GOBLET CELL DIFFERENTIATION IN HUMAN COLORECTAL CANCER CELL LINES

Haoyu Liu

Kellogg College

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

Trinity Term 2017

Weatherall Institute of Molecular Medicine

University of Oxford

DEDICATION

To the loving memory of my father.

Dad, not a day goes by I don't miss you in the past ten years. I wish you were here with me and would be proud to know the completion of my PhD in University of Oxford.

To the continuous support from my mother, a respectable lady who has supported the whole family by herself since the death of my father. Thank you for all your unconditional love.

ABSTRACT

Goblet cells are one of the three fully differentiated lineages in human colonic crypts. They play an important role in protecting epithelial cells from direct contact with luminal contents by secreting mucus, which consists of MUC2, TFF3 and other constituents. The goblet cell differentiation, however, is largely dysregulated in human colorectal cancers. Abundant expression of MUC2 is often observed in signet ring carcinomas and mucinous carcinomas that are associated with unsatisfactory prognosis. Despite the significant roles of goblet cells, its genetic nature and exact regulatory mechanisms remain incompletely understood.

In this thesis, the goblet cell differentiation at mRNA and protein levels was characterised in a panel of 64 human colorectal cancer cell lines. Microarray analysis on the bulk population of these cell lines reveals the genes differentially expressed in goblet cell-positive cell lines, including AKR1B10, AGR3 and the cellular surface protein CA12 (**Chapter 3**). In addition, a novel protocol is developed for isolation and sequencing of RNA from the fixed and FACS purified cells. Using this protocol, the first goblet cell transcriptome is profiled in LS180 cancer cell line, and the goblet cell-specific genes are identified, including TFF3 and SPDEF (**Chapter 4**). The co-expression patterns of TFF3 and MUC2 is further investigated. In a subset of cancer cell lines, TFF3 recognises the goblet cells that cannot produce MUC2, suggesting additional regulatory mechanisms are required for its expression. The dysregulated regulation of TFF3 may provide additional evidence in colorectal cancer classification. The transcriptional regulation of ATOH1 and SPDEF on goblet cell differentiation is also demonstrated. Knock-down of either transcriptional factor decreases the goblet cell numbers, while double knock-down completely depletes goblet cell formation, suggesting the co-operative regulation of SPDEF and ATOH1 on goblet cell differentiation. In addition, CA12-positive but not -negative cells can give rise to goblet cells with the expression of MUC2, TFF3 and SPDEF. This indicates CA12 may act as a potential cellular surface marker to identify goblet cell progenitors (**Chapter 5**).

In summary, this thesis screens goblet cells in colorectal cancer cell lines and characterise the first goblet cell transcriptome, which provides the foundation to understand regulatory control of goblet cell differentiation. (354 words)

ACKNOWLEDGEMENTS

I am extremely grateful to my supervisor Professor Sir Walter Bodmer for giving me the opportunity to work in his laboratory, as well as his continuous guidance throughout my DPhil project. None of my work would have been possible without his supervision. I have been overwhelmingly motivated by his enthusiasm, inspiring ideas, tremendous support, and the spirit of a true scientist. It is my great honour and pride for life of working with Prof Sir Walter Bodmer.

I would like to express my sincere gratitude to all those who have helped me make this thesis completed, particularly Laura Colling, Tom Barnes, Dr Chi Zhang and Prof Valentine Macaulay for their thorough revision and valuable suggestions. Furthermore, I am really thankful to Dr Neil Ashley, for his close collaboration and fruitful discussion. They have always been my big support in Oxford throughout the past three years, especially when the rain set in my life.

I am also grateful to other members in the Cancer and Immunogenetics Laboratory. Huge thanks to Dr Djamila Ouaret for helping me make the right experimental decisions. Great thanks to Dr Mustak Ibn Ayub, Jens Puschhof, Peter Kalugin and Marahaini Musa. You are all extraordinary scientists, and it has been a privilege to work with you.

I would like to thank Kevin Clark, Sally Clark and Craig Waugh for the expertise in cell sorting, Christoffer Lagerholm for the expertise in fluorescent microscope, and Nikolaos Barkas, Nicki Grey and Emmanouela Repapi for the expertise in bioinformatics. I also appreciate the opportunity to discuss with Dr Neil Ashley, Prof Valentine Macaulay, Dr Lai Mun Wang and Prof Xin Lu for my transfer of status and confirmation.

I would like to thank my mother and other family members Zhaohui Li and Haiyan Zhang for their love and support. I thank all my friends, Dr Chen Zhao, Dr Wei Liu, Dr Mei-Yi Sun, Dr Yanchun Peng, Lina Guo, Koon Hwee Ang, Zetian Gao, Dr Feng Zhou, Haochao Huang, Bo-Han Zhang, Yuning Cai, Wei Wu and Professor Yan Zhao. Last but not the least, a special thanks to Dr Chi Zhang, as my trusted friend, thank you for always being on my side, helping me through various situations with kind encouragement and selfless dedications.

Thank life for bringing me to Oxford, and I am sure it will take me to a bright future!

DECLARATION

I, Haoyu Liu, hereby declare that the work on which this thesis is based is my original work and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. The work is original, except where indicated by reference in the text.

Signed: Date:

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

CHAPTER 2

MATERIALS AND METHODS

CHAPTER 3

SCREENING OF GOBLET CELL DIFFERENTIATION IN A PANEL OF 64 HUMAN COLORECTAL CANCER CELL LINES

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AKR1B10	Aldo-Keto Reductase Family 1 Member B10
AGR2	Anterior gradient protein 2 homolog
AGR3	Anterior gradient protein 3 homolog
Atoh1	Atonal homolog 1 (Drosophila)
bHLH	Basic Helix-Loop-Helix
BMI1	B lymphoma Moloney murine leukaemia virus insertion region 1
BMP	Bone morphogenetic protein
CA12	Carbonic anhydrase XII
CBRG	Computational Biology Research Group
cDNA	Complementary DNA
CDX1	Caudal-type homeobox transcription factor 1
CDX2	Caudal-type homeobox transcription factor 2
CPM	Count per million
CRC	Colorectal cancer
CRUK	Cancer Research UK
CSC	Cancer stem cells
DBZ	Dibenzazepine
DEPC	Diethyl pyrocarbonate
DLL1	Delta Like Canonical Notch Ligand 1
DLL2	Delta Like Canonical Notch Ligand 2
DLL3	Delta Like Canonical Notch Ligand 3
DLL4	Delta Like Canonical Notch Ligand 4
DMEM	Dulbecco's Modified Eagle's Medium
DMSO	Dimethyl suphoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DSL	Delta/Serrate/Lag-2
ER	Endoplasmic reticulum
FACS	Fluorescence activated cell sorting
FBS	Fetal bovine serum
FCGBP	Fc fragment of IgG binding protein
FOXJ1	Forkhead box protein J1
FS	Forward scatter
Hath1	Atonal homolog 1 (human)
Hes1	Hairy and enhancer of split 1 (Drosophila)
HP	Hyperplastic polyp
µg	Microgram

µL	Microlitre
µm	Micro-meter
µM	Micro-molar
KLF4	Kruppel-like factor 4
logFC	log-fold change
mAb	Monoclonal antibody
Math1	Atonal homolog 1 (murine)
mg	Milligram
mL	Milliliter
mRNA	Messenger ribonucleic acid
MUC2	Mucin 2
NICD	Notch intracellular domain
NKX3.1	NK3 Homeobox 1
ng	Nanogram
nM	Nanomolar
pAb	Polyclonal antibody
PBS	Phosphate buffered saline
PDI	Protein disulphide isomerase
PFA	Paraformaldehyde
PI	Propidium iodide
PTS	Proline, Threonine and Serine
rpm	Revolutions per minutes
SEM	Standard error of mean
siRNA	Small interfering RNA
SPDEF	SAM pointed domain-containing Ets transcription factor
SSC	Saline Sodium Citrate
SSP	Sessile serrated polyp
STAR	Spliced Transcripts Alignment to a Reference
TFF3	Trefoil factor 3
TR	Tandem repeats
TSA	Traditional serrated adenoma
vWF	Von Willebrand factor
WFA	Wisteria Floribunda

CHAPTER 1

INTRODUCTION

1.1 Colorectal anatomy and crypt microarchitecture

1.1.1 Anatomy of the colon and rectum

The colorectum is the last part of human gastrointestinal tract before the anus. The length of entire colorectum is approximately 150cm, with a diameter that varies on individuals from 2.5 to 5cm. Its main function is to absorb water and salts, to act as a temporary storage for stool before elimination from the body and to provide the environment for gut microbes for fermentation of undigested food.

The human colorectum consists of eight segments. The first six segments, i.e. caecum, ascending (right) colon, hepatic (right) flexure, transverse colon, descending (left) colon and splenic (left) flexure, belong to the colon; while the latter two, i.e. sigmoid colon and rectum, belong to the rectum (Aigner and Fritsch, 2010). Rising from the cecum, the ascending colon travels towards the right upper quadrant to the liver under-surface, where it connects to the transverse colon via the hepatic flexure. Crossing the midline, transverse colon and descending colon are connected at the splenic flexure. Passing along the left side of abdomen, the S-shaped sigmoid colon connects the descending colon to the rectum. Though colon and rectum are identified largely anatomically, macroscopic differences were described including the presence of regular haustra or pouches, and small tissue folds in colon, in contrast to the comparatively smooth lining of rectum (Aigner and Fritsch, 2010). Several differences between left- and right-sided colon cancers have been illustrated regarding their associated

pathological and genetic alteration in colonic neoplastic transformation. The right-sided colon cancers are more frequently observed in older women, with higher rate of comorbidities and mortality (Benedix, et al., 2010). Pathologically, a higher percentage of poor differentiation and worse prognosis are observed in right-sided colon cancers, while no significant difference is found regarding synchronous distant metastases (Benedix, et al., 2010). Notably, a higher proportion of signet-ring cell carcinoma and mucinous adenocarcinoma was illustrated in right- than left-sided colon cancers (4% vs 1%, p-value < 0.01) (Nawa, et al., 2008). Genetically, the right-sided colon cancers show higher microsatellite instability (MSI) (Iacopetta, 2002; Jass, et al., 2002; Sugai, et al., 2006), while left-sided colon cancers present higher chromosomal instability (CIN) (Lindblom, et al., 2001). In addition, the right-sided colon cancers present a higher incidence of CpG island methylator phenotype (CIMP) (Iacopetta, 2002; Sugai, et al., 2006). These observations suggest the existence of different predominant mechanisms of colorectal carcinogenesis in different parts of colons, the CIMP+/MSI+ groups in right-sided colons and CIN in left-sided colons (Iacopetta, 2002; Lindblom, et al., 2001). Furthermore, the MSI+ cancers in the right-sided colons presented higher CD8/CD3 mRNA levels, supporting the cytotoxic nature of lymphocyte infiltration and immunogenicity in MSI+ colon cancers.

1.1.2 Crypt microarchitecture and stem cell niche

The basic functional unit of colorectal epithelium is the crypt. It consists of a single layer of columnar epithelial cells and forms the finger-like micro-invaginations

extending into the lamina propria (Humphries and Wright., 2008). The organisation of millions of crypts in normal colons largely increases the functional surface areas. The crypt turnover is at a high rate of usually 4-5 days, which is determined by the activity of tissue-specific stem cells that reside at the bottom of colonic crypts (**Figure 1.1**) (Humphries and Wright., 2008; Medema and Vermeulen, 2011).

The stem-cell niche represents a specialised anatomic area where a microenvironment is provided to support the self-renewal and lineage commitment of stem cells. Within niches, the multipotent colonic stem cells are surrounded by mesenchymal cells that are of a myofibroblast lineage. The pericryptal myofibroblasts regulate the maintenance of stem-like phenotype and the differentiation of colonic stem cells by secreting growth factors of several signalling pathways (Garrett et al., 2010). In particular, WNT ligands secreted by myofibroblasts can bind to the Frizzled receptors on the colonic stem cell surface and play an important role in maintaining the stem-like phenotype (Fevr et al., 2007; Clevers, 2006). By interacting with myofibroblasts, colonic stem cells are capable of self-renewing and differentiating into the transient amplifying cells.

Transient-amplifying cells reside immediately adjacent to the stem cells and obtain a greater proliferative capability at the frequency of approximately two divisions a day (Crosnier et al., 2006). In the intermediate differentiated state, these cells obtain plasticity at some level with regard to their eventual lineage commitment. After 4-5 rapid divisions, transient amplifying cells migrate towards the upper compartment of

the crypt and terminally differentiate into the three lineages: the absorptive enterocytes, the hormone-secreting enteroendocrine cells and the mucus-producing goblet cells (**Figure 1.1**) (Marshman et al. 2002; Snippert et al. 2010). The organised crypt microarchitecture and balanced lineage differentiation are key in maintaining the colonic homeostasis.

Notably, Paneth cells represent a unique type of fully differentiated cells that maintain asepsis of intestinal crypts (Bevins, et al., 2011) via secreting the anti-microbial alpha defensin 5 (DEFA5) (Salzman, et al., 2007; Ouellette, et al., 2011) and definsin 6 (DEFA6) (Jones, et al., 1993), as well as enzymes including lysozymes and phospholipase A2 (Kiyohara, et al., 1974). Paneth cells usually present in the small intestine and can be occasionally found in the normal ascending colon and other places under certain diseases (Porter et al., 2002; Ayabe et al., 2004). Paneth cells are described 'metaplastic' when observed in the areas where they are not normally present, including descending colons, sigmoid and rectum (Tanaka, et al., 2001). Meanwhile, increased number of Paneth cells, in term of 'Paneth cell hyperplasia', can be observed in proximal colons. Paneth cells often occur in inflammatory bowel disease, including ulcerative colitis and and Crohn's disease, and tend to linger on after inflammation has resolved and crypt structure improved (Surawicz et al., 1994). Thus the Paneth cell metaplasia is thought as a sign of longstanding colitis inflammation history (Tanaka, et al., 2001). Though the exact molecular mechanisms under Paneth cell metaplasia and hyperplasia remain unclear, the nucleotide-binding oligomerization domain-containing

protein 2 (NOD2) is suggested to abundantly expressed by metaplastic Paneth cells in inflammatory bowel diseases, with a role of activating Paneth cellular reponse by interacting bacterial lipopolysaccharide (Simmonds, et al., 2014; Lala, et al., 2003; Ogura, et al., 2003).

