP3	0.819	3.16E-11
ZEB1	-0.819	3.16E-11
RAB25	0.819	3.16E-11
AP1M2	0.819	3.16E-11
---	-0.819	3.16E-11
AP1M2	0.819	3.16E-11
GALNT3	0.755	9.89E-10
IRF6	0.753	1.82E-09
EPCAM	0.753	1.82E-09
DSC2	0.753	1.82E-09
PKP3	0.753	1.82E-09
ST14	0.753	1.82E-09
F11R	0.753	1.82E-09
BSPRY	0.753	1.82E-09
TMEM125	0.753	1.82E-09
C1orf172	0.753	1.82E-09
---	0.753	1.82E-09
EPS8	0.702	1.45E-08
F11R	0.702	1.45E-08
PRSS8	0.687	2.19E-08
---	-0.687	2.19E-08
IRF6	0.685	4.93E-08
MAP7	0.685	4.93E-08
MAP7	0.685	4.93E-08
LAD1	0.685	4.93E-08
TMC4	0.685	4.93E-08
RAB39B	-0.685	4.93E-08
---	0.685	4.93E-08
KRT19	0.682	9.50E-08
SPINT1	0.682	9.50E-08
CNKSR1	0.682	9.50E-08
CDS1	0.682	9.50E-08
KRT8	0.682	9.50E-08
SYT11	-0.682	9.50E-08
F2RL1	0.682	9.50E-08
RAB17	0.682	9.50E-08
ESRP2	0.682	9.50E-08
MMP16	-0.682	9.50E-08
DSC2	0.682	9.50E-08
ESRP2	0.682	9.50E-08

MARVELD3	0.682	9.50E-08
ST14	0.656	1.25E-07
REEP2	-0.656	1.25E-07
S100A14	0.656	1.25E-07
OVOL1	0.656	1.25E-07
CBLC	0.652	1.15E-07
MACC1	0.652	1.15E-07
AASS	-0.632	2.01E-07
VIM	-0.629	5.89E-07
TJP2	0.629	5.89E-07
CLDN7	0.629	5.89E-07
MAP7	0.629	5.89E-07
DNAJB5	-0.629	5.89E-07
DENND2D	0.629	5.89E-07
ESRP1	0.629	5.89E-07
METRNL	0.629	5.89E-07
CDS1	0.629	5.89E-07
PCDHB7	-0.629	5.89E-07
RAB39B	-0.629	5.89E-07
DYNC2H1	-0.62	4.78E-07
ELMO3	0.616	7.32E-07
MSRB3	-0.616	7.32E-07
LOC100506844	-0.616	7.32E-07
---	-0.616	7.32E-07
MUC20	0.612	4.98E-07
VIM	-0.609	2.12E-06
MAPK13	0.609	2.12E-06
GPR56	0.609	2.12E-06
CNTNAP1	-0.609	2.12E-06
SYK	0.609	2.12E-06
CDS1	0.609	2.12E-06
CAMK4	-0.609	2.12E-06
JUP	0.606	4.50E-06
HLX	-0.606	4.50E-06
HOOK2	0.606	4.50E-06
MAL2	0.606	4.50E-06
FZD8	-0.606	4.50E-06
TMEM139	0.591	1.65E-06
KRT18	0.582	4.08E-06
ANK2	-0.582	4.08E-06
ESRP1	0.582	4.08E-06
LTBP2	-0.582	4.08E-06
SYNE1	-0.58	3.20E-06
APLP1	-0.58	3.20E-06
ZEB1	-0.58	3.20E-06
KRTCAP3	0.58	3.20E-06

MARVELD3	0.58	3.20E-06
FGFR4	0.565	4.14E-06
AKT3	-0.565	4.87E-06
---	0.565	4.87E-06
CDH3	0.558	4.91E-06
SV2A	-0.55	1.96E-05
PKP2	0.55	1.96E-05
MAPK13	0.55	1.96E-05
TRAF3IP2	0.55	1.96E-05
ADAP1	0.55	1.96E-05
FA2H	0.55	1.96E-05
SERINC2	0.55	1.96E-05
ADAP1	0.55	1.96E-05
TMEM30B	0.549	1.12E-05
PWWP2B	0.549	1.12E-05
TTLL7	-0.549	1.12E-05
ANKRD22	0.549	1.12E-05
CTSH	0.541	1.93E-05
PPARG	0.541	1.93E-05
FHL1	-0.541	1.93E-05
CTBP2	0.541	1.93E-05
BSPRY	0.541	1.93E-05
FBXO17 /// SARS2	-0.541	1.93E-05
RBM47	0.541	1.93E-05
---	-0.541	1.93E-05
TRPC1	-0.541	1.27E-05
ATP8B2	-0.541	1.27E-05
DNAJC18	-0.541	1.27E-05
TRIM31	0.527	1.72E-05
ERBB3	0.527	1.92E-05
---	-0.526	7.71E-05
KIAA0495	-0.526	7.71E-05
CLDN4	0.526	7.71E-05
SNAP25	-0.526	7.71E-05
RGS4	-0.526	7.71E-05
RGS4	-0.526	7.71E-05
MYO1D	0.526	7.71E-05
SERINC5	0.526	7.71E-05
FAM5C	-0.526	7.71E-05
TSPAN13	0.526	7.71E-05
EN1	-0.526	7.71E-05
STAP2	0.526	7.71E-05
SHROOM3	0.526	7.71E-05
RNF150	-0.526	7.71E-05
SHROOM3	0.526	7.71E-05
---	-0.526	7.71E-05