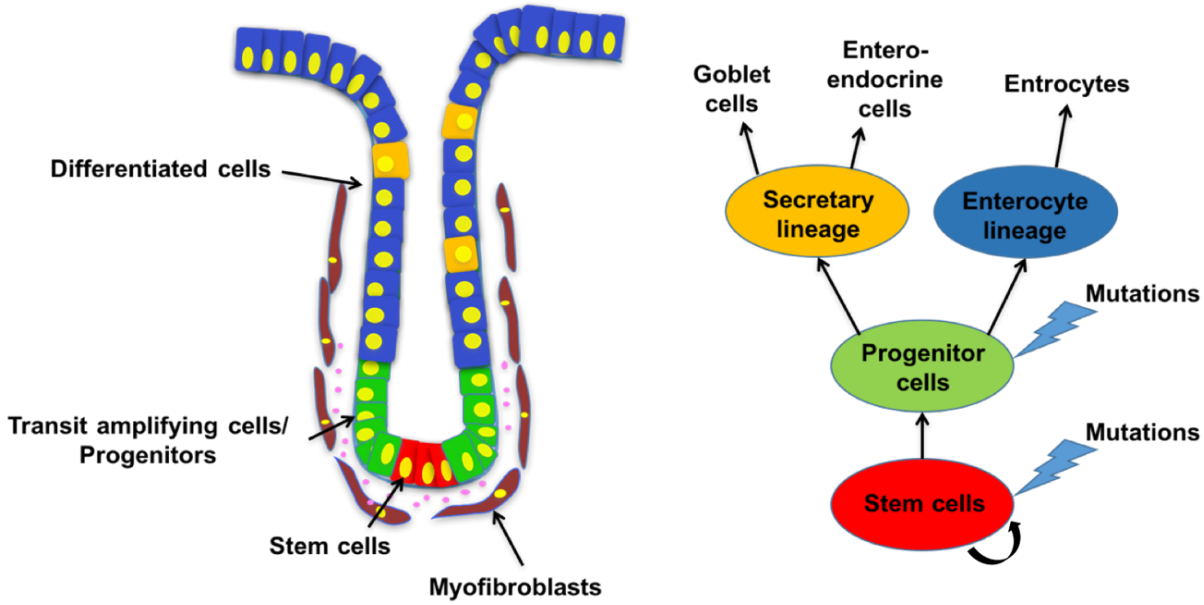


Figure 1.1 Colonic crypt microarchitecture.
Colonic stem cells reside at the bottom of the crypt, surrounded by myofibroblasts. Directly adjacent to stem cells, transient amplifying cells migrate towards the upper compartment of crypt and differentiate into secretory or enterocyte lineages. The accumulation of genetic mutations in normal stem cells or progenitor cells provide the survival advantages and result in the generation of cancer stem cells.

1.1.3 Stem cell differentiation under 3-dimensional (3D) culture

It has been indicated that adenocarcinomas are driven by cancer stem cells that maintain the ability to self-renew and differentiate into all three lineages (Ashley, 2011). When seeded and cultured in 3D conditions using Matrigel, a single cell suspension from specific colorectal cancer cell lines can either form crypt-like colonies that consist of polarised cells towards the central lumen, or the small solid non-lumen forming colonies (Richman and Bodmer, 1988; Delbuono et al., 1991; Yeung et al., 2010; Yeung et al., 2011). Cells obtained from the lumen-forming colonies can further give rise to both new lumen colonies as well as small non-lumen colonies, whereas cells derived from the small non-lumen colonies can not form lumens (Yeung et al., 2010). In addition, cells from the non-lumen colonies showed higher tumourigenic capability in mouse xenografts than lumen-forming colonies (Yeung et al., 2010). This suggests a feature of differentiation, i.e. non-lumen forming colonies are comprised of poorly differentiated cells. The lumen formation can be used as an important *in vitro* model to characterise the differentiation of cancer stem cells in the differentiating cell lines (Ashley et al., 2013). Notably, lumen formation can be used to help identify cancer stem cells in those lines that do differentiate (Ashley, 2011), but that does not say anything about identifying cancer stem cells in the cell lines that do not differentiated. For example, in the cancer cell line HCT116, virtually all the cells are cancer stem cells for that particular cancer (Yeung et al., 2010).

1.2 Goblet cells in normal colons

1.2.1 Goblet cells: history, morphology and functions

Goblet cells were originally identified in the epithelial layer of small intestine by Henle in 1837, and were first found to secret mucus by Leydig in 1857 (Young et al., 2013). These cells were termed as 'goblet' due to their goblet-like morphology. Goblet cells are highly polarised (**Figure 1.2**), with nucleus, endoplasmic reticulum (ER) and Golgi apparatus at the basal compartment. And the remaining portion of goblet cells is filled with the mucin-containing granules that are anatomically under the microvilli-shape apical plasma membrane (Ross and Pawlina, 2011).

The mucin, especially the goblet cell-specific Mucin 2 (MUC2), plays a predominant role in maintaining goblet cell morphology. In MUC2-deficient mice, cells with goblet-like morphology were no longer identifiable, while other goblet cell products, e.g. intestinal trefoil factor (TFF3), continued to be expressed (Velcich et al., 2002). In the TFF3-deficient mice, however, the goblet-morphology cells could still be identified in spite of smaller granules that contain no mucins (Taupin and Podolsky, 2003).

The major function of goblet cells is to secret mucus that covers the mucosa and serves as the front line of innate immune defence against microbes and their products in the gastrointestinal tract (Verdugo, 1990; Allen et al., 1982). The proportion of goblet cells varies largely depending on the locations and microbial distribution (Kim and Lo, 2010).

The goblet cell proportion among all differentiated lineages increases from 4% in duodenum to 16% in the distal colon. This is comparable with the increase in microbial number from the proximal intestine to the colon (Karam, 1999; Kim and Lo, 2010). Also, the germ-free mice showed fewer and smaller intestinal goblet cells than the mice raised in the normal condition (Deplancke and Gaskins, 2001). This indicates the colonic microbes can affect the goblet cell dynamics possibly via releasing bioactive factors. These results suggest a close interaction between microbial activities and goblet cell differentiation.

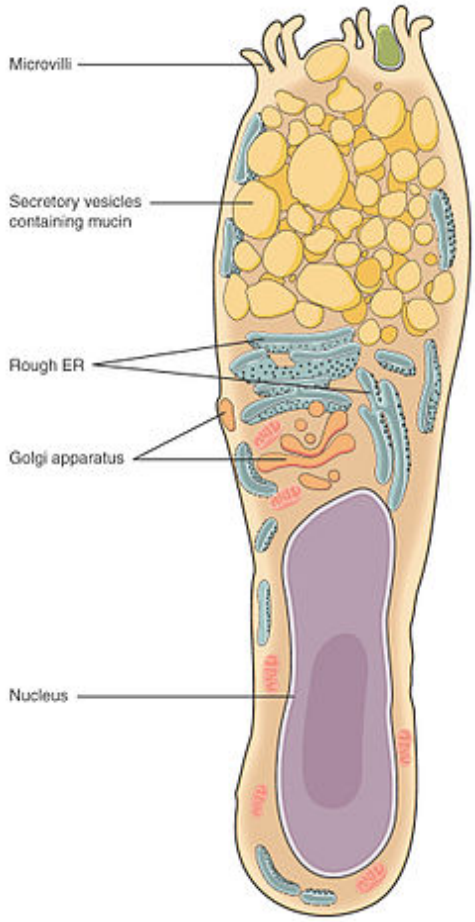


Figure 1.2 Goblet cell structure
The goblet cell is polarised with nucleus and other organelles accumulated at the cellular basal, and the remainder of the cells is filled by mucin-containing granules. This figure is adapted from Wikipedia.

1.2.2 Mucus organisation and MUC2

The mucus was organised into two layers in the human colon – the outer layer that is loosely organised and easy to penetrate by microbes, and the inner layer, which forms a condensed manner structure firmly adherent to epithelial surface (Atuma et al., 2001; Brownlee et al., 2003). As shown in **Figure 1.3A**, the inner mucus layer is estimated to be about 50um in the rat colon (approximately 100um in human colon), while the outer layer is approximately 100um containing bacteria. The relative thickness of outer and inner mucus layers also seems to depend on the microbial activities. In germ-free mice, the outer mucus layer seems relatively thicker compared to the mucus layer of conventionally raised mice (Johansson et al., 2008).

The predominant component of mucus is MUC2, a mucin that is recognised as a canonical goblet cell marker (Buisine et al., 1998; Ajioka et al., 1997). The glycoprotein MUC2 consists of approximately 5200 amino acid residues. The predicted molecular weight of MUC2 peptide backbone is ~550 kilo Dalton (Tytgat, et al., 1996), while the oligosaccharide side chains contribute to the even larger molecular weight of MUC2, which has not been fully determined. The schematic domains of MUC2 are presented in **Figure 1.3**. The central region of MUC2 is flanked by two tandem repeat regions (TRs) (Lang et al., 2007). The first TR consists of 21 repetitions of irregularly repetitive amino acid motifs that are rich in Proline, Threonine and Serine (PTS domains). The second TR consists of the perfectly repeated sequence of 23 amino acids (PTTTPITTTTTVTPTPTPTGTQT) (Lang et al., 2004) (**Figure 1.3B**). The N- and C-

terminal ends of MUC2 are rich in cysteine residues, sharing similar organisation with the von Willebrand factor (vWF), which is important in cellular adhesion and wound healing (Gum et al., 1994). The N-terminus of MUC2 contains three domains D1-D3 and an incomplete vWF domain D', while the C-terminus of MUC2 consists of vWF domain D4. The cysteine residues in these regions are important in the mucin packing and assembly before secretion (Pelaseyed et al., 2015).

After translation, the MUC2 peptides dimerise via C-terminal covalent disulphide bonds in the endoplasmic reticulum (ER) of goblet cells (Johansson et al., 2011). The correct disulphide bond-formation depends on the balanced activity of the ER, and the incorrect mucin assembly was reported to be correlated with ER stress and spontaneous inflammation resembling ulcerative colitis in mice (Heazlewood et al., 2008). Also, the anterior gradient protein 2 homolog (AGR2), a member of the disulphide isomerase (PDI) family of ER proteins, is closely involved in mucus secretion (Park et al., 2009). AGR2 was reported to covalently bind to the C-terminus of MUC2 (Park et al., 2009), but Bergström and colleagues could not reproduce this observation (Bergström et al., 2014). Even though the covalent binding between AGR2 and MUC2 is still controversial, AGR2 plays a significant role in the production of a functional mucus layer. AGR2-deficient mice produced a less mature mucus layer and resulted in colitis development (Park et al., 2009; Zhao et al., 2010; Bergström et al., unpublished).

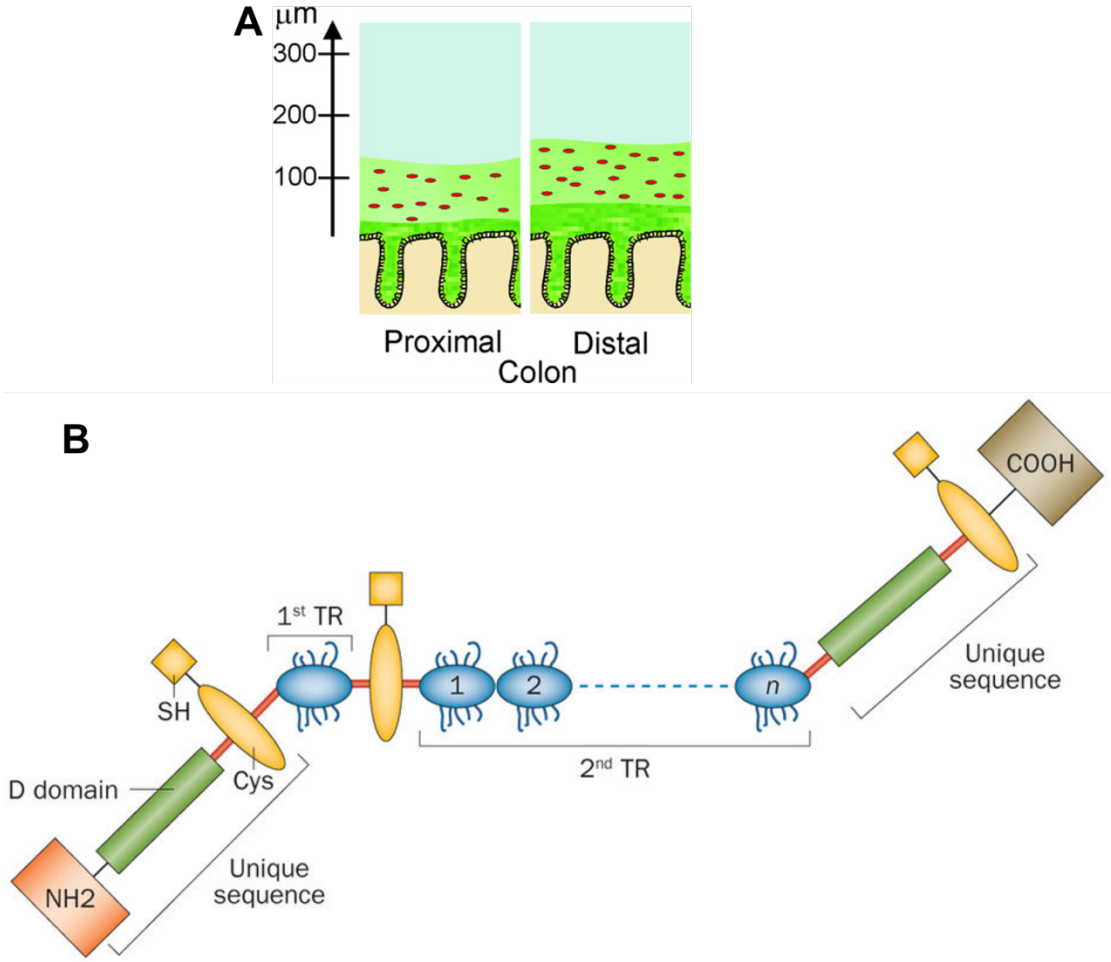


Figure 1.3 Schematic colonic mucus layers and MUC2 structure

(**A**) Representation of the inner mucus layer in the dark green, and outer mucus layer in light green, with bacteria in red dots. The left axis represents the mucus thickness as measured in the rat colon (Modified from Pelaseyed et al., 2015). (**B**) Schematic structure of a MUC2 monomer. The D domain at the N-terminal region is rich in disulphides, followed by the central part of two tandem repeats (TR). The 1st TR consists 21 repetitions of irregular amino acid motifs that are rich in Serine, Threonine and Proline. The 2nd TR consists of variable numbers of 23 amino acid repetitions. Cysteine residue-rich regions are located between N-terminal and 1st TR, between 1st TR and 2nd TR, and at the C-terminal region. (Adapted from Theodoratou et al., 2014)

MUC2 is heavily O-glycosylated at the central tandem repeats through the addition of N-Acetylgalactosamine monosaccharides (GalNAc) when exiting the ER and entering the Golgi apparatus (Bennett et al., 2012). Once in the Golgi apparatus, monosaccharides are added to the backbones of MUC2 central tandem repeats and form the complex oligosaccharides, which accounts for the weight of mucin (Larsson et al., 2011; van der Post et al., 2013).

The dysregulated glycosylation of mucin is a common feature of colorectal cancer and was reported to occur at the early stage of colorectal carcinogenesis (Theodoratou et al., 2014)., *Wisteria Floribunda* (WFA), a lectin that binds specifically to terminal N-Acetylgalactosamine (GalNac) of MUC2, was patented for distinguishing benign hyperplastic polyps (HPs) from pathologically significant polyps, including sessile serrated polyps (SSPs), traditional serrated adenomas (TSAs) and mucinous cancers based on their altered glycosylation patterns (Yeung, Patent WO 2015/198065 A1). In the trans-Golgi apparatus, the N-terminus of MUC2 is further covalently connected to form trimers via disulphide bonds (Ambort et al., 2012; Godl et al., 2002). A putative plain sheet of net-like MUC2 polymers might be formed via the C-terminal dimers and N-terminal trimers (Johansson et al., 2010). However, there is no direct experimental evidence of such net-like structures, although the dimeric and trimeric connections were suggested to contribute to the insolubility of MUC2 (Axelsson et al., 1998; Herrmann et al., 1999).

Before secretion, it is essential to pack MUC2 in a condensed form within goblet cell granules. This condensation requires a low pH and high Ca^{2+} concentrations. When the granules are released via exocytosis, a dramatic volume expansion (>1000-fold) of MUC2 occurs. This expansion is triggered by the increased pH and low level of calcium ions inside the granules (Ambort et al., 2012).

1.2.3 Other mucus components

Besides MUC2, other major mucus components secreted by goblet cells in colon include FCGBP (Johansson, et al., 2011; Johansson, et al., 2008; Harada, et al., 1997), TFF3 (Fernández-Estívariz, et al., 2003), SPINK4 (Metsis, et al., 1992; Wapenaar, et al., 2007), ZG16 (Pelaseyed, et al., 2014). Here we give a brief introduction to two well-characterised mucus components, FCGBP and TFF3.

1.2.4.1 FCGBP

Fc-γ binding protein (FCGBP) was originally identified to specifically bind to the Fc portion of IgG. It is a mucin-like protein that is composed of 5405 amino acid residues. It consists of 13 vWF D domains, but does not have the PTS domain in MUC2. In human gastrointestinal epithelium, FCGBP is highly expressed in the goblet cell granules (Harada et al., 1997).

Recent proteomic analyses of the two mucus layers in human colons suggested that FCGBP is strongly connected with MUC2 protein via disulphide bonds in single or

multiple vWF D domains in the inner firm mucus layer of the colon (Johansson et al., 2009). Moreover, FCGBP was also reported to form heterodimers with another important mucus component trefoil factor 3 (TFF3) via disulphide bonds (Albert et al., 2010). These observations indicated the significance of the covalent connection of FCGBP in crosslinking and stabilising mucus organisation. This was further confirmed by the depletion of FCGBP in the rat colon during development of the colitis induced by dextran sulphate sodium treatment (Feng et al., 2007).

1.2.4.2 TFF3

The trefoil factor family consists of three small secreted peptides with a molecular weight ranging from 6.5 to 12 kDa (Kim and Lo, 2010). All the three trefoil factors are important in maintaining epithelial restitution throughout gastrointestinal tract but with distinct anatomical distribution. TFF1 is expressed in the gastric surface foveolar cells, while TFF2 is highly expressed in mucous neck cells (Kim and Lo, 2010; Taupin and Podolsky, 2003). TFF3 (previously known as 'intestinal trefoil factor') is abundantly expressed in the theca of colorectal goblet cells. Besides intestines, TFF3 is also produced in the salivary glands (Jagla et al., 1999), endocrine pancreas (Jackerott et al., 2006), uterus (Wiede et al., 2001), vagina (Madsen et al., 2007), urinary tract (Rinnert et al., 2010), hypothalamus (Jagla et al., 2002), Vater's ampulla (Paulsen et al., 2005), esophagus (Kouznetsova et al., 2007), conjunctiva (Langer et al., 1999), respiratory tract (Wiede et al., 1999), efferent tear ducts (Paulsen et al., 2002), and gastric antrum

and cardia (Kouznetsova et al., 2007; Kouznetsova et al., 2004). Although the trefoil factors are highly expressed by various types of secretory cells (Albert et al., 2010), TFF3 is indicated as a goblet cell marker in human colons (Gött, et al., 1996; Bergstrom, et al., 2008)

The cysteine-rich peptide of TFF3 consists of 59 amino acid residues, including 7 cysteine residues (Wong et al., 1999; Hoffman et al., 2001; Taupin et al., 2003). The integrity of mucus and the intestinal homeostasis rely on the regulated expression of TFF3, which plays a significant role in the mucosal repair and regeneration processes (Albert et al., 2010; Hoffman et al., 2001). For example, it was documented that TFF3 is able to facilitate 'restitution', the cell migration-based tissue repair, via chemotaxis (Chwieralski et al., 2004). This statement is further supported by the observation that injury in distal and proximal gastrointestinal tracts can result in the up-regulated TFF3 expression (Mashimo et al., 1996). Mice deficient in trefoil factors showed the disrupted mucosal healing and extensive colitis after epithelial injury (Mashimo et al., 1996).

As a cysteine-rich peptide, TFF3 can form a homodimer via cysteine-57 (Kinoshita, et al. 2000; Chinery et al., 1995), and the three-dimensional structure of both TFF3 monomers and homodimers were characterised (Lemercinier et al., 2001; Muskett et al., 2003). Both TFF3 homodimers and monomers were reported to show mitogenic effects *in vitro* (Kinoshita et al., 2000; Oertel et al., 2001). However, only TFF3

homodimers but not monomers improved the colitis in two different rat colitis models induced by dextran sulphate sodium and mitomycin C respectively (Poulsen et al., 2005). And this effect could only be achieved by the luminal application, but not the parenteral delivery (Poulsen et al., 2005). Similarly, only TFF3 homodimers were identified to have anti-apoptotic effects *in vitro* (Kinoshita et al., 2000). This was also confirmed by Taupin and colleagues via the exogenous TFF3 expression in the human colorectal carcinoma cell line HCT116 and the nontransformed rat intestinal epithelial cell line IEC-6 (Taupin et al., 2000). Their research also showed that the TFF3-deficient mice presented the increased colonocyte apoptosis without changes in receptor-related (TNFR/Fas) or stress-related (Bcl-family) cell death regulators (Taupin et al., 2000).

As described before, TFF3 is able to form heterodimers with FCGBP, whose N-terminal portion is covalently connected to MUC2, thus forming a trimeric structure together (Johansson et al., 2009). It is noteworthy that the expressions of MUC2 and TFF3 are not regulated co-ordinately in the rat intestine (Matsuoka et al., 1999). The precise mechanisms that regulate MUC2 and TFF3 separately have not yet been well characterised.

1.3 Regulation of goblet cell differentiation

1.3.1 Notch pathway

The Notch pathway is highly conserved across almost all multicellular species (VanDussen et al., 2012) and has been shown to play a central role in regulating stem

cell maintenance and intestinal homeostasis by interacting with several developmental pathways - a synergy between Notch and Wnt signalling was suggested to induce the intestinal adenomas formation, especially in the colon (Fre et al., 2008). In the human colon, the Notch pathway consists of four types of receptors, i.e. Notch 1-4, and several ligands: Jagged1, Jagged2, DLL1, DLL3 and DLL4 (Artavanis-Tsakonas et al., 1999). Notch ligands are type I transmembrane proteins of the Delta/Serrate/Lag-2 (DSL) family. DLL1 and DLL4 are predominantly expressed on the cellular surface of goblet cells at the colonic crypt bottom, which provide an essential niche for stem cell activity (Rothenberg et al., 2012).

As shown in **Figure 1.4**, Notch is activated upon the binding of its ligands from an adjacent cell (Gray et al., 1999). When engaging a ligand, the extracellular domain of Notch is cleaved by an ADAM-family metalloprotease (van Tetering et al., 2009). The cleaved Notch extracellular domain together with the bound ligand enters the ligand-expressing cells through endocytosis. However, the specific mechanism and functions of this process are not fully understood. Following the first metalloprotease cleavage, a second cleavage occurs on the intracellular domain of Notch receptors by γ-secretase (Desbordes and López-Schier, 2005). This cleavage liberates the Notch intracellular domains (NICD) into the Notch receptor-expressing cell, which then translocate into the nucleus and forms a transcriptional complex with the transcriptional factor RBPJ (Recombination Signal Binding Protein for Immunoglobulin Kappa J Region), also known as CSL or CBF1. The transcriptional complex further recruits transcriptional

co-activators, e.g. MAML-1 (Mastermind-like protein 1), which can be activated upon NICD-binding (Qiao et al., 2009; Desbordes and López-Schier, 2005). This results in the expression of Notch responsive genes, in particular *Hes1*, *Hes5* and *Hey1*.

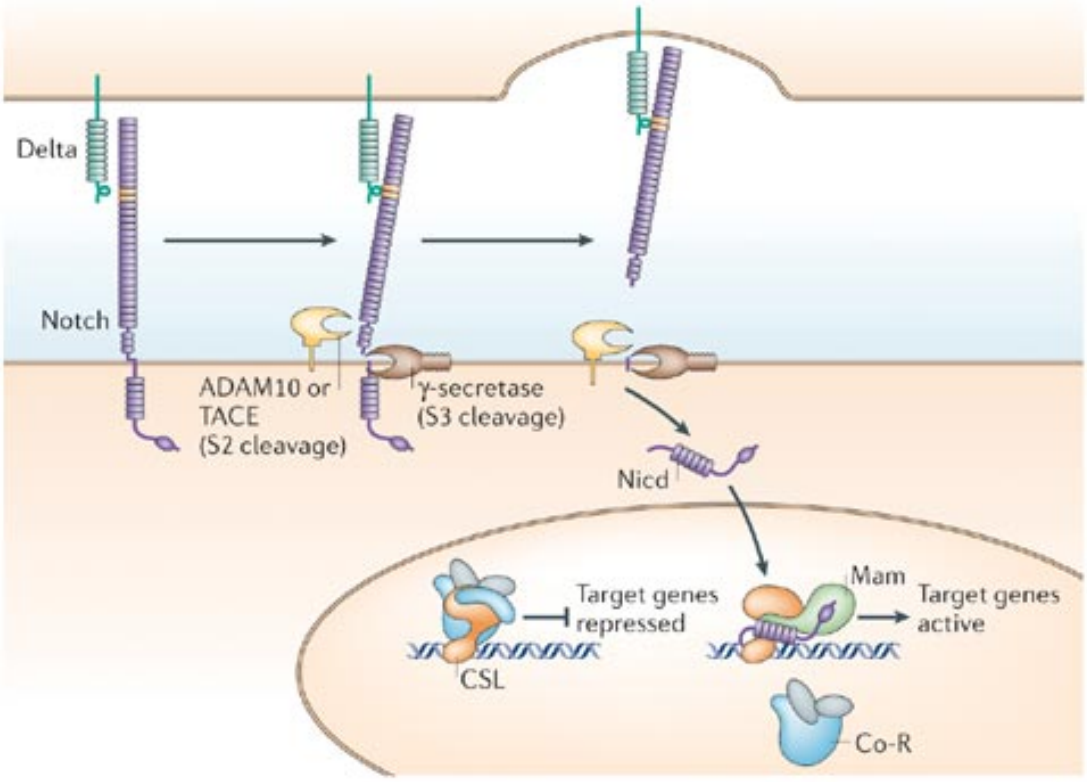


Figure 1.4 Notch signalling transduction.
Upon the binding of ligands, Notch receptors are cleaved twice by metalloprotease and γ-secretase, which releases NICD at the cytoplasmic side. NICD translocates into nucleus and replaces the co-repressor by recruiting co-activators, e.g. CSL. This forms the transcriptional complex that binds and activates the target genes.
This figure is adapted from Nature Reviews Molecular Cell Biology (Bray, 2006)

Genetically engineered mice showed a Notch-controlled binary cell fate switch in intestinal progenitors between secretory and absorptive lineages (Jensen et al., 2000). The activated Notch signalling promotes differentiation of progenitor cells towards the enterocyte lineage over the secretory lineage (Jensen et al., 2000; Fre et al., 2005; Gerbe et al., 2011; Pellegrinet et al., 2011). Gut-specific conditional knockout of RBPJ leads to the accumulation of post-mitotic goblet cells at the cost of proliferating progenitor cells (van Es et al., 2005). In contrast, forced expression of NICD in mice gave rise to the increased number of immature goblet cell and decreased goblet cell differentiation (Fre et al., 2005). Notch pathway blockade using γ-secretase inhibitors DBZ (Dibenzazepine) (van Es et al., 2005) or DAPT (N-[N-(3,5-difluorophenacetyl)-l-alanyl]-S-phenylglycine t-butyl ester) (Dovey et al., 2001; Geling, 2002), disrupted the stem cell homeostasis, by largely increasing the goblet cell differentiation and down-regulating the expression of Hes1, a key basic helix loop helix (bHLH) transcriptional factor that is directly responsive to Notch pathway. The striking increase of goblet cell numbers under γ-secretase inhibitor treatment mimicked the observation in RBPJ-deficient mice. Moreover, genetic knockout of various Notch receptors and ligands, including Notch 1 and Notch 2 (Riccio et al., 2008), and DLL1 and DLL4 (Pellegrinet et al., 2011), leads to an increased number of goblet cells and decreased proliferation of the crypt. Similar phenotypes were observed by using a combination of blocking antibodies against Notch1 and Notch2 receptors (Wu et al., 2010). Taken together, these observations suggest a significant role of Notch pathway as a gatekeeper to regulate intestinal goblet cell differentiation.

1.3.2 ATOH1

Atonal homolog 1 (ATOH1, its homolog known as 'Hath1' in human and 'Math1' in mouse) is a helix-loop-helix transcriptional factor and is negatively regulated by *Hes1*, the Notch target gene (**Figure 1.5)** (Yang et al., 2001; Shroyer et al., 2007). Inhibition of ATOH1 by the Notch pathway provides an important mechanism for the cell fate switch between secretory lineage and enterocyte lineage (Noah and Shroyer, 2013).

As previously described, the forced overexpression of Notch signalling results in increased Hes1 activity and decreased Math1 expression, and further depletion of all secretory lineages (**Figure 1.5)** (van der Flier and Clevers, 2009). By contrast, depletion of ATOH1 resulted in the loss of goblet cells, enteroendocrine cells and Paneth cells in mouse small intestines (Yang et al., 2001). This observation was further confirmed by a small intestine-specific Math1-knockout in mice that presented a normal crypt-villi microarchitecture solely formed by enterocytes (Shroyer et al., 2007). This demonstrated that ATOH1 is the key mediator of Notch pathway in regulating the differentiation of all secretory lineages (Shroyer et al., 2013). Notably, when the Notch pathway was blocked via DBZ or knocking out RBPJ in Math1-deficient mice, crypt structure, proliferation and active intestinal stem cells remained, but goblet cell differentiation was disrupted in absence of Math1 (van Es et al., 2010; Kim et al., 2011; Kazanjian et al., 2011). It was further suggested by Noah and Shroyer that in absence of ATOH1, Notch-Hes1 signalling is dispensable for absorptive lineage commitment in intestine-specific Math1-null mice, while its primary role lies on the ATOH1-

mediated goblet cell differentiation (Kazanjian et al., 2010; Noah and Shroyer, 2013). In addition, the enforced ATOH1 expression in mouse intestines resulted in the secretory lineage enforcement at the cost of enterocytes. This also led to the dysregulated crypt microarchitecture with proliferating cell migrating towards the villi (VanDussen and Samuelson, 2010).

There may also be a feedback loop from ATOH1 on Hes1. In the Math1-deficient mice, an up-regulated Hes1 expression was observed, even with the genetic mutation of *RBPJ*, the Notch-pathway component upstream to Hes1 (Kazanjian et al., 2010).

Taken together, these observations suggested the reciprocal roles of ATOH1 and Notch-Hes signalling on the cell fate switch between enterocytes and secretory cells. The Notch-mediated Hes1 expression represses ATOH1, thus directing progenitors towards enterocytes. In the cells that escape Notch activation, ATOH1 expression occurs, which directs the secretory lineage specification, though the specific trigger has not been fully understood.

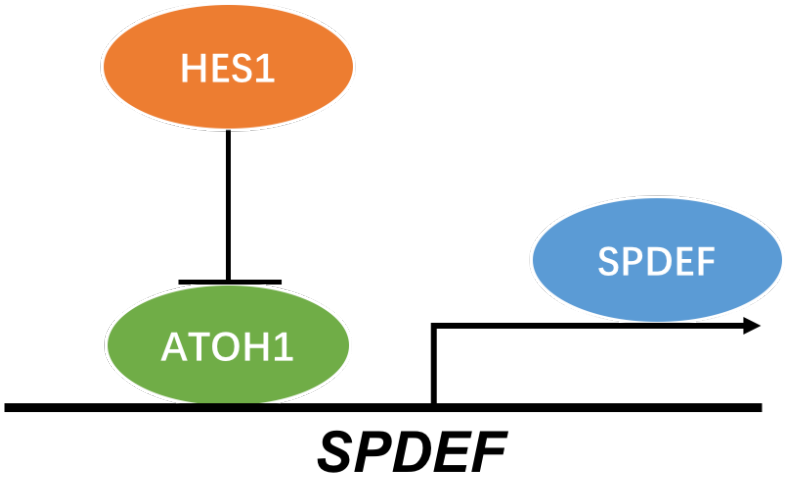


Figure 1.5 Interaction between ATOH1 and SPDEF in mouse small intestines
HES1, a direct Notch pathway target gene, negatively regulates ATOH1 expression and further represses goblet cell differentiation. ATOH1 mediates the Notch inhibition on secretory lineage by binding to the core promoter region of SPDEF.