ARSB	-0.526	7.71E-05
MSRB3	-0.526	7.71E-05
---	-0.526	7.71E-05
NEGR1	-0.526	7.71E-05
KIAA0495	-0.526	7.71E-05
SIRPB1	-0.522	0.000195
LOC641365	-0.522	0.000195
---	-0.522	0.000195
VAMP8	0.522	0.000195
ITGB3	-0.522	0.000195
CACNA2D1	-0.522	0.000195
HTATIP2	0.522	0.000195
NRG2	-0.522	0.000195
LSR	0.522	0.000195
SMARCA4	0.522	0.000195
EFEMP2	-0.522	0.000195
PDLIM3	-0.522	0.000195
CHRNA3	-0.522	0.000195
SPINT2	0.522	0.000195
UGT2B28	-0.522	0.000195
CHRNA3	-0.522	0.000195
SMARCA4	0.522	0.000195
IGFBP7	-0.522	0.000195
SMARCA4	0.522	0.000195
SMARCA4	0.522	0.000195
CTAG1A /// CTAG1B	-0.522	0.000195
---	-0.522	0.000195
MARVELD2	0.522	0.000195
ENTPD7	-0.522	0.000195
LOC400464	-0.522	0.000195
LOC641365	-0.522	0.000195
---	-0.522	0.000195
FXYD3	0.52	3.29E-05
SYK	0.52	3.29E-05
RBM47	0.52	3.29E-05
GRHL2	0.52	3.29E-05
GLI3	-0.52	3.29E-05
OSBPL6	-0.52	3.29E-05
---	-0.52	3.29E-05
FAM55C	-0.518	3.00E-05
ANKRD22	0.518	3.00E-05
GPX2	0.512	3.01E-05
TMPRSS2	0.509	3.74E-05
LRRC8C	-0.509	3.74E-05
FHL1	-0.506	6.90E-05
ZNF225	-0.506	6.90E-05

CMTM3	-0.506	6.90E-05
ARHGAP8 /// PRR5-ARHGAP8	0.506	6.90E-05
MSRB3	-0.502	0.000105
MYO6	0.502	0.000105
ERCC1	-0.502	0.000105
AREG	0.502	0.000105
ANK3	0.502	0.000105
SYT11	-0.502	0.000105
MYO5C	0.502	0.000105
C1orf106	0.502	0.000105
EHF	0.502	0.000105
AK5	-0.502	0.000105
MEX3B	-0.502	0.000105
EHF	0.502	0.000105
EHF	0.502	0.000105
ANKRD50	-0.502	0.000105
FAM89A	0.502	0.000105
SHISA2	-0.502	0.000105
EHF	0.502	0.000105
---	-0.502	0.000105
TJP3	0.502	0.000105

Table A.12: List of the 200 top probesets most associated with the *ZEB1* gene in the cell lines, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Phenotype	*r*	p-value
RAB25	-1	3.99E-16
AP1M2	-1	3.99E-16
AP1M2	-1	3.99E-16
GALNT3	-0.929	4.35E-14
IRF6	-0.919	1.63E-13
EPCAM	-0.919	1.63E-13
DSC2	-0.919	1.63E-13
PKP3	-0.919	1.63E-13
ST14	-0.919	1.63E-13
F11R	-0.919	1.63E-13
BSPRY	-0.919	1.63E-13
TMEM125	-0.919	1.63E-13
C1orf172	-0.919	1.63E-13
---	-0.919	1.63E-13
CDH1	-0.869	1.68E-12
F11R	-0.869	1.68E-12
IRF6	-0.843	1.45E-11
MAP7	-0.843	1.45E-11
PKP3	-0.843	1.45E-11
MAP7	-0.843	1.45E-11
LAD1	-0.843	1.45E-11
TMC4	-0.843	1.45E-11
---	-0.843	1.45E-11
---	0.843	1.45E-11
KRT19	-0.833	5.66E-11
SPINT1	-0.833	5.66E-11
CNKSR1	-0.833	5.66E-11
CDS1	-0.833	5.66E-11
KRT8	-0.833	5.66E-11
SYT11	0.833	5.66E-11
F2RL1	-0.833	5.66E-11
RAB17	-0.833	5.66E-11
ESRP2	-0.833	5.66E-11
MMP16	0.833	5.66E-11
DSC2	-0.833	5.66E-11
ESRP2	-0.833	5.66E-11
MARVELD3	-0.833	5.66E-11
CDH1	-0.819	3.16E-11
ST14	-0.819	3.16E-11
REEP2	0.819	3.16E-11
S100A14	-0.819	3.16E-11