1.3.3 SPDEF

SPDEF (SAM pointed domain containing ETS transcription factor) is a member of the ETS (E26 transformation specific) transcription factor family. The transcriptional factor SPDEF is featured by its preferential DNA binding site on the GGAT core sequence over the ETS-family consensus GGAA. SPDEF, originally identified to be expressed in prostate epithelial cells, trans-activates prostate-specific antigen promoters by interacting with NKX3.1 in an androgen-dependent manner (Oettgen et al., 2000; Chen et al., 2002). It was later identified in other locations, including gastric, breast, airway, small bowel and colon (Yamada et al., 2000). The expression of SPDEF in prostate and breast cancers was shown to be largely reduced compared that in the normal prostate and breast epithelium (Sood et al., 2007). SPDEF was also indicated to repress cellular migration and invasion, and to serve as a potential tumour suppressor (Turner et al., 2008; Feldman et al., 2003).

SPDEF was suggested as a tumour suppressor by repressing β-Catenin activity (Noah et al., 2013). By examining over 500 human CRC samples and over 80 normal controls, Noah and colleagues identified the loss of SPDEF in ~85% tumours and its correlation with the normal-adenoma-adenocarcinoma progression (Noah et al., 2013). This observation is also confirmed by mouse models – the *Spdef*$^{-/-}$ mice showed the

development of ~3-fold more colonic tumours under DSS treatment than the wild-type (Noah et al., 2013).

In the human colon, SPDEF was suggested to express in an ATOH1-dependent manner (Shroyer et al., 2007). Consistent with ATOH1, the expression of SPDEF was also negatively regulated by Notch signalling, and the DBZ-induced Notch blockade up-regulated the expression of SPDEF, which is depending on ATOH1 (Kazajian et al., 2010). Recent ChIP-Seq data suggested a direct binding of ATOH1 on the core promoter of SPDEF (Lo et al., 2017).

Taken together, despite most data above from only murine studies, ATOH1 and SPDEF may serve as important transcriptional factors downstream of the Notch signalling pathway in human as well. Further characterisations were conducted to investigate their regulation on the goblet cell differentiation and maturation in this thesis.

1.4 Goblet cells in colorectal cancers

Colorectal cancer is a malignancy that develops in the colon or rectum (Reya et al., 2001). It is one of the leading cause for cancerous death worldwide, especially in Western countries. In the UK, there are over 38,000 newly diagnosed patients per year, with an annual morbidity of more than 15,000 (Siegel et al., 2014). The geographical distribution of colorectal cancer in Western countries may be associated with diet habit including a high intake of red meat and refined grains and sugars (Chan et al., 2010).

The risk of colorectal cancer increases with age. Nearly 60% of colorectal cancer cases occur in the people over 70 years old (CRUK, Bowel Cancer; NCI, Colon and Rectal). Risk is increased with a positive family history. The family history involves the inherited conditions or syndromes with altered genetic expression, including Familial adenomatous polyposis (FAP) and hereditary non polyposis colon cancer (HNPCC) (also known as Lynch syndrome) (Jasperson et al., 2010).

The accumulation of genetic mutations over time is widely recognised for the adenoma-carcinoma progression of colorectal cancers (Vogelstein et al., 1988; Smith et al., 2002). An effective diagnosis and endoscopic removal of early-stage polyps have been shown to reduce the further development of colorectal cancer (Winawer et al., 1993).

There are two types of colorectal cancers that are highly involved with aberrant goblet cell differentiation and MUC2 production: mucinous carcinoma and signet ring

carcinoma (Kim and Lo, 2010). These two types of colorectal cancers should be clearly distinguished from each other. Mucinous carcinoma refers to the tumour with abundant mucin secretion that accounts for at least 50% of its total volume. Signet ring carcinoma, on the other hand, represents the tumour with over 50% cells containing intracellular mucins (Bosman et al., 2010). Compared to signet ring carcinoma (5-year survival rate 9-30%) (Belli et al., 2014), mucinous carcinoma is associated with better prognosis in colon (5-year survival rate of ~32.5%) (Chand et al., 2014). Mucinous carcinoma is more frequently observed in the colon than the small bowel and representing 6-19% colorectal cancers (Kim and Lo, 2010). Compared to non-mucinous cancers, the mucinous cancers are more invasive, usually with higher frequency of lymph node metastasis, microsatellite instability and BRAF mutations (Tanaka et al., 2006).

The expression and glycosylation of MUC2 is dysregulated in human colorectal cancers. An abundant amount of MUC2 was expressed and secreted in mucinous carcinomas, while in non-mucinous carcinomas, goblet cell differentiation and MUC2 expression are largely reduced (Kim, Y. S. et al., 2010). The high expression of MUC2 in mucinous carcinomas resulted from an altered genetic and epigenetic regulation of MUC2, e.g. promoter hypomethylation (Okudaira, K. et al., 2010) or up-regulated expression and increased binding of secretory lineage-specific transcriptional factor ATOH1 (Leow, C. C. et al., 2004; Park, E. T. et al., 2006). In addition to the decreased MUC2 expression, non-mucinous carcinomas also displayed an altered crypt morphology and cellular migration in the $MUC2^{-/-}$ mice (Velcich, A. et al., 2002).

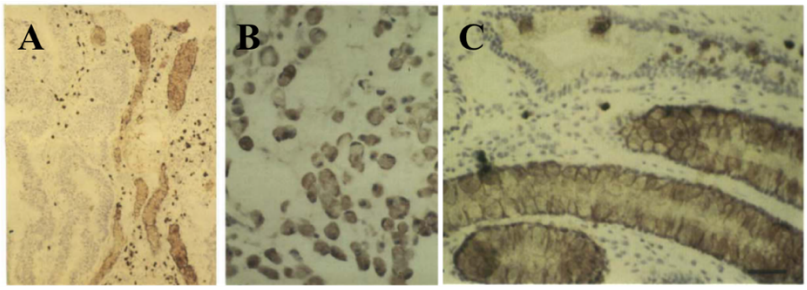
In both mucinous and non-mucinous carcinomas, the truncated glycan structures of MUC2 and the high expression of neo-glycans, e.g. sialylLeA and sialylLeX antigens, were observed (Hollingsworth et al., 2004; Andrianifahanana et al., 2006; Kim et al., 2008; Byrd et al., 2004). In addition, it has been suggested that high levels of goblet cell differentiation were observed in colorectal cancer cell lines with a higher metastastic capacity (Byrd, J. C. et al., 2004).

1.5 PR5D5, an in-house mAb targeting goblet cells

The in-house antibody PR5D5 is able to identify goblet cells in tissue and in cultured cells. In 1987, a panel of 12 monoclonal antibodies were produced by immunising and boosting the BALB/c mice with normal mucosal scrapings, cell membranes and the HT29 colon carcinoma cell line.

One of these monoclonal antibodies, PR5D5, was found to react specifically with the mucus within goblet cells without cross-reacting with the columnar cells (**Figure 1.6D**). In the normal colonic crypt, PR5D5 only stained the goblet cells adjacent to the unstained carcinomas (**Figure 1.6A**). In the signet ring carcinoma, all cells were PR5D5-positive (**Figure 1.6B)**. In the metastatic polyp, only occasional PR5D5-positive cells were identified, compared to the normal crypt. (**Figure 1.6C**). The reduction of the mucus-reacting PR5D5 antibody reactivity in the malignant polyps is consistent with the reduced number of goblet cells in colorectal carcinomas, in which the differentiation of goblet cells is largely dysregulated.

Thus, the in-house PR5D5 mAb is widely used throughout this thesis, because of its exceptional affinity and specificity towards goblet cells and the large volume in stock. The molecular target of PR5D5 will be characterised in **Chapter 3**.



D

	Group 1: MAbs reactive with mucus determinants ± other cell constituents						Group 2: MAbs reactive with cell constituents other than mucus					
	Antibody						Antibody					
	PR.4D4	PR.5D5	PR.3A5	PR.4D1	PR.5C5	PR.4D2	PR.1A3	PR.4B10	PR.3B10	PR.4C5	PR.4D6	PR.6B5
Specificity												
Goblet cells												
Mucus	+	+	+	+	+[1]	+	−	−	−	−	−	−
Cytoplasm	−	−	+	+	+	+	−	−	+	−	+	−
Cell membrane	−	−	+	+	+	+	−	+	+	+	+	+
Columnar absorptive cells												
Cytoplasm	−	−	+	+	+	+	+[2]	+	+	+	+	−
Cell membrane	−	−	+	+	+	+	+	+	+	+	+	+
Microvillous brush border	−	−	+	+	+	+	+	−	+	−	−	−
Neuroendocrine cell granules	−	−	−	−	−	−	−	−	−	−	−	−
Epithelial basement membrane	−	−	−	−	−	−	−	−	−	+	−	−
IgG subclass	1	1	1	1	1	3	1	2a	1	1	1	1
Resistance of antigenic determinants to formalin fixation	−	+	+	+	+	+	−	−	+	−	−	−

[1]Mucus of some goblet cells.–[2]Reactivity at upper poles of cells; in most crypts reaction strongest at upper crypt and surface epithelium.

Figure 1.6 Representative staining with PR5D5

Representative PR5D5 staining in (**A**) the normal colonic crypt, (**B**) signet ring carcinoma, and (**C**) the edge of a metastatic polyp. (**D**) Among the PR-series monoclonal antibodies, PR5D5 presented highly specific staining with mucus in goblet cells, and showed resistance to antigenic determinants to formalin fixation. (Adapted from Richman and Bodmer, 1987)

1.6 Understanding gap and technical barrier

Several gaps still exist in understanding goblet cell differentiation, including the lack of goblet cell surface markers, the absence of goblet cell transcriptomic profile and the further characterisation of key transcriptional factors in the context of human colorectal cancer cell lines.

Recent advances in transcriptomic characterisation, including mRNA microarray and RNA-seq, have largely deepened the understanding of cells at various differentiation states. The research of goblet cell differentiation, however, is largely limited by the lack of goblet cell surface markers. Using intracellular markers, e.g. MUC2, to identify goblet cells requires fixation, permeabilisation and intracellular staining. It can easily result in irreversible RNA degradation through intra- and extra-cellular RNase contamination. In addition, the small overall proportion of goblet cells in human colorectal cancer cell lines results in a long sorting period adding another layer of complexity in RNA preservation. This stands as a technical barrier to understanding goblet cell differentiation in colorectal cancers.

To optimise the RNA extraction while maintaining staining patterns for precise sorting of goblet cells, it is important to choose suitable fixatives and permeabilisation agents alongside with additional RNase inhibitors to minimise RNA degradation. There are two major types of fixatives commonly used in the intracellular staining. The first are coagulation-based dehydrants, including methanol, ethanol and acetone. These

fixatives disrupt the hydrophobic bonds, denature proteins and change their solubility and conformation by exposing the internal hydrophobic groups at 4°C. The second type are the cross-linking reagents, i.e. the aldehydes, including paraformaldehyde, glutaraldehyde and potassium permanganate.

After fixation, the permeabilisation of cell membranes allows antibodies to access the intracellular antigens. Two types of permeabilisation reagents that are commonly used - the organic solvents and the detergents. Organic solvents include methanol, ethanol and acetone. This type of reagent permeabilises cellular membranes by dissolving cell surface lipids. As described before, they can also precipitate proteins and be used as both fixative and permeabilisation reagents at the same time. Detergents, such as saponin, can permeabilise cell membrane by selectively removing cellular cholesterol. Triton, another widely used detergent, contains the hydrophilic head groups that unselectively interact with cell membrane lipids. These fixative and permeabilisation agents, however, cannot destroy the structure of RNases or inactivate them.

Thus, there is a significant technical barrier by using traditional intracellular staining methods for RNA isolation from the fixed and sorted samples. This thesis will explore the optimisation to overcome this barrier and characterise the goblet cell transcriptomic profile that retains the potential for discovery of novel surface marker discovery providing insights into goblet cell fate decision.

1.7 Aims and Objectives

1) In order to fully understand the transcriptomic nature of goblet cells that can only be identified by the intracellular marker, this project aims to establish and optimize a novel protocol to isolate RNA from the fixed and FACS-enriched cells;

2) This project aims to find out novel goblet cell markers or key regulators by characterising and analysing the transcriptome of goblet cells in human colorectal cancer cell lines;

3) Based on the goblet cell screening in a panel of human colorectal cancer cell lines, as well as transcriptome characterisation, this project aims to illustrate the effects of key regulators on the differentiation of goblet cells in colorectal cancers

CHAPTER 2

MATERIALS AND METHODS

2.1 Reagents and suppliers

Unless otherwise stated, inorganic chemicals and reagents that are not included in standard kits were supplied from Sigma-Aldrich (Poole, UK). Tissue culture flasks, plates and tubes were purchased from Corning (New York, USA). Tissue culture medium, including DMEM, RPMI and IMDM, and supplementary antibiotics were obtained from Invitrogen (Paisley, UK).

2.2 Two-dimensional cell culture

2.2.1 Colorectal cancer cell lines

The human colorectal cancer cell lines C10, C106, C125PM, C2BBe1, C32, C80, C84, C99, CACO2, CAR1, CC20, CL40, COLO201, COLO320DM, COLO678, CX1, DLD1, GP2D, GP5D, HCA46, HCA7, HCT116, HCT15, HDC111, HDC114, HDC142, HDC57, HDC73, HDC82, HRA19, HT29, HT55, JHCOLOY1, JHSKREC, LIM1215, LIM1863, LIM2405, LOVO, LS1034, LS123, LS174T, LS180, NCIH508, NCIH747, OXCO1, OXCO3, PMFKO14, RCM1, RKO, RW2982, RW7213, SKCO1, SNUC2B, SW1222, SW1417, SW403, SW48, SW480, SW837, SW948, T84, TITTKB, VACO10MS and WIDR were obtained from the cryogenic storage of the Cancer and Immunogenetics Laboratory (CIL), Weatherall Institute of Molecular Medicine (WIMM), Oxford, UK.

All the cell lines, except for the suspension cell lines COLO201 and LIM1863, are adherent. The C80, C84 and C99 cell lines were originally established by Dr David Bicknell at CIL. The SW1222 cell line was gifted from Dr Meenhard Herlyn of the Wistar Institute, Philadelphia, USA. The remaining cell lines were acquired from the American Type Culture Collection (ATCC) (Manassas, Virginia, USA), the Deutsche Sammlung von Mikroorganismen und Zellkulturen, GmbH (DSMZ) (Braunschweig, Germany), or the European Collection of Cell Cultures (ECACC) (Salisbury, UK).

2.2.2 Cell culture conditions

All human colorectal cancer cell lines were cultured with medium in regular-attached flasks or plates. Specifically, C10, C106, C125PM, C80, C84, C99, CAR1, COLO320DM, HCA46, HDC111, HDC114, HDC142, HDC57, HDC73, HDC82, HRA19, OXCO1, OXCO3 and VACO10MS were cultured with IMEM medium. The cell lines LIM1215, LIM2405, NCIH508, NCIH747 and SNUC2B were cultured in RPMI medium. The remaining cell lines were grown in DMEM medium. The complete media were prepared by supplementing the media with 10% fetal bovine serum (FBS) (Biosource, Nivelle, Belgium) and 1% penicillin/streptomycin (Invitrogen).

2.2.3 Cell culture maintenance

To avoid potential cross-contamination, maximum 12 cell lines were cultured at any one time; For each single batch, maximum 6 flasks of cell lines were transferred into sterile hood for cell passage or harvesting. Cell lines were carefully checked and

examined to ensure their identity. For each batch, the cell lines were snap frozen immediately after recovery and afterwards every 2-3 weeks for DNA extraction, and SNP analysis for each batch was conducted by Dr Wilding in Bodmer's Lab.

All cells in flasks and plates were incubated at 37°C in a humidified 10% CO2 atmosphere before reaching 60-80% confluence. For the adherent cell lines, once the cells got confluent as observed under microscope, the cells were washed using 1x phosphate buffered saline (PBSA) and trypsinised with 1x trypsin/EDTA for around five minutes at 37°C. The tripsinised cell suspension was then neutralised and re-suspended in 10mL complete media containing serum, and splitted at ratios of 1:5 to 1:20 dilution ratio depending on cell line growth rates for subculture and future experiments.

For the suspension cell lines, cells were harvested and passaged when cell clumps were visible under microscope. After vigorous pipetting in order to disaggregate cell clumps, the cells were centrifuged at 1,000 rpm for five minutes and re-suspended in 10mL medium for downstream experiments.

2.2.4 Cell counting

The cell number was routinely counted by loading 20uL of cell suspension into a disposable cell counting chamber and inserted into the Cellometer Auto T4 cell counter (Nexcelom Biosciences, USA). Images were automatically captured from eight

individual areas of the cell chamber, and counted based on the cellular size and morphology, followed by calculation of the concentration in cells per mL.

2.2.5 Cell storage and retrieval

For cell cryo-preservation in liquid nitrogen, one million cells were washed and centrifuged at 1,000 rpm for five minutes. Supernatant was removed and the cell pellet was re-suspended in 0.5mL freesing buffer (90% FBS and 10% dimethyl sulfoxide (DMSO)) before transferring into sterile cryovials (Corning, USA). Subsequently, the cryovials were transferred into Mr.Frosty Freezing Container (Thermo Fisher Scientific, UK) and put in the -80°C freezer overnight before transferring into liquid nitrogen.

For the cell retrieval, 10 mL complete medium was added to 0.5 mL cell suspension after thawing at 37°C water bath. The cell suspension was then transferred into a flask and put in the incubator over night to allow the attachment of cells. To remove the remaining DMSO, the medium was changed the next day with respective complete medium. All freshly thawed cells were used only after at least one passage.

2.2.6 Mycoplasma contamination testing

Cell culture were routinely tested for Mycoplasma contamination using MycoAlert® Mycoplasma Detection Assay (Lonza, Rockland, Maine, USA) in Bodmer laboratory. In brief, 1mL medium was withdrawn from cells being cultured for at least 72 hours. After centrifuging at 1,200rpm for five minutes to pellet cell debris, 100uL cleared supernatant was transferred into each well of a 96-well white-walled microplate

combined with 100uL MycoAlert Reagent. The supernatant-MycoAlert Reagent mixture was then incubated at room temperature for five minutes, followed by the measurement of background fluorescence (Reading A) using a FLUOstar OPTIMA Luminometer (Offenburg, Germany). Subsequently, 100uL of MycoAlert Substrate was added into each well and incubated for ten minutes, after which the fluorescence was measured again (Reading B).

The ratio of Reading B and Reading A was calculated. Cells were regarded to be negative for mycoplasma contamination if the ratio was less than 1. Cells would be cultured for another week and retested if ratios were between 1 and 2. Cells with the ratio greater than 2 would be considered to be contaminated and discarded.

2.2.7 Notch γ-Secretase Inhibitor Dibenzazepine treatment

Dibenzazepine (DBZ), with ≥95% purity by high-performance liquid chromatography (HPLC), was purchased from Merck Millipore (UK). DMSO was used as a solvent for a final concentration of 300nM for DBZ treatment. For preparation of DBZ treatment, cells were harvested, washed and counted as previously described, and 2,500 – 5,000 cells were seeded into each well of a 96-well plate depending on their growth rates. Medium was removed on the following day, and 200nM DBZ, or DMSO as a treatment control was added - the concentration is determined from van Es, et al., 2005 and previous research in CIL. Cells were cultured for varying amounts of time, (3 days, 5 days, or 7 days) before being fixed for immunostaining.

2.2.8 Transient siRNA transfection

Transient siRNA transfection was performed in vitro using Lipofectamine RNAiMAX (Invitrogen, UK). Pre-designed siRNAs against ATOH1and SPDEF were obtained from Santa Cruz Biotechnology (UK) and Dharmacon (UK), respectively. Three different sequences of the Stealth RNAi siRNA against MUC2 were purchased from Thermo Fisher Scientific (UK). To re-suspend siRNA, tubes containing siRNA were briefly spun down and an appropriate volume of RNase-free water was added to prepare the stock concentration of 10 μM. After pipetting up and down 3-5 times, the solution was placed on an orbital shaker for half an hour at room temperature and the concentration was verified using UV spectrophotometry. All siRNAs were stored at -20 °C and used at a final concentration of 50nM.

3uL siRNA (10uM) was added into 150uL Opti-MEM medium, and mixed thoroughly with 9uL Lipofectamine RNAiMAX reagent diluted in 150uL Opti-MEM medium. The mixture was incubated for five minutes at room temperature. A 10uL mixture for 96-well plates and 250uL mixture for 6-well plates was transferred into an empty well. Subsequently, 3 x 10^5 cells (6-well plate) or 1 x 10^4 cells (96-well plate) in antibiotics-free medium with 10% FBS were added to each well and gently mixed. The antibiotic-free medium was replaced by complete medium the next day.

2.3 Three-Dimensional Cell Culture

Matrigel (Becton Dickinson, UK) was thawed on ice and diluted 1:1 with ice-cold DMEM complete medium. 40uL DMEM/Matrigel mixture was then gently loaded into each well of a 96-well plate or a Nunc Lab Tek Chamber Slide (Sigma-Aldrich, UK) before incubating at 37°C for at least 30 minutes to solidify the Matrigel as a base for three dimensional cell culture.

Single cell suspension was attained by filtration through 20um filters (Celltrics, Partec GmbH). 500 single cells were re-suspended in 20uL freshly prepared DMEM/Matrigel mixture, and added onto the solidified Matrigel base. The cells in Matrigel were then incubated at 37°C for another 30 minutes and culture medium was then added to the wells. Cells were grown at 37°C for 12-14 days to allow colony formation with the medium changed twice weekly.

Once colonies were formed, the medium was removed and 100uL 4% (v/v) paraformaldehyde in phosphate buffered saline (PBS) was added and incubated at room temperature for 15mins to fix the cells. The fixed cells were then washed three times with 100uL PBS. Subsequently, 100uL 0.1% Triton-X was added and incubated at room temperature for 10 minutes for permeabilisation. After washing twice with wash buffer (PBS with 2% FBS), the colonies were stained with primary and secondary antibodies along with 100µl of 1:1000 diluted TRITC-Phalloidin and incubated at 4°C

in the dark overnight. After another washing step, the colonies were stained with 100µl of 10 µg/ml 4'6- diamidino-2-phenylindole (DAPI, Sigma) at 4°C in the dark for 10 minutes before being imaged under a Zeiss LSM510 laser scanning confocal microscope (Carl Ziess Ltd, UK).

2.4 Fluorescent Activated Cell Sorting (FACS)

2.4.1 Antibody labelling

For the surface antigen staining, cells were trypsinised, washed and filtered through 20um filter to obtain single cell suspension, which was then aliquoted into FACS tubes containing a million cells each. Cells were stained with an appropriate dilution of primary antibody for 30 minutes on ice, followed by washing with PBS and centrifuging at 1,200rpm centrifugation for 7 minutes. Cells were then re-probed with secondary antibodies at appropriate dilution ratios to target primary antibodies or serve as an isotype control, and incubated for 30 minutes in the dark on ice. After another washing step, cells were then re-suspended in 500uL wash buffer with DAPI or Propidium iodide (PI) or DRAQ7 (Abcam, UK) for live/dead staining and were ready for flow cytometry analysis.

For intracellular staining, cells were fixed and permeabilised by adding ice-cold methanol dropwise into the cell suspension while vortexing thoroughly and incubating on ice in the dark for one hour. Subsequent antibody labelling was performed in the same way as for the surface staining.

2.4.2 Flow cytometry analysis and sorting

The CyAn ADP analyser (Beckman Coulter, UK) was used to assess the protein expression profiles under the data collection software Summit 4.3 (Cytomation) or FlowJo V10 (FlowJo, LLC). Unstained samples were used to optimise the voltage setting, and an isotype control was used to exclude the non-specific binding of antibodies to Fc receptors. In the analysis, doublets and debris were gated out from the analysis using pulse width at a forward scatter gain and side scatter gain of 3.6 and 1.0 at 450 Volts respectively. For fluorescence detection, FITC and Alex Fluo 488 were detected under the 488-nm laser gain of 1.0 at 400 Volts, while APC was detected under the 633-nm red laser gain of 1.0 at 775 Volts. Fluorescence logarithms and histograms were displayed to distinguish the positively staining populations.

The BD FACSAria III (DB Biosciences) and SH800Z Cell Sorter (SONY Biotechnology Inc.) were used for live cell sorting. After proper staining and filtration through a 20um filter as described before, cells were sorted using the integrated Summit software. Single cells were obtained by gating out the debris and cellular aggregates using pulse width, and dead cells were excluded based on the staining with viability reagents, including DAPI, PI and DRAQ7. The top 5% positively or negatively stained cells were sorted and collected into 1mL complete medium for subsequent experiments.

2.5 Immunofluorescent staining

Cells were plated in 96-well plate at density of 5 x 10^3 (or 2.5 x 10^3 for the fast-growing cell lines) cells per well. Cells were allowed to adhere by incubating overnight at 37°C, and incubated for an additional five days. Medium was changed on the third day.

For immunofluorescent staining, the medium in each well was discarded. Adhered cells were washed with 200μL PBS and fixed with 4% paraformaldehyde for 15 minutes at room temperature. After three washes with PBS, the fixed cells were permeabilised with 0.1% Triton Buffer for 10 minutes in the dark at room temperature, and then washed with PBS once. In order to avoid non-specific binding, the cells were blocked with 200μL wash buffer and incubated at 4°C for at least one hour. Following the blocking step, the combined staining reagents of lectins Wisteria Floribunda (WFA) or Dolichos Biflorus Agglutinin (DBA) and different MUC2 antibodies were diluted, added to each well and incubated in the dark at 4°C for at least one hour. The cells in each well were then washed with 200μL wash buffer and then incubated with corresponding secondary antibodies at a dilution of 1:200 (at the final concentration of 10μg/mL) in the dark at 4°C for one hour. After another wash with 200μL wash buffer, cells were stained with 50μL 1:5000 diluted DAPI solution and incubated in the dark at 4°C for 10 minutes. The plate was then observed under the Zeiss Observer z1, Carl Zeiss (Oncology Microscopes, WIMM, University of Oxford) and images were

captured using the Microscope Software ZEN (Zeiss, UK). Images were exported and further quantified and analysed using Fiji (ImageJ, UK).

Alexa Fluor™ 488 Antibody Labelling Kit (Thermo Fisher) was used to conjugate primary antibodies. The dilutions and sources of primary antibodies used in immunostaining and flow cytometry are listed in **Table 2.1**.

Antibody Name/Target	Primary Ab type, clone and supplier	Primary Ab Dilution	2nd antibody and dilution
PR5D5	Mono, N/A, In-house	1:200	1:200 Alexa-fluor 488 and 555; Goat anti-mouse APC
MUC2-D	Mono, M7313, Dako	1:200	1:200 Alexa-fluor 488 and 555
MUC2-N	Mono, 996/1, Invitrogen	1:200	1:200 Alexa-fluor 488 and 555
TFF3	Mono, ab109104, abcam	1:200	1:200 Alexa-fluor 488 and 555
SPDEF	Poly, sc-166846, Santa Cruz	1:100	1:200 Alexa-fluor 488 and 555
ATOH1	Poly, 21215-1-AP, Proteintech	1:500	1:200 Alexa-fluor 488 and 555
CA12-M	Mono, ab54917, abcam	1:200	1:200 Alexa-fluor 488 and 555
CA12-R	Poly, HPA008773, Sigma	1:50	1:200 Alexa-fluor 488 and 555

Table 2.1 Dilutions of antibodies used in the immunostaining and flow cytometry

2.6 RNA methods

2.6.1 RNA extraction

RNeasy Mini Kit (Qiagen, UK) was used to extract total RNA following the manufacturer's protocol. In brief, the pellet of cells was lysed in Buffer RLT (Qiagen) to disrupt the plasma membranes of cells and organelles to release the total RNA. The lysates were homogenised to decrease their viscosity by transferring them into a QIAshredder spin column with a collection tube, centrifuging for 2 minutes at full speed. The same volume of 70% ethanol was then added to optimise the conditions for RNA binding to the membrane of the RNeasy Mini spin column. The mixture was then loaded onto a RNeasy® mini column with a 2mL collection tube. Total RNA that binds efficiently to the column membrane and contaminants, were washed away with wash buffer RPE (Qiagen). The RNA was then eluted in RNase-free water and stored at -80°C for further analysis.

2.6.2 Microarray gene expression analysis

Bodmer's lab has generated the microarray mRNA expression data of all corresponding colorectal cancer cell lines. The cell line microarray analysis has been conducted by previous researchers in the Bodmer's lab over the past decade. Briefly, microarray mRNA expression data were acquired from the Paterson Institute for Cancer Research (Wong et al 2006). 10 μg RNA extracted from each colorectal cancer cell line were sent to the Molecular Biology Core Facility (Paterson Institute for Cancer Research) for

Affymetrix U133 plus 2 microarrays. Briefly, mRNA from each sample was reverse transcriptionally converted into the double-strand cDNA, which was then labelled with biotins. 10ug fragmented cDNA were hybridised to the probes (more than 54,612 transcripts) on the Affymetrix U133 Plus 2.0 oligonucleotide arrays. The microarray expression data was then normalised using Robust Multichip Analysis algorithm at Molecular Biology Core Facility, Paterson Institute for Cancer Research.

The starting point of microarray gene expression analysis in this thesis is a gene list from the analysis described above. Using this list, the further differential expression analysis using the cell line microarrays in this thesis was performed under a default setting in Partek® Genomics Suite, essentially using the student t-test to determine the significance between the mean levels of expression of each probe set between the goblet cell-positive cell lines in contrast to the goblet cell-negative cell lines. The genes with a 5-fold change or greater and p-value of <0.01 were considered as significantly differentially expressed. Those genes were further analysed by reviewing online gene Human Protein Atlas and PubMed databases and comparing expression profiles in other cell lines.

2.6.3 RNA Extraction from the fixed goblet cells

After trypsinisation and a single washing step with 10mL PBS, a million LS180 cells were fixed with 1mL freshly prepared 4% PFA and incubated on ice for 15 minutes. After spinning and removal of supernatant PFA, the cell pellet was washed once with

1mL RNase-free wash buffer (1x PBS, 1% RNase-free BSA and 0.01% RNasin Plus RNase inhibitor) before being treated with permeabilisation buffer (0.1% saponin with 1% RNase-free BSA (Gemini, UK) and 0.005% RNasin Plus RNase inhibitor) on ice for 10 minutes. Subsequently, cells were stained with 1:200 PR5D5 and 1:200 anti-mouse APC secondary antibody diluted in RNase-free wash buffer.