OVOL1	-0.819	3.16E-11
VIM	0.781	4.26E-10
TJP2	-0.781	4.26E-10
CLDN7	-0.781	4.26E-10
MAP7	-0.781	4.26E-10
DENND2D	-0.781	4.26E-10
ESRP1	-0.781	4.26E-10
CDS1	-0.781	4.26E-10
ELMO3	-0.775	3.50E-10
LOC100506844	0.775	3.50E-10
VIM	0.751	3.87E-09
MAPK13	-0.751	3.87E-09
GPR56	-0.751	3.87E-09
CNTNAP1	0.751	3.87E-09
SYK	-0.751	3.87E-09
CDS1	-0.751	3.87E-09
JUP	-0.74	1.69E-08
HOOK2	-0.74	1.69E-08
MAL2	-0.74	1.69E-08
FZD8	0.74	1.69E-08
PRSS8	-0.736	2.61E-09
ZEB1	0.736	2.61E-09
KRTCAP3	-0.736	2.61E-09
MARVELD3	-0.736	2.61E-09
KRT18	-0.729	5.94E-09
EPS8	-0.729	5.94E-09
ESRP1	-0.729	5.94E-09
TMEM30B	-0.702	1.43E-08
CBLC	-0.702	1.43E-08
MACC1	-0.702	1.43E-08
ANKRD22	-0.702	1.43E-08
SV2A	0.686	8.02E-08
PKP2	-0.686	8.02E-08
MAPK13	-0.686	8.02E-08
TRAF3IP2	-0.686	8.02E-08
FA2H	-0.686	8.02E-08
SERINC2	-0.686	8.02E-08
RAB39B	0.686	8.02E-08
PPARG	-0.685	4.93E-08
BSPRY	-0.685	4.93E-08
FBXO17 /// SARS2	0.685	4.93E-08
FXYD3	-0.671	6.20E-08
GRHL2	-0.671	6.20E-08
KIAA0495	0.65	8.01E-07
CLDN4	-0.65	8.01E-07
SNAP25	0.65	8.01E-07

RGS4	0.65	8.01E-07
RGS4	0.65	8.01E-07
MYO1D	-0.65	8.01E-07
SERINC5	-0.65	8.01E-07
TSPAN13	-0.65	8.01E-07
SHROOM3	-0.65	8.01E-07
RNF150	0.65	8.01E-07
SHROOM3	-0.65	8.01E-07
---	0.65	8.01E-07
MSRB3	0.65	8.01E-07
---	0.65	8.01E-07
KIAA0495	0.65	8.01E-07
CMTM3	0.646	2.80E-07
MSRB3	0.646	2.80E-07
---	0.646	2.80E-07
ARHGAP8 /// PRR5-ARHGAP8	-0.646	2.80E-07
SIRPB1	0.637	4.32E-06
---	0.637	4.32E-06
VAMP8	-0.637	4.32E-06
ITGB3	0.637	4.32E-06
CACNA2D1	0.637	4.32E-06
HTATIP2	-0.637	4.32E-06
NRG2	0.637	4.32E-06
LSR	-0.637	4.32E-06
SMARCA4	-0.637	4.32E-06
PDLIM3	0.637	4.32E-06
SPINT2	-0.637	4.32E-06
SMARCA4	-0.637	4.32E-06
IGFBP7	0.637	4.32E-06
SMARCA4	-0.637	4.32E-06
SMARCA4	-0.637	4.32E-06
---	0.637	4.32E-06
MARVELD2	-0.637	4.32E-06
LOC400464	0.637	4.32E-06
---	0.637	4.32E-06
MSRB3	0.633	7.87E-07
ERCC1	0.633	7.87E-07
DNAJB5	0.633	7.87E-07
SYT11	0.633	7.87E-07
MYO5C	-0.633	7.87E-07
C1orf106	-0.633	7.87E-07
EHF	-0.633	7.87E-07
AK5	0.633	7.87E-07
MEX3B	0.633	7.87E-07
EHF	-0.633	7.87E-07
EHF	-0.633	7.87E-07

ANKRD50	0.633	7.87E-07
METRNL	-0.633	7.87E-07
EHF	-0.633	7.87E-07
RAB39B	0.633	7.87E-07
SERPINB1	-0.618	6.75E-07
AKT3	0.618	6.75E-07
AKT3	0.618	6.75E-07
DTX4	-0.618	6.75E-07
AKT3	0.618	6.75E-07
RECK	0.612	1.19E-06
CLDN11	0.612	1.19E-06
SYNE1	0.612	1.19E-06
APLP1	0.612	1.19E-06
FSD1	0.612	1.19E-06
---	0.612	1.19E-06
ATP8B2	0.595	1.81E-06
---	-0.595	1.81E-06
ANK2	0.589	4.68E-06
ERCC1	0.589	4.68E-06
RGS4	0.589	4.68E-06
ARHGAP8 /// PRR5-ARHGAP8	-0.589	4.68E-06
COBL	-0.589	4.68E-06
SALL2	0.589	4.68E-06
KIAA0495	0.589	4.68E-06
LTBP2	0.589	4.68E-06
SIRPA	0.583	1.07E-05
PGCP	0.583	1.07E-05
KCNK1	-0.583	1.07E-05
KCNK1	-0.583	1.07E-05
TGFA	-0.583	1.07E-05
GFPT2	0.583	1.07E-05
MST1R	-0.583	1.07E-05
CD70	0.583	1.07E-05
SEL1L3	-0.583	1.07E-05
SEL1L3	-0.583	1.07E-05
DKK3	0.583	1.07E-05
ANGPTL4	0.583	1.07E-05
CAMK4	0.583	1.07E-05
C3orf55	0.583	1.07E-05
AKT3	0.582	4.10E-06
PWWP2B	-0.582	4.10E-06
TTLL7	0.582	4.10E-06
FAM55C	0.573	4.37E-06
ANKRD22	-0.573	4.37E-06
CMTM3	0.554	1.18E-05
MAPK12	0.554	1.18E-05

S100P	-0.554	1.18E-05
RBM47	-0.554	1.18E-05
DYNC2H1	0.554	1.18E-05
SH2D3A	-0.554	1.18E-05
GLI3	0.554	1.18E-05
RASSF8	0.554	1.18E-05
OSBPL6	0.554	1.18E-05
ACSL5	-0.553	9.63E-06
ACSL5	-0.553	9.63E-06
PRR15	-0.553	9.63E-06
---	0.553	9.63E-06
TTC22	-0.553	9.63E-06
ST3GAL3	0.55	1.96E-05
SERPINE1	0.55	1.96E-05
ANK2	0.55	1.96E-05
FHL1	0.55	1.96E-05
CTBP2	-0.55	1.96E-05
RBM47	-0.55	1.96E-05
LRRN4	0.537	0.00012
MSRB3	0.537	0.00012
SNAP25	0.537	0.00012

Table A.13: List of the 200 top genes most associated with the *CDH1* gene in the TCGA primary tumors, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Transcript	*r*	p-value
Metazoa_SRP	-0.806	2.00E-04
MIR548AY	-0.684	0.00494
MIR550A1	-0.64	0.000223
AC009446.1	-0.611	0.00255
EPCAM	0.559	2.30E-43
Y_RNA	-0.545	0.00099
AP003461.1	-0.543	0.00864
AL390726.1	-0.507	0.0144
HMGN2P32	-0.505	2.03E-14
RNA5SP152	-0.498	0.000208
LINC02210-CRHR1	-0.494	0.00353
DOPEY1	0.494	8.91E-34
USP7	0.49	1.04E-33
TEAD1	0.474	2.13E-31
ZNF91	0.468	1.26E-30
RNA5SP295	-0.467	0.0311
KDM7A	0.465	2.47E-30
ROCK2	0.46	6.76E-30
AC012146.4	-0.455	0.0594
VDAC1	0.453	7.74E-29
SH3D19	0.453	7.93E-29
SH3BGRL2	0.451	2.44E-28
MIR4322	-0.449	0.00589
DYNC1LI2	0.447	3.52E-28
CEACAM5	0.445	6.93E-28
RNU4-40P	-0.444	0.0126
CHD9	0.443	1.83E-27
ACSL5	0.443	9.15E-28
TM9SF3	0.44	2.88E-27
RNA5SP297	-0.44	0.0129
ATP13A3	0.44	4.46E-27
TNFRSF18	-0.44	2.43E-27
AL139317.1	-0.433	0.00574
FNBP1L	0.432	5.17E-26
Y_RNA	-0.431	0.0971
AP1G1	0.431	5.06E-26
AP001922.4	-0.43	0.00121
RPS29P9	-0.429	1.66E-05
MAP7	0.429	8.88E-26
RBBP6	0.428	6.07E-26
NFAT5	0.427	7.28E-26

PHOSPHO2	0.427	1.32E-25
MIR4285	-0.427	0.0317
WWP1	0.427	1.38E-25
FAM135A	0.424	3.39E-25
AKAP9	0.424	3.39E-25
PAPD5	0.424	1.93E-25
ADD3	0.424	1.93E-25
CMTR2	0.424	4.40E-25
BAZ2B	0.424	3.34E-25
PRLR	0.423	2.27E-25
CD46	0.423	3.27E-25
SPRR2C	-0.423	0.0409
RNU4-36P	-0.422	0.00874
RNA5SP354	-0.422	0.00413
ZNF664	0.422	8.00E-25
CRCT1	-0.422	0.000311
AL391903.3	-0.421	0.0103
RN7SL664P	-0.42	2.51E-07
RNU6-3P	-0.42	0.000288
VMO1	-0.419	4.75E-25
TGOLN2	0.419	4.83E-25
TTC14	0.419	1.10E-24
NFE2L2	0.419	1.63E-24
UBN2	0.419	7.91E-25
ESRP1	0.418	9.27E-25
TRMT1L	0.418	3.08E-24
FNIP1	0.416	3.35E-24
BRWD3	0.416	2.24E-24
B9D2	-0.414	2.22E-24
HOOK1	0.414	4.67E-24
ATG101	-0.414	4.35E-24
MRPL55	-0.414	6.56E-24
RBAK	0.413	4.69E-24
RBM12	0.413	4.69E-24
KIF5B	0.411	4.05E-24
ZNF638	0.411	8.08E-24
RBM27	0.411	1.09E-23
SRPK1	0.411	8.43E-24
EEA1	0.41	9.96E-24
PIKFYVE	0.409	1.09E-23
POF1B	0.408	2.30E-23
PHF6	0.408	3.39E-23
FKTN	0.408	2.13E-23
CXorf23	0.408	2.13E-23
DLG1	0.407	2.65E-23
PUM2	0.407	4.98E-23