One-hundred PR5D5-positive or -negative cells were sorted into each of the strip tubes containing protein lysis buffer (5uL PKD buffer, 1:4 Proteinase K and ERCC RNA Spike-In Mix). The low cell number in each tube greatly reduced the sorting time to minimise RNA degradation. Seven tubes of PR5D5-positive cells and seven tubes of PR5D5-negative cells were prepared for RNA isolation. The miRNeasy FFPE Kit was used to lyse cells and purify RNA.

2.6.4 Smart-seq2 RNA sequencing

The sequencing libraries were prepared as described in Picelli et al., 2013 and Ramsköld et al., 2012. In brief, the oligo-dT primers were used to reverse transcribe the polyA-positive total mRNA, which was then extended with several non-template nucleotides by the terminal transferase activity of Monoley Murine Leukaemia Virus Reverse Transcriptase. These extended nucleotides were base-paired with the SMARTer II A oligos (ClonTech, UK), which were further reverse transcribed and formed a full length cDNA. It should be noted that the directional cDNA analysis is indicated to give rise to the 3' end bias (Head, et al., 2014; Harbers, 2008; Conesa, et

al., 2016). The cDNA libraries were quantified using Bioanalyser 2100 (Agilent Genomics, UK). The cDNA libraries were pre-amplified by using KAPA HotStart HIFI 2× ReadyMix (KAPA Biosystems, UK) and incubating PCR master at 95 °C for 1min, followed by 15 cycles (95 °C 15 s, 65 °C 30 s, 68 °C 6 min) and an extension at 72 °C for 10min. The directional cDNA analysis is indicated to give rise to the 3' end bias

The cDNA library was sent to the Single Cell Facility and the Next Generation Seq Facility at the WIMM, where 1ng amplified cDNA library of each sample was used for tagmentation reaction (25uL 2x Tagment DNA Buffer and 5uL Tagment DNA Enzyme) and incubated at 55 °C for 5 mins. After purification using DNA Clean & Concentrator-5 kit (Zymo Research) from the Illumina Nextera XT library preparation platform, the sample was sent to the Next Generation Seq Facility (WIMM) for Illumina HiSeq 2000. 5-20 million mapped 75-base paired-end reads were used to quantify the medium to significantly expressed genes in the 14 RNA-seq samples of the highly specified goblet cells (Conesa, et al., 2016).

2.7 Statistical methods

2.7.1 General data analysis

GraphPad Prism Software (La Jolla, California, USA) was used for data analysis. All data are shown as mean ± SEM. Chi-square test and Fisher's exact test were used for 2x2 contingency tables.

2.7.2 RNA sequencing analysis

The quality of RNA-seq raw data (in 'fastq' format) was examined using FastQC (Andrews et al., 2010). After quality control assessment, the fastq files were aligned to GRCh37 (hg19) using the package Spliced Transcripts Alignment to a Reference (STAR) (Dobin et al., 2013) from the server of Computational Biology Research Group (CBRG). The assigned reads were counted by featureCounts. The lowly expressed genes were filtered out and only the genes whose count per million (CPM) values were more than 5 in at least 7 samples were used for further analysis. The sequencing depth and RNA composition were normalised using the *calNormFactors* function of the edgeR package (Robinson et al., 2010). A multi-dimensional scale analysis was performed using CPM values to evaluate the variance between samples. The *topTags* function of the edgeR package was then used to identify the differentially expressed genes and apply the Benjamini-Hochberg correction (Robinson et al., 2011). Generally, the Bonferroni test is considered to correct all input p-values equally while Benjamini-Hochberg test corrects p-values based on their p-value rankings. Thus, in multiple correction on more than 20,000 human protein coding genes in this case, Bonferroni test is more likely to produce false negatives and discard potential significant differentially expressed genes, while Benjamini-Hochberg test is believed to have better control of FDR (McDonald, 2009; Hair, 1984). Because the general aim of RNA-sequencing analysis in this thesis is exploratory to identify the differentially expressed genes in goblet cells, the default Benjamini-Hochberg multiple hypothesis test in the

topTags function of edgeR was used rather than Bonferroni (McDonald, 2009; Noble, 2009).

CHAPTER 3

SCREENING OF GOBLET CELL DIFFERENTIATION IN A PANEL OF 64 HUMAN COLORECTAL CANCER CELL LINES

3.1 Introduction

In human colorectal cancers, the expression and glycosylation of MUC2 are frequently dysregulated. Specifically, mucinous carcinomas express an abundant amount of MUC2, while in non-mucinous carcinomas, goblet cell differentiation and MUC2 expression are largely reduced (Kim et al., 2010). In the mucinous carcinomas, the altered genetic and epigenetic regulation of MUC2, e.g. the hypomethylation of MUC2 core promoter (Okudaira et al., 2010) or up-regulated expression of secretory lineage-specific transcriptional factor ATOH1 (Leow et al., 2004; Park et al., 2006), leads to the increased mucin production. In non-mucinous carcinomas, deficient MUC2 expression, aberrant crypt morphology and altered cellular migration are observed (Velcich et al., 2002). The truncated glycan structures and the high expression of neo-glycans, e.g. sialylLeA and sialylLeX antigens, can be seen in both mucinous and non-mucinous carcinomas (Hollingsworth et al., 2004; Andrianifahanana et al., 2006; Kim et al., 2008; Byrd et al., 2004). In addition, it has been suggested that high levels of goblet cell differentiation were observed in colorectal cancer cell lines with higher metastasis capacity (Byrd et al., 2004).

Despite the important functions of goblet cells and MUC2, there is no surface marker available for goblet cells, and the regulatory mechanisms of goblet cell differentiation involved in colorectal cancers have been poorly elucidated. This is therefore a barrier

in the understanding of the genetic nature of goblet cells and how they become dysregulated in colorectal carcinogenesis.

To address this problem, **Chapter 3** characterises goblet cell differentiation in colorectal cancers by using a large panel of human colorectal cancer cell lines and their microarray expression data. The results presented in this chapter first validate MUC2 as the molecular target of PR5D5, an in-house antibody targeting goblet cells, which allows screening for goblet cells in a panel of 64 colorectal cancer cell lines at protein level by using PR5D5 and another commercial MUC2 antibody (Dako, UK). The screening results were then compared with mRNA expression of MUC2 from microarray data, which led to the categorisation of goblet cell differentiation in human colorectal cancer cell lines. Based on the goblet cell categorisation, microarray analysis has identified the genes that are differentially expressed in the goblet cell-positive cell lines, including CA12, a cellular surface protein, which will be further discussed in **Chapter 5**.

3.2 Results

3.2.1 PR5D5, an in-house antibody specifically targeting goblet cells

PR5D5 was raised against the fresh human colorectal mucosa (Richman and Bodmer, 1987). It was used throughout this thesis due to its high specificity and affinity against goblet cells, and a large amount of stock. Although the monoclonal antibody PR5D5 has been reported in several publications (Albaugh et al., 1992; Campbell et al., 1994; Nair et al., 2003; Ashley et al., 2013; Yeung, Patent WO 2015/198065 A1), its molecular target has not yet been clearly characterised. Therefore, the first aim was to clarify the target of PR5D5.

3.2.1.1 PR5D5 staining in the normal colon and colorectal cancer cell line

The staining pattern of PR5D5 was demonstrated via immunofluorescent staining in normal human colonic tissue FFPE slides and the colorectal cancer cell line RW7213 (**Figure 3.1**). Along the epithelial layer of the normal colonic crypt, goblet cells were strongly labelled with fluorescent tagged PR5D5 (data from Dr Neil Ashley) (**Figure 3.1 A**). In addition, RW7213 expresses a high level of MUC2 identified in microarray data. Single cells from RW7213 were seeded in Matrigel for 3D culture, which was reported as a valid *in vitro* model to characterise colorectal cancer stem cell differentiation (Neil Ashley et al., 2013). Lumens were identified by Phalloidin staining, and goblet cell-specific mucus was strongly labelled in the granules within goblet cells or secreted towards the lumen (**Figure 3.1 B**). These staining patterns were consistent

with previously published data (Albaugh et al., 1992; Campbell et al., 1994; Ashley et al., 2013; Yeung, Patent WO 2015/198065 A1), and confirmed that PR5D5 monoclonal antibody specifically targets a protein expressed specifically in goblet cells.

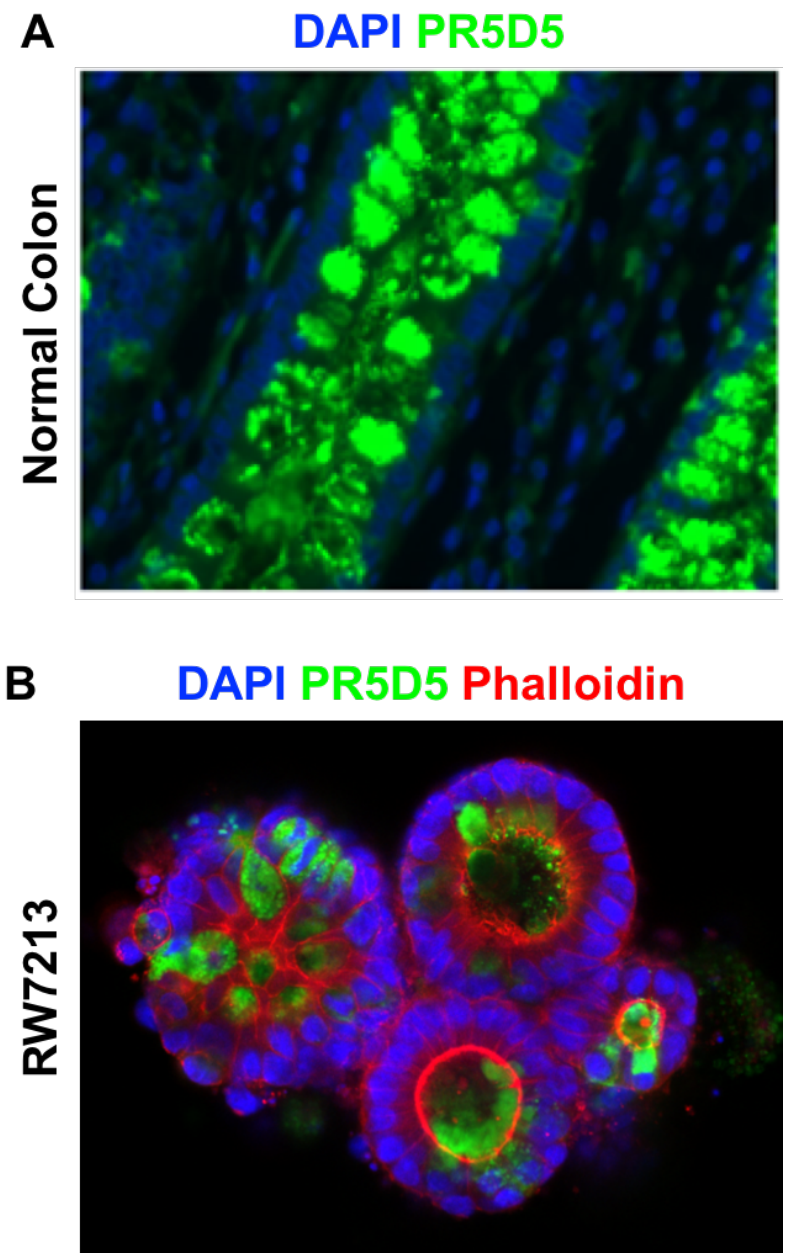


Figure 3.1 PR5D5 staining in goblet cells of the human normal colonic crypt and the colorectal cancer cell line RW7213

(**A**) Immunofluorescent staining of cryostat sections of normal colon tissues with PR5D5 (green) for goblet cells and DAPI (blue) for cell nucleus. (**B**) Single cell suspensions of human colorectal cancer cell line RW7213 were embedded in Matrigel and grown for 14 days before being fixed and labelled with DAPI (blue, 1:5000), PR5D5 (green, 1:200) and Phalloidin (red, 1:10000) for F-actin staining.

3.2.1.2 Validation of PR5D5 targeted protein by co-staining with MUC2 antibodies

It has previously been indicated by Western Blot that PR5D5 recognises a protein of approximately 700 kilo Daltons (Campbell et al., 1994), which is the similar molecular weight of the glycosylated MUC2 protein. Together with the PR5D5 staining in normal crypts and cancerous lumens (**Figure 3.1**), it is therefore reasonable to hypothesise that PR5D5 targets MUC2, the goblet cell-specific mucin.

The molecular target of PR5D5 monoclonal antibody was further supported by co-staining with two commercially available MUC2 antibodies, namely MUC2-D (Dako, UK) and MUC2-N (Neomarker, UK). From the datasheets, MUC2-D recognises a 29-amino acid synthesised peptide which contains one unit of the 23-amino acid tandem repeat (PTTTPITTTTTVTPTPTPTGTQT) and another four amino acids in the next repeat, while MUC2-N targets the recombinant protein of five tandem repeats of MUC2. The MUC2-D and MUC2-N antibodies were conjugated using Alexa Fluor™ 488 Antibody Labelling Kit (Thermo Fisher) before co-staining with PR5D5. As shown in **Figure 3.2A**, immunofluorescent co-staining with PR5D5 monoclonal antibody largely overlapped with the staining using MUC2-D antibody in the colorectal cancer cell line LS180. Similarly, MUC2-N and PR5D5 co-labelled the same goblet cells in LS180 (**Figure. 3.2B**). This confirmed the PR5D5 specificity against goblet cells, and indicated MUC2 as its molecular target.

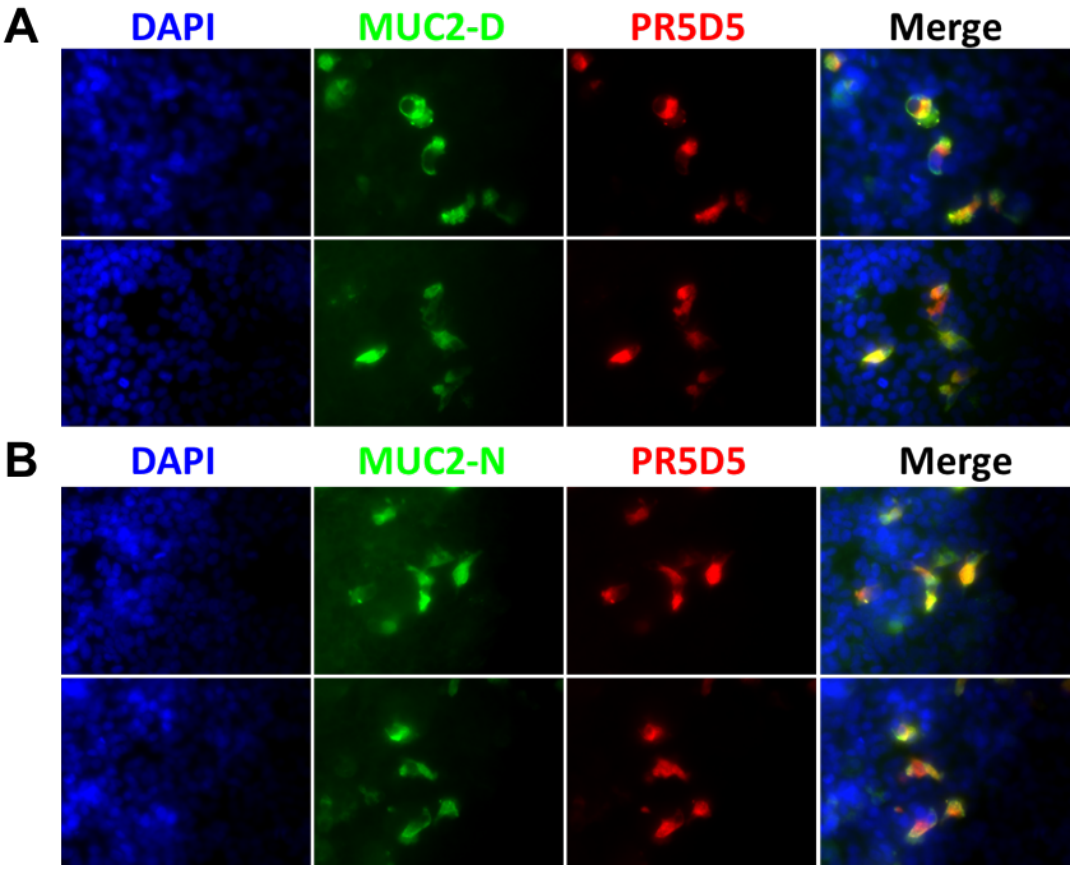


Figure 3.2 Co-staining of PR5D5 with MUC2-D and MUC2-N
5,000 cells of human colorectal cancer cell line LS180 were seeded into each well of 96-well plates, and grown for three days before fixation and intracellular co-staining with PR5D5 (red, 1:200) and (**A**) MUC2-D (green, 1:200) or (**B**) MUC2-N (green, 1:100). DAPI (blue) was used for nucleus staining.