RRN3	0.406	2.51E-23
RPS12P5	-0.406	0.00348
SENP6	0.406	2.12E-23
SEC63	0.406	1.81E-23
ERI2	0.406	4.80E-23
ZNF518A	0.405	3.58E-23
GSPT1	0.404	6.26E-23
FAM212A	-0.404	3.99E-23
ZNF585A	0.403	5.39E-23
CNOT11	0.403	9.57E-23
TRNAU1AP	-0.403	4.02E-23
AC109479.3	-0.402	8.17E-06
SOS1	0.402	7.40E-23
STRN	0.401	1.39E-22
GOPC	0.4	1.13E-22
SNORA80A	-0.399	5.28E-06
NIPBL	0.399	1.02E-22
FASTKD1	0.399	3.19E-22
RANBP2	0.399	3.19E-22
STT3B	0.399	3.19E-22
ATP9A	0.398	9.52E-23
MED13L	0.398	2.60E-22
IPMK	0.398	2.60E-22
JRKL	0.398	2.60E-22
ZNF181	0.397	3.42E-22
ATRX	0.397	2.74E-22
LYL1	-0.397	1.90E-22
SF3B1	0.396	7.89E-22
SRFBP1	0.396	7.89E-22
ICE1	0.395	2.89E-22
ZNF112	0.395	3.59E-22
MIR548N	-0.395	0.00395
MBTD1	0.394	6.78E-22
NR1D2	0.394	6.78E-22
MON2	0.394	6.15E-22
PREPL	0.394	8.17E-22
HELZ	0.393	5.06E-22
CTNND1	0.393	3.79E-22
DDX3X	0.392	8.06E-22
PHACTR2	0.391	8.20E-22
CHCHD5	-0.391	1.03E-21
PRKAA1	0.39	1.90E-21
MACC1	0.39	1.73E-21
ITCH	0.39	2.62E-21
SBF2	0.39	2.62E-21
TMEM170A	0.39	2.62E-21

POLR2L	-0.39	1.74E-21
LTN1	0.389	1.24E-21
CCDC122	0.389	1.72E-21
STRBP	0.388	1.69E-21
SMN1	0.388	2.68E-21
ZNF675	0.388	1.73E-21
SCYL2	0.388	4.74E-21
TOP1	0.388	1.75E-21
ETS2	0.387	1.88E-21
CEP295	0.387	1.88E-21
DNAJC2	0.387	4.25E-21
SNX4	0.387	4.25E-21
NOSTRIN	0.387	2.89E-21
RBM41	0.387	2.55E-21
REV3L	0.387	4.29E-21
BATF3	-0.387	4.29E-21
TMEM161B	0.387	4.29E-21
RSPRY1	0.386	2.96E-21
AL353691.2	-0.386	1.46E-08
ZNF146	0.386	3.48E-21
ZNF292	0.386	3.48E-21
KTN1	0.386	7.96E-21
UFL1	0.385	3.29E-21
CTCF	0.385	1.73E-20
TSN	0.384	1.23E-20
EAF1	0.384	9.18E-21
HERC4	0.384	9.18E-21
ZNF445	0.384	4.26E-21
RNA5SP194	-0.384	0.0177
TRIP12	0.384	6.74E-21
FANCF	0.384	6.74E-21
SCARF1	-0.383	2.11E-20
RBBP5	0.383	1.70E-20
TXNDC9	0.383	1.04E-20
NINJ1	-0.383	4.23E-21
RC3H1	0.383	8.52E-21
NAALADL2	0.383	4.59E-21
UBR3	0.382	9.90E-21
VTRNA1-3	-0.382	2.98E-05
PUS7L	0.381	1.07E-20
NDUFS7	-0.381	1.32E-20
GLS	0.381	1.32E-20
CWC22	0.381	1.32E-20
SPAST	0.381	1.89E-20
PIK3CB	0.381	1.89E-20
PPP4R3B	0.381	1.89E-20

ZNF253	0.381	1.04E-20
NF1	0.381	1.09E-20
MMP25-AS1	-0.381	1.09E-20
ZNF493	0.381	8.11E-21
ANKRD26	0.38	7.29E-21
AC093227.1	0.38	1.14E-20
ACAP2	0.38	2.08E-20
MAP3K2	0.38	2.08E-20
TAOK1	0.38	2.44E-20
SMARCAD1	0.38	2.44E-20
ARHGAP42	0.38	1.91E-20
ZNF107	0.38	1.91E-20
PAXBP1	0.379	1.25E-20
RAB10	0.379	3.72E-20
BZW2	0.379	3.72E-20
PEX13	0.379	3.72E-20
ZNF852	0.378	1.46E-20
UPF2	0.378	1.71E-20
KIAA1109	0.378	1.56E-20
JMJD1C	0.377	3.56E-20
ARMCX5-GPRASP2	0.377	3.56E-20

Table A.14: List of the 200 top genes most associated with the *NKX2-3* gene in the TCGA primary tumors, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Transcript	*r*	p-value
FENDRR	0.534	6.20E-39
FOXF1	0.503	8.59E-35
F2RL2	0.491	5.59E-33
FOXF2	0.48	1.60E-31
TCF21	0.478	2.02E-31
SALL1	0.466	8.00E-30
LINC00686	-0.456	0.0707
EPHA7	0.423	1.60E-24
LAMC3	0.42	2.13E-23
GFRA1	0.419	8.17E-24
ITIH5	0.414	4.91E-24
HPSE2	0.409	6.63E-23
OR7E47P	0.399	9.13E-22
ADAMTS8	0.398	2.99E-22
LRRN4CL	0.388	6.11E-21
CH25H	0.382	1.62E-20
KIAA1755	0.381	2.40E-20
CXCL14	0.373	1.43E-19
MIR549A	-0.372	0.12
SLC35F1	0.371	2.92E-19
AC097639.2	-0.367	0.368
ROBO2	0.361	3.99E-18
MIR486-1	-0.36	0.0294
EMID1	0.353	8.89E-18
GLP2R	0.351	7.27E-17
LRRC4B	0.344	1.69E-16
COL9A1	0.341	2.33E-16
ISLR2	0.332	1.05E-15
MIR4499	0.32	0.09
MYOCD	0.32	5.86E-15
AL080285.2	0.319	0.0253
PDGFRA	0.318	2.87E-14
MIR3162	-0.317	0.174
Y_RNA	0.316	0.0875
FAM163A	0.316	5.66E-14
EDNRB	0.316	4.31E-14
MTND6P22	-0.315	2.69E-06
GREM2	0.314	3.00E-14
PLEKHH2	0.314	3.44E-14
SCUBE1	0.312	6.56E-14
PCDH18	0.311	3.28E-14