3.2.1.3 Knock-down of MUC2 decreases PR5D5 staining on goblet cells

MUC2 is considered to be the dominant component of colonic mucus and co-stains strongly with PR5D5. Knock down of MUC2 was therefore performed to evaluate its effect on PR5D5 staining in colorectal goblet cells. **Figure 3.3** illustrates the PR5D5 and MUC2 staining after transient silencing of MUC2 in the cell line LS180 using the MUC2 Stealth RNAi^{TM} siRNA (Cat. 1299001, Thermo Fisher Scientifics). With three independent replicates, **Figure 3.3A** demonstrates the representative FACS staining patterns of PR5D5 after the LS180 cells were treated with varying concentrations of siMUC2_1. When the cells were not treated with siRNA, the PR5D5-positive cell proportion is 7.25% compared to the isotype control. PR5D5 staining decreased in response to the siRNA treatment against MUC2 in a dose-dependent manner. The PR5D5-positive proportion reduced to 3.05% at the siMUC2_1 amount of 1.25pmol, and remained at similar level (2.99%) at the siMUC2_1 amount of 2.5pmol. At 5pmol siMUC2_1 per well, FACS analysis showed that PR5D5-positive cells of LS180 decreased to less than 1% of the analysed sample.

As shown in **Figure 3.3B**, using different siRNA sequences against the MUC2 gene, PR5D5 and MUC2-D staining decreased in a similar dose-dependent manner, while the positive cell proportion remained comparable to the control when using scrambled siRNA. The qRT-PCR analysis could have been conducted to further confirm the decreased MUC2 expression under siRNA-mediated knock-down if time had allowed.

This observation was further strengthened by the immunofluorescent staining under siRNA knock down of MUC2 as shown in **Figure. 3.3 C**. The proportion of cells labelled with PR5D5 and MUC2-D decreased co-ordinately with increasing concentration of siRNA, while both antibodies targeted the same goblet cells.

These results provided strong evidence that PR5D5 staining was inhibited when the MUC2 expression was silenced. This suggests that MUC2 protein is the target of the PR5D5 monoclonal antibody.

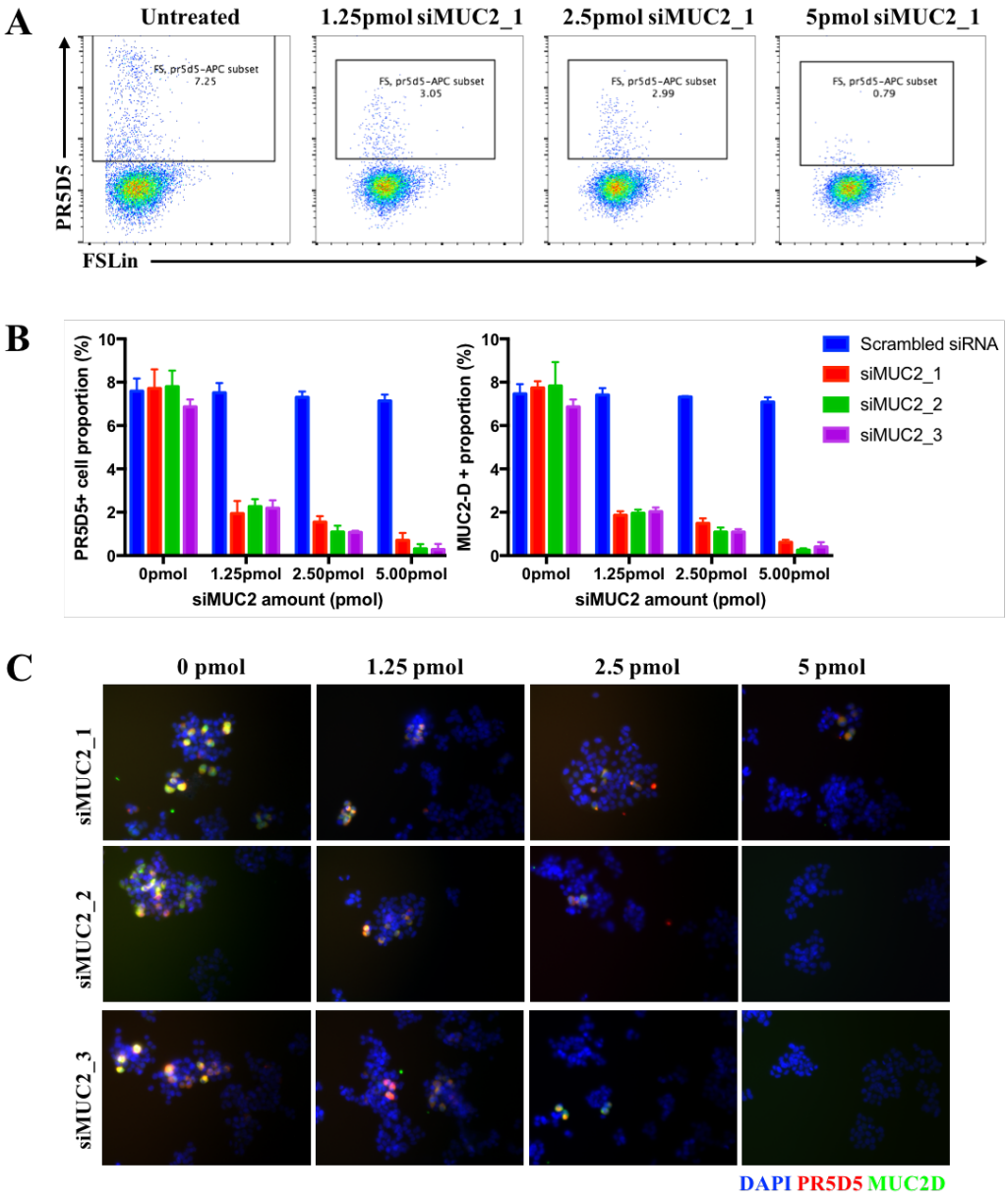


Figure 3.3 PR5D5 staining reduced when knock down *MUC2*

30,000 cells of human colorectal cancer cell line LS180 were seeded into each well of 6-well plates, and treated with scrambled siRNA (5 pmol) as control and varying amounts of three different siRNAs against MUC2 (0, 1.25, 2.5 and 5 pmol) for 24 hours. Culturing media were then changed and cells were allowed to grow for 48 hours before fixation and intracellular staining with PR5D5 and MUC2-D for FACS analysis. (**A**) Representative PR5D5 FACS staining under siMUC2_1 treatment. (**B**) Decreased FACS staining with PR5D5 and MUC2D under siRNA-mediated MUC2 knock down were summarised (Error bar: mean±SD, n=3). (**C**) 5,000 cells of human colorectal cancer cell line LS180 were seeded into each well of 96-well plates and treated with siRNA in the same way as described above, followed by fixation and co-staining with PR5D5 (red, 1:200) and MUC2-D (green, 1:200).

3.2.1.4 Competitive binding assay using PR5D5 decreases MUC2D staining

The target specificity of PR5D5 antibody was further confirmed by a competitive binding assay against MUC2-D. PR5D5 at varying dilution ratios (1:200, 1:50, 1:12.5, 1:3.125, 1:1) was incubated with the colorectal cancer cell line LS180 overnight at 4 °C before adding the MUC2-D (1:200 dilution) that was conjugated using the Alexa Fluor™ 488 Antibody Labelling Kit (Thermo Fisher). MUC2-D stained the mucus within goblet cells when cells were incubated with a 1:200 dilution of PR5D5. The binding of MUC2-D to its antigen, the tandem repeats of MUC2 protein, significantly decreased with an increasing concentration of PR5D5 (**Figure 3.4**). The staining with MUC2-D was almost undetectable when fixed LS180 cells were incubated with 1:1 diluted PR5D5. These results indicated the binding of PR5D5 antibody overlapped or interfered with the MUC2-D binding site.

Taken together, these lines of evidence show that the MUC2 protein is the target of the PR5D5 antibody, confirming the specificity of PR5D5 on goblet cells.

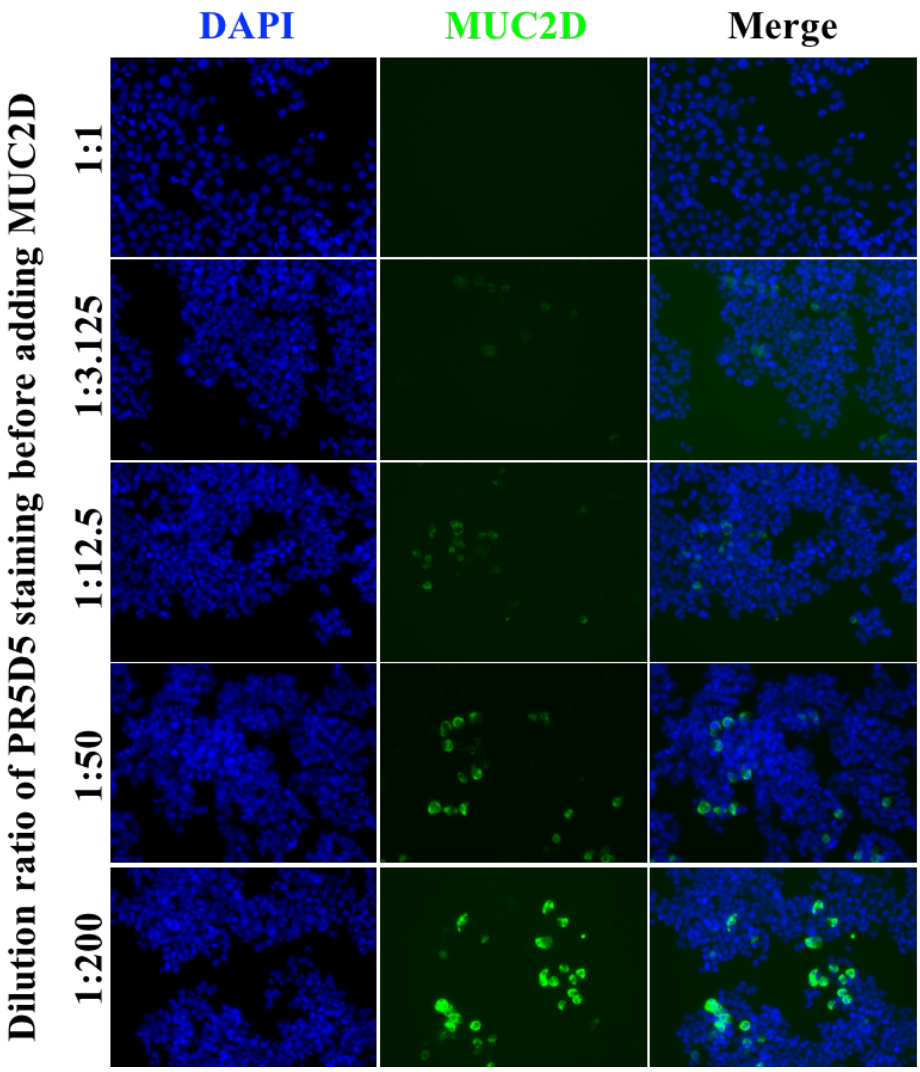


Figure 3.4 Competitive binding between PR5D5 and MUC2-D
5000 cells of LS180 were seeded into each well of 96-well plates and grown for three days before fixation and permeabilisation. Fixed cells were incubated with varying concentration of PR5D5 antibody at 4°C overnight. Cells were then incubated with the MUC2D antibody that was conjugated with Alexa Fluor 488 at 4°C overnight.

3.2.2 Screening of MUC2 expression in 64 colorectal cancer cell lines

The Cancer and Immunogenetics Laboratory has access to 142 human colorectal cancer cell lines and their microarray expression profiles (unpublished data). This has provided valuable resources to examine goblet cell differentiation and determine the differentially expressed genes in the goblet cell differentiating cell lines.

3.2.2.1 mRNA expression levels of MUC2 from microarray analysis

MUC2 mRNA expression was examined in a panel of 142 human colorectal cancer cell lines. **Figure 3.5** shows the microarray expression intensity of MUC2 ranked from low to high.

As shown in **Figure 3.5A**, five out of 64 cell lines (3.52%): SNU1746, CL40, RW7213, NCIH508 and NCIH498, showed significantly high MUC2 expression with more than 900 microarray expression arbitrary units (AU) (separated by the green line in **Figure 3.5A**). These five cell lines, labelled with '***' in **Table 3.1**, represented a novel sub-group of colorectal cancers with excessive MUC2 production and may potentially be signet ring carcinomas cell lines or mucinous carcinomas cell lines. In fact, SNU1746 is known to be derived from signet ring carcinomas (Ku et al., 2010). Besides these five cell lines, the MUC2 expression level in the remaining cell lines is smaller than 700AU.

Figure 3.5B is plotted on a different scale to show the pattern of MUC2 expression at lower levels and omitting the highly expressing cell lines. An obvious gap can be observed at the MUC2 expression level between 700AU and 900AU. As shown in **Figure 3.5B**, 19 out of 142 cell lines (13.38%), labelled with '**' in **Table 3.1**, showed MUC2 expression levels between 200AU and 700AU. These cell lines, together with the previous five, were considered as the cell lines with MUC2 positivity at the mRNA level. A difference in the MUC2 expression increasing rates, as separated by the blue line, can be observed at the level of 200AU in **Figure 3.5B**. There are 51 out of 142 cell lines (35.92%) with MUC2 expression levels between 100AU to 200AU. These cell lines, labelled with '*' in **Table 3.1**, were considered as the cell lines with intermediate MUC2 mRNA expression. The remaining 67 cell lines with MUC2 expression levels below 100AU (47.18%), as shown by the red line in **Figure 3.5B,** were considered as low MUC2 expressing cell lines and labelled with '-' in **Table 3.1**. The classification based on the mRNA expression of MUC2 was further confirmed through a screening of goblet cell differentiation at the protein level.