CCL11	0.31	6.16E-14
AC093001.1	-0.304	0.000391
STMN2	0.303	3.90E-13
ADAM33	0.301	6.71E-13
CNNM1	0.301	4.87E-13
VIPR2	0.3	7.17E-13
NEGR1	0.297	3.59E-12
AC110814.1	0.296	5.50E-13
RNU5B-3P	-0.296	0.0621
SERTAD4-AS1	0.294	1.54E-12
EFCC1	0.292	1.65E-12
LINC01082	0.288	1.33E-11
COL4A5	0.288	3.19E-12
WBSCR17	0.287	9.38E-12
FRZB	0.286	3.54E-12
LINC00452	-0.286	0.00408
C1QTNF2	0.285	6.24E-12
WFDC1	0.281	7.94E-12
LCE3D	-0.279	0.531
CD1C	0.279	1.43E-11
AC010641.1	-0.279	0.401
TNFSF11	0.275	3.18E-11
INSC	0.275	4.46E-11
ZNF734P	-0.275	0.571
AMPH	0.272	5.40E-11
AC100803.2	0.27	6.73E-11
CHL1	0.269	1.20E-10
DNASE1L3	0.269	8.92E-11
BMP5	0.269	3.72E-10
TYRP1	0.269	3.39E-10
SLC24A3	0.268	1.17E-10
PLD4	0.268	8.73E-11
CD1A	0.268	2.51E-10
CLEC2A	-0.267	0.0768
CHRNA1	0.267	3.41E-10
IGKV3-25	-0.267	0.219
SKOR2	0.267	0.00631
RGS7BP	0.267	5.67E-10
WNT4	0.266	1.12E-10
SCARA5	0.265	1.97E-10
AC007881.1	-0.264	0.265
DERL3	0.264	5.12E-10
MAB21L2	0.264	5.12E-09
AC040170.1	0.263	5.62E-10
LSAMP	0.262	3.11E-10
AFF3	0.262	3.05E-10

AGT	0.262	1.17E-09
ANKRD35	0.262	4.42E-10
SRC	0.261	7.77E-09
LINC01915	0.26	9.81E-10
PLAT	0.259	3.36E-10
LUZP2	0.259	1.46E-09
AL645608.1	0.258	5.42E-10
GDNF	0.257	1.95E-09
AP001790.1	-0.257	0.0106
PDPN	0.255	7.85E-10
CCDC3	0.253	1.03E-09
LDLRAD4	0.252	1.61E-09
MAPK10	0.252	3.45E-09
LINC00973	-0.251	0.000832
WNT5A	0.249	1.73E-09
SMIM24	0.248	3.25E-09
LINC02470	-0.247	0.0453
AP003071.4	0.244	9.91E-09
IL1RL1	0.244	6.23E-09
SLC26A3	0.244	3.22E-09
SCARNA21	-0.244	0.405
RNU6-25P	0.243	0.108
RSPO2	0.242	3.49E-08
NRG1	0.241	6.90E-09
INA	0.24	4.38E-08
MASP1	0.239	1.24E-08
AC092691.1	-0.238	0.348
AC103996.1	-0.236	0.23
CPED1	0.235	1.16E-08
PRSS8	0.235	6.14E-08
SLC27A6	0.235	7.02E-06
RNU6-1255P	-0.233	0.352
IGHA2	0.231	2.12E-08
RNA5SP92	0.231	0.31
COL4A6	0.23	5.22E-08
SNORA31	-0.23	0.44
C15orf59	0.23	2.95E-08
SEMA5A	0.23	3.96E-08
STOX2	0.229	3.22E-08
MGLL	0.229	9.67E-08
TRPC4	0.228	8.52E-08
SHOX2	-0.228	9.20E-08
ACKR1	0.227	5.47E-08
GLRB	0.227	1.20E-07
COL21A1	0.225	1.19E-07
IHH	0.225	9.25E-08

TMSB15A	0.224	1.12E-07
VIP	0.223	9.71E-08
MIR3972	-0.223	0.209
BHLHE22	0.221	1.04E-07
TNN	0.219	2.16E-07
IGHD1-26	-0.218	0.389
CCL22	0.218	1.54E-07
ACVRL1	0.217	1.72E-07
RNU6-36P	0.216	0.143
AC091286.1	0.215	0.0476
AC120498.1	-0.214	0.527
MIR205HG	-0.213	0.0471
CD207	0.213	6.07E-07
SNORD116-14	-0.212	0.183
SUMO2P17	-0.212	1.00E-06
TNFSF18	0.211	9.28E-07
UNC5C	0.21	7.54E-07
SNORA51	0.21	0.00059
RNA5SP297	0.209	0.305
ADAMDEC1	0.208	5.06E-07
ARHGAP20	0.208	1.01E-06
RPS26P4	0.208	0.0834
SHISA2	0.208	5.32E-07
CD1E	0.207	1.04E-06
AC105460.2	-0.207	0.0167
CEBPA	0.206	8.15E-07
AL034346.1	0.206	1.74E-05
HDC	0.205	1.04E-06
IGFL3	-0.205	8.18E-05
CHODL	0.205	2.55E-06
AC125603.1	-0.204	0.00238
RNA5SP277	0.204	0.455
AC105460.1	-0.204	0.000715
GYG2	0.202	1.35E-06
HHLA2	0.202	1.35E-06
NEMP2	-0.202	1.35E-06
MEOX1	0.202	1.19E-06
HHIP	0.202	1.21E-06
SORT1	0.202	2.26E-06
MIR6825	0.202	0.298
JCHAIN	0.201	1.17E-06
CNGA3	0.201	2.08E-05
FEZF2	0.201	0.00787
NLGN4X	0.201	2.80E-06
AXIN1	0.201	0.00026
CACFD1	0.201	0.00026

FAM83E	0.2	6.02E-06
snoU13	0.199	0.607
H3F3AP6	-0.199	4.79E-06
MIR3667	0.199	0.804
AP000439.2	0.199	2.18E-06
DKK3	0.199	1.87E-06
AC103563.4	-0.199	0.317
ESPN	0.199	2.73E-06
LINC02315	-0.198	0.0297
SLC44A4	0.198	5.54E-06
SHANK2	0.198	5.23E-06
ZNF536	0.197	6.48E-06
ADARB2	0.197	4.64E-06
CACNA1D	0.197	4.25E-06
C3orf36	0.196	3.65E-06
snoU13	-0.196	0.262
Y_RNA	-0.196	0.0675
ALDH2	0.196	1.21E-05
DAB2IP	0.196	1.21E-05
FAM72D	-0.195	3.67E-06
AC144831.1	0.195	1.21E-05

Table A.15: List of the 200 top genes most associated with the *AOC3* gene in the TCGA primary tumors, sorted by the absolute value of the association coefficient *r* between them. P-value from χ^2 test with Yates's correction for continuity is shown.

Gene or Transcript	*r*	p-value
LMOD1	0.821	3.46E-92
DNAJB5	0.748	7.90E-77
HSPB8	0.72	3.55E-71
CNN1	0.715	2.20E-70
TNS1	0.708	4.75E-69
SYNPO2	0.706	1.86E-68
KCNMB1	0.704	3.00E-68
MYLK	0.693	5.00E-66
POPDC2	0.689	2.06E-65
KANK2	0.689	2.52E-65
JPH2	0.688	3.00E-65
AP000892.3	0.684	1.87E-64
SLIT3	0.674	1.68E-62
RBPMS2	0.673	2.33E-62
MYH11	0.669	1.09E-61
JAM3	0.666	5.35E-61
CRYAB	0.659	9.19E-60
ACTG2	0.657	1.61E-59
SPARCL1	0.656	2.27E-59
LHFP	0.654	6.16E-59
MGP	0.637	3.25E-56
CILP	0.633	3.19E-55
BVES	0.633	2.86E-55
FXYD6	0.631	3.47E-55
LINC00702	0.631	1.29E-53
FHL1	0.614	6.45E-52
LDB3	0.596	7.03E-49
PKD2	0.596	3.04E-49
OSMR	0.589	4.29E-48
ATP1A2	0.587	3.60E-46
PLIN4	0.587	9.95E-48
C7	0.583	4.81E-47
FILIP1	0.582	5.10E-47
CASQ2	0.58	1.03E-42
THBS4	0.575	5.88E-46
MMRN2	0.573	1.45E-45
IL1R1	0.571	2.53E-45
COL14A1	0.568	9.56E-45
SYNM	0.561	8.89E-44
PSD	0.56	1.22E-43
PDE1B	0.56	1.22E-43

CAVIN2	0.556	4.63E-43
KCNK3	0.554	4.84E-41
CLDN11	0.546	2.10E-41
CSRP1	0.543	7.12E-41
DES	0.536	3.07E-40
PYGM	0.536	1.35E-39
HTR2B	0.535	3.77E-37
BCHE	0.535	1.64E-36
CXCL12	0.533	9.28E-40
C1QTNF7	0.533	4.82E-39
ATP2B4	0.533	1.64E-39
ITGA7	0.522	3.95E-38
SNAP25	0.52	4.35E-35
MBNL1-AS1	0.512	1.44E-36
MYOM1	0.512	9.65E-37
EBF1	0.51	1.77E-36
C16orf45	0.51	1.79E-36
CD93	0.509	2.74E-36
PPP1R12B	0.499	7.18E-35
NPR1	0.498	9.14E-35
SVIL	0.494	3.34E-34
LIMS2	0.493	3.02E-34
CYP1B1	0.493	4.77E-34
DIXDC1	0.49	1.09E-33
AC005180.1	0.489	1.00E-20
CHST10	0.484	5.07E-33
SPOCK1	0.482	9.21E-33
ADAMTS1	0.481	1.42E-32
CDH5	0.478	5.28E-32
CFL2	0.475	1.46E-31
COL16A1	0.468	5.72E-31
GPX3	0.467	8.64E-31
AQP1	0.467	8.67E-31
SRPX	0.466	9.74E-31
PECAM1	0.465	1.70E-30
C10orf10	0.465	1.25E-30
SCN1B	0.463	3.05E-30
ZCCHC24	0.458	1.05E-29
FBXO32	0.456	1.56E-29
FABP4	0.453	6.32E-28
AC007881.1	-0.452	0.00962
PDE2A	0.451	1.96E-28
CTGF	0.448	1.69E-28
FBXL22	0.448	3.37E-28
CD34	0.447	2.66E-28
AC036108.3	0.444	2.33E-26

OR51E2	0.44	3.38E-27
ISLR	0.439	2.21E-27
PDGFRL	0.438	3.15E-27
ARHGAP1	0.437	4.22E-27
ENTPD1	0.435	1.11E-26
SH2D3C	0.434	1.08E-26
KIF1A	0.43	3.11E-25
NKX3-2	0.428	6.68E-26
FMO2	0.427	6.76E-24
ABCA9	0.423	9.84E-25
AC106786.1	0.421	6.46E-22
ELK3	0.416	1.44E-24
SFRP4	0.415	1.08E-24
CAV1	0.412	4.60E-24
CRIP2	0.408	7.99E-24
TUSC5	0.404	5.88E-12
MIR6870	-0.402	0.00139
BEX4	0.401	4.43E-23
MIR3926-1	-0.401	0.00475
NTM	0.4	4.91E-23
PI16	0.4	9.96E-22
RNA5SP92	-0.399	0.0126
CSGALNACT2	0.398	1.01E-22
SFRP2	0.398	9.87E-23
NOVA2	0.397	1.19E-22
ADIPOQ	0.396	4.14E-09
SH3PXD2B	0.396	1.67E-22
CXorf36	0.391	1.29E-21
Y_RNA	-0.389	0.0147
DLG4	0.387	1.64E-21
UST	0.386	2.12E-21
SYNGR1	0.386	2.10E-21
OXSM	-0.385	1.99E-21
MRRF	-0.383	3.39E-21
MIR646	-0.383	0.0601
CYR61	0.381	4.88E-21
PRG4	0.381	2.83E-20
KDR	0.379	1.01E-20
C15orf59	0.378	1.44E-20
COL10A1	0.375	2.09E-20
TCF7L1	0.375	4.51E-20
TBX1	0.374	9.16E-20
GUCY1B3	0.371	5.96E-20
FAM180A	0.37	2.87E-19
C1orf54	0.37	7.40E-20
NCAM2	0.368	3.45E-19

PNMA1	0.367	1.61E-19
PCDHB3	0.367	3.89E-19
C1QTNF3	0.364	3.14E-19
MIR3162	-0.364	0.074
AC104689.1	-0.363	0.00273
ESAM	0.363	4.85E-19
CLEC3B	0.363	4.97E-19
SNCG	0.362	4.98E-19
AL050335.1	-0.362	0.0831
THRSP	0.361	8.06E-14
PHLDA3	0.36	6.10E-19
TACR2	0.359	8.49E-19
NR2F2	0.359	1.14E-18
SERPINE1	0.359	8.43E-19
RNU6-534P	-0.359	0.000123
SFXN4	-0.358	1.18E-18
AC053503.4	0.358	6.17E-13
MIR6855	-0.357	0.00116
AC093627.4	0.357	2.07E-18
RNU6-101P	-0.356	0.00336
LINC02265	-0.354	0.00013
PLIN1	0.353	1.14E-17
COG2	-0.348	1.09E-17
CASP5	-0.347	1.15E-17
RNU1-75P	-0.343	0.0812
BMPR1B	0.342	9.25E-16
RNU6-964P	-0.342	0.00445
SCN3B	0.339	2.37E-16
AL356320.1	-0.339	0.000125
AC092435.1	-0.339	0.0274
EDNRA	0.339	8.46E-17
PCDHGC3	0.338	8.81E-17
PCDHB4	0.337	1.42E-16
RTL8C	0.336	1.27E-16
FZD4	0.336	3.87E-16
MRPL37	-0.335	1.38E-16
ANOS1	0.334	1.72E-16
MIR4514	-0.334	0.0136
ZBTB46	0.332	3.25E-16
MAD2L1	-0.332	3.22E-16
FAM19A5	0.332	3.13E-16
SHQ1	-0.332	3.49E-16
PLPP4	0.331	3.92E-16
KIF17	0.329	5.99E-16
LINC00955	-0.329	7.76E-09
FNDC5	0.329	7.18E-16

RAD18	-0.328	6.54E-16
PCAT19	0.328	6.59E-16
PTPRN	0.326	1.04E-15
IL18BP	0.326	1.04E-15
ZNF418	0.325	4.75E-15
SIRPA	0.324	1.39E-15
MIR5192	-0.324	0.00227
SAMD4A	0.323	1.88E-15
OLFM1	0.322	3.14E-15
AC000111.2	-0.321	0.00123
snoU13	-0.32	0.0186
ATP5C1	-0.32	3.98E-15
LINC00987	0.32	4.69E-15
ZDHHC23	-0.319	4.58E-15
CORIN	0.319	4.11E-15
TM6SF1	0.319	6.32E-15
CRCT1	0.318	0.00808
MAPKAPK5	-0.317	6.13E-15
GJC1	0.317	1.49E-14
UMPS	-0.317	7.87E-15
LINC02022	-0.317	0.0771