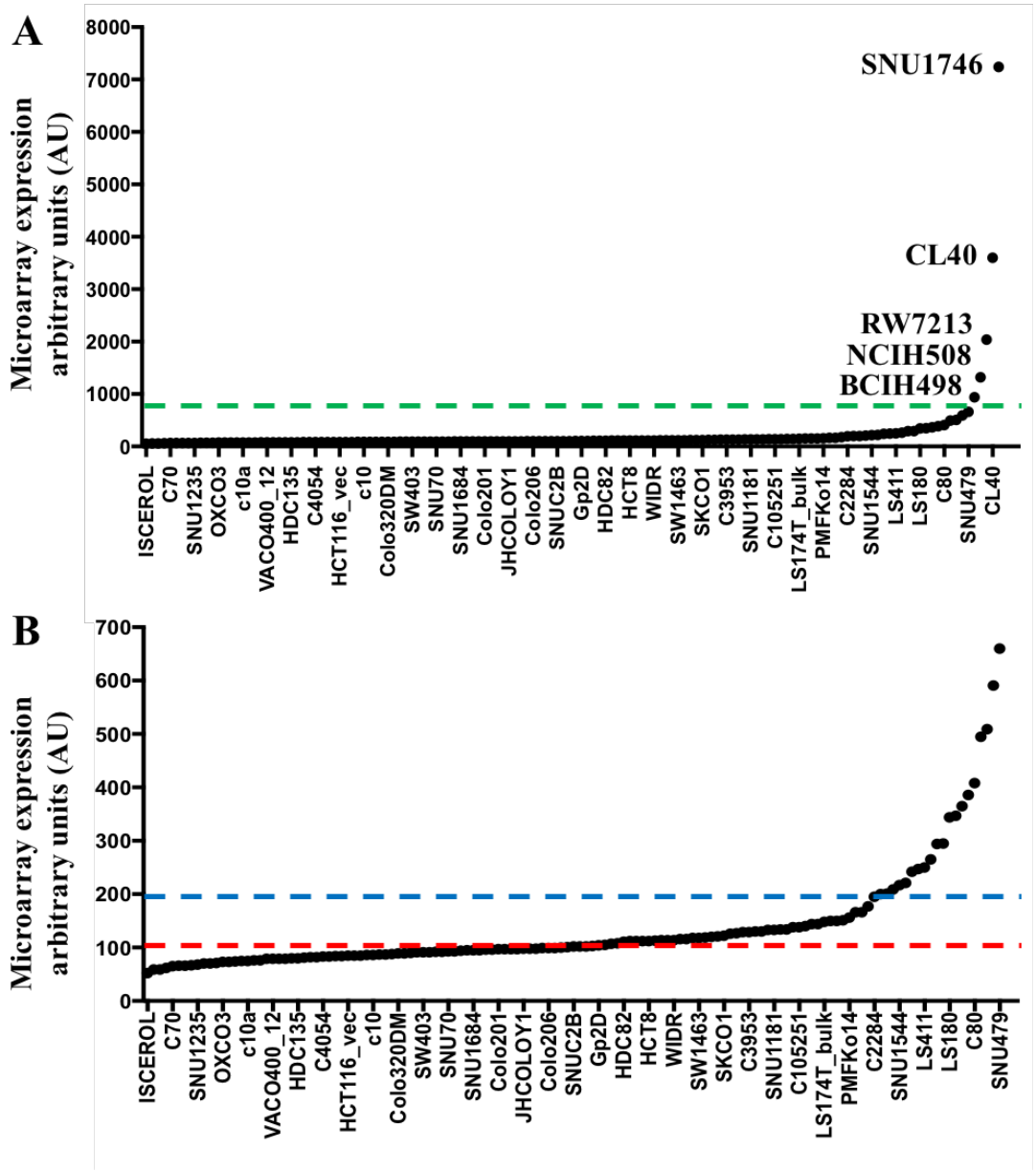


Figure 3.5 MUC2 microarray expression across 142 human colorectal cancer cell lines

MUC2 mRNA expression were measured by Affymetrix Human Genome U133 plus 2.0 microarray. (**A**) Among the whole panel of 142 cell lines, MUC2 is highly expressed in SNU1746, CL40, RW7213, NCIH508 and NCIH498 with more than 900AU. These cell lines are classified as MUC2 positive at the message level. (**B**) A difference in the increase rate of MUC2 expression can be observed at the level of 200AU in the remaining 137 cell lines. The cell lines with MUC2 expression level more than 200AU are classified as MUC2 positive at message level. The cell lines with MUC2 expression of 100-200AU were considered as intermediate. The cell lines with MUC2 expression smaller than 100AU were considered as MUC2 low at mRNA level.

3.2.2.2 PR5D5 and MUC2 antibody staining in colorectal cancer cell lines

In order to systematically screen goblet cell differentiation and MUC2 expression at protein level, two MUC2 antibodies, PR5D5 and MUC2-D, were used in immunofluorescent staining in a panel of 64 colorectal cancer cell lines.

Figure 3.6 shows representative staining of cell lines with positive, intermediate and negative reactivity with PR5D5 and MUC2-D. RW7213 and LS180 were categorised as goblet cell positive cell lines, with a subset of cells that show strong staining. The positively stained areas were largely polarised at one end of the goblet cells. JHCOLOY1 and C84 showed a small number of cells with some reactivity against PR5D5 or MUC2-D antibodies, and were characterised as goblet cell intermediate cell lines. RKO was characterised as a goblet cell negative cell line due to undetectable positive staining with both antibodies.

For PR5D5 immunofluorescent staining, 17 out of 64 cell lines (26.56%), namely C125PM, CL40, CX1, HCA46, HDC114, HDC57, HDC73, HT29, KM2012, LOVO, LS174T, LS180, NCIH508, RW7213, SW1222, VACO10MS and WIDR (labelled with '**' in **Table 3.1**), showed a proportion of greater than 5% positively stained cells. Another five cell lines (7.81%) (labelled with '*' in **Table 3.1**), in term of C80, C84, HDC111, JHCOLOY1 and LIM1863, showed various positive proportion of <5% PR5D5 staining. The remaining 42 colorectal cancer cell lines (65.63%) (labelled with

'-' in **Table 3.1**) did not show PR5D5 positive staining. The representative PR5D5 staining of all 64 cell lines is summarised in **Figure Appendix.1**.

The MUC2-D antibody was characterised to target the unglycosylated form of MUC2 with the subcellular distribution near nucleus from previous study (Dr Trevor Yeung, unpublished data). MUC2-D demonstrated a staining pattern highly correlating to PR5D5 staining. 18 out of 64 cell lines (28.13%) (labelled with '**' in **Table 3.1**) showed a strongly positive MUC2-D staining, including the cell line HDC111 that was classified as intermediate in PR5D5 staining. This is possibly due to the unglycosylated form of MUC2 that can be detsected by MUC2-D but not PR5D5. The five cell lines (7.81%), C80, C84, JHCOLOY1, LIM1863 and RW2982 showed intermediate staining reactivity with MUC2-D antibody. They were labelled with '*' in **Table 3.1**. The remaining 41 cell lines (64.06%) (labelled with '-' in **Table 3.1**) showed little immunochemical reactivity with MUC2-D antibody. The representative MUC2-D staining of all 64 cell lines was summarised in **Figure Appendix.2**.

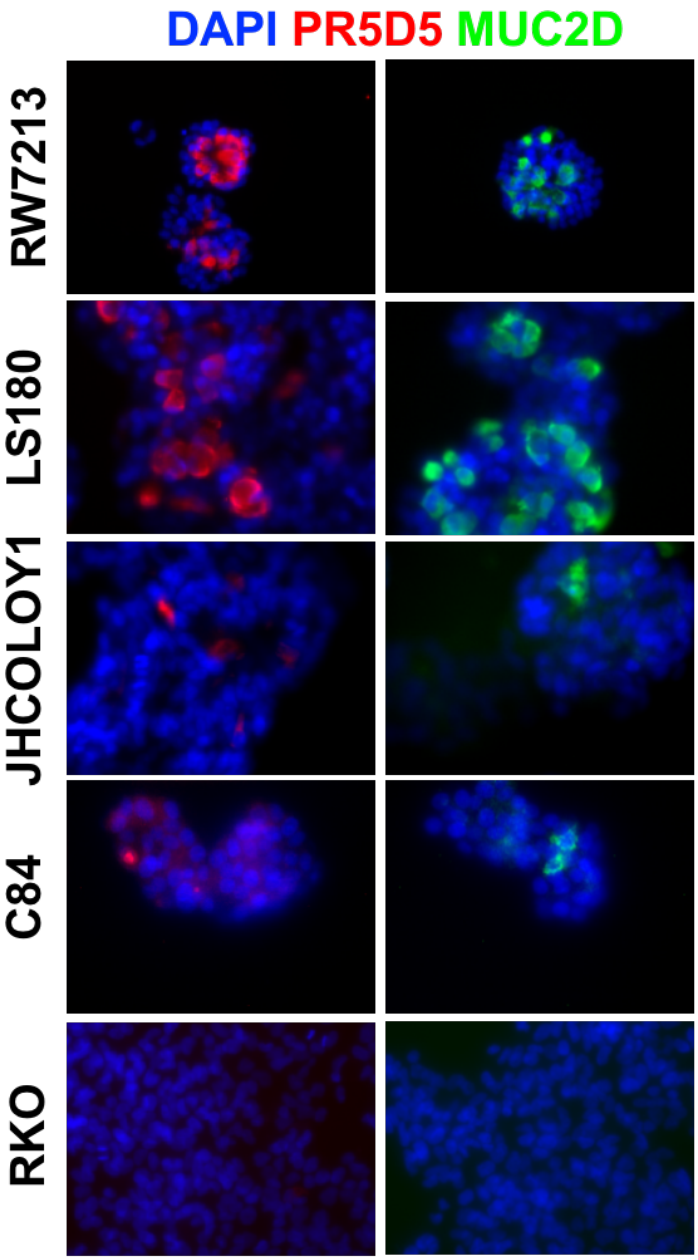


Figure 3.6 Representative PR5D5 and MUC2-D staining in goblet cell-positive, -intermediate and –negative cell lines.
5000 cells of each cell line were seeded into 96-well plates and grown for three days before intracellular staining using PR5D5 (1:200, left column) or MUC2-D (1:200, right column) antibodies. RW7213 and LS180, as representative goblet cell-positive cell lines, showed a subset of cells with clear strong reactivity against PR5D5 and MUC2-D. JHCOLOY1 and C84, with occasional staining, represent goblet cell-intermediate cell lines. RKO, as an example of goblet cell-negative cell lines, showed no reactivity with PR5D5 or MUC2-D.

Goblet cell differentiation was characterised at mRNA and protein levels in 64 human colorectal cancer cell lines, and the microarray expression and immunostaining data were cross-checked and summarised in **Table 3.1**. If both PR5D5 and MUC2-D staining are high ('**') and microarray expression is not negative (not '-'), the cell lines were categorised as 'high goblet cell differentiation (+)'. If all three PR5D5, MUC-2 staining and microarray expression are negative ('-'), the cell lines will be categorised as 'low goblet cell differentiation (-)'. The remaining cell lines were categorised as 'intermediate goblet cell differentiation (+/-)'.

Table 3.1 (the following page) Categorisation of goblet cell differentiation at mRNA and protein levels in 64 human colorectal cancer cell lines
The microarray expression and PR5D5/MUC2-D staining was summarised. For the cell lines with '***' or '**' in PR5D5 and MUC2-D staining, and not '-' in microarray, they were characterised as 'High GC differentiation (+)'. For the cell lines with '-' in all staining and microarray, they were characterised as 'Low GC differentiation (-)'. The remaining cell lines were characterised as 'Intermediate GC differentiation (+/-)'.

High GC differentiation (+)				Intermediate GC Expression (+/-)				Low GC differentiation (-)			
Cell Line	**PR5D5**	**MUC2D**	**Microarray**	**Cell Line**	**PR5D5**	**MUC2D**	**Microarray**	**Cell Line**	**PR5D5**	**MUC2D**	**Microarray**
C125PM	**	**	**	C106	-	-	*	C10	-	-	-
CL40	**	**	***	C80	*	*	**	C2BBe1	-	-	-
CX1	**	**	*	C84	*	*	**	C32	-	-	-
HCA46	**	**	**	Caco2	-	-	*	C99	-	-	-
HDC114	**	**	**	CAR1	-	-	*	CC20	-	-	-
HDC73	**	**	**	GP2D	-	-	*	COLO201	-	-	-
HT29	**	**	*	HDC111	*	**	*	COLO320DM	-	-	-
KM2012	**	**	*	HDC57	**	**	-	COLO678	-	-	-
LOVO	**	**	**	HDC82	-	-	*	DLD1	-	-	-
LS174T	**	**	**	HT55	-	-	*	GP5D	-	-	-
LS180	**	**	**	JHCOLOY1	*	*	-	HCA7	-	-	-
NCIH508	**	**	***	LIM1215	-	-	*	HCT116	-	-	-
RW7213	**	**	***	LIM1863	*	*	*	HDC142	-	-	-
SW1222	**	**	*	LS1034	-	-	*	HRA19	-	-	-
WIDR	**	**	*	LS123	-	-	*	JHSKREC	-	-	-
				PMFKO14	-	-	*	LIM2405	-	-	-
				RCM1	-	-	*	OUMS23	-	-	-
				RW2982	-	*	*	OXCO1	-	-	-
				SKCO1	-	-	*	OXCO3	-	-	-
				SNUC2B	-	-	*	RKO	-	-	-
				SW480	-	-	*	SW1417	-	-	-
				SW837	-	-	*	SW403	-	-	-
				SW948	-	-	*	SW48	-	-	-
				T84	-	-	**	TITTKB	-	-	-
				VACO10MS	*	**	*				

3.2.3 Identification of differentially expressed genes in high goblet cell differentiation cell lines

In order to determine the goblet cell differentiation-associated genes, the Partek® Genomics Suite® software was used to examine the differential expression profiles of the high (+) versus low (-) goblet cell colorectal cancer cell lines (**Table 3.1**).

A volcano plot was used to visualise the microarray expression analysis. MUC2, as a positive control of microarray analysis, is expressed 3.8 fold higher in high goblet cell-positive cell lines. Notably, the p-value of MUC2, the prime candidate gene from previous result, is 0.0317, which does not have to be corrected for multiple comparisons. FCGBP, another goblet cell marker, also shows a 5.18-fold higher expression in goblet cell-positive cell lines with corrected p-value of 0.0182. TFF3 is expressed 2.48-fold higher in goblet cell-positive cell lines, though the p-value corrected for multiple comparisons is 0.404. Notably, the expression of ST6GALNAC1 (Alpha-N-acetylgalactosaminide alpha-2,6-sialyltransferase 1), a glycotranferase involved in the sialyl Tn antigen transfer during MUC2 O-glycosylation, is expressed 6.08-fold higher in goblet cell-positive cell lines (corrected p-value = 0.053). Detailed up-regulated genes in goblet cell-positive and -negative cell lines can be seen in **Table Appendix.1** and **Table Appendix.2** respectively.

Three unique genes, after correcting for multiple tests and setting the false discovery rate FDR cut-off to 0.01, showed significant expression with corrected p-value<0.01

and fold change>5 in high goblet cell differentiation cell lines from microarray analysis results (**Figure 3.7**). These three genes are Aldo-Keto Reductase Family 1 Member B10 (AKR1B10), Anterior Gradient 3 (AGR3) and Carbonic Anhydrase 12 (CA12). Together with MUC2, these genes were highly differentially expressed between high versus low goblet cell differentiation cell lines and these represent potential biomarkers for goblet cell differentiation (Figure 3.8A). The expression levels of these genes in most cell lines are well above 200AU, and the detailed mRNA expression levels of the identified genes were investigated in the microarray data as shown in **Figure 3.8B**.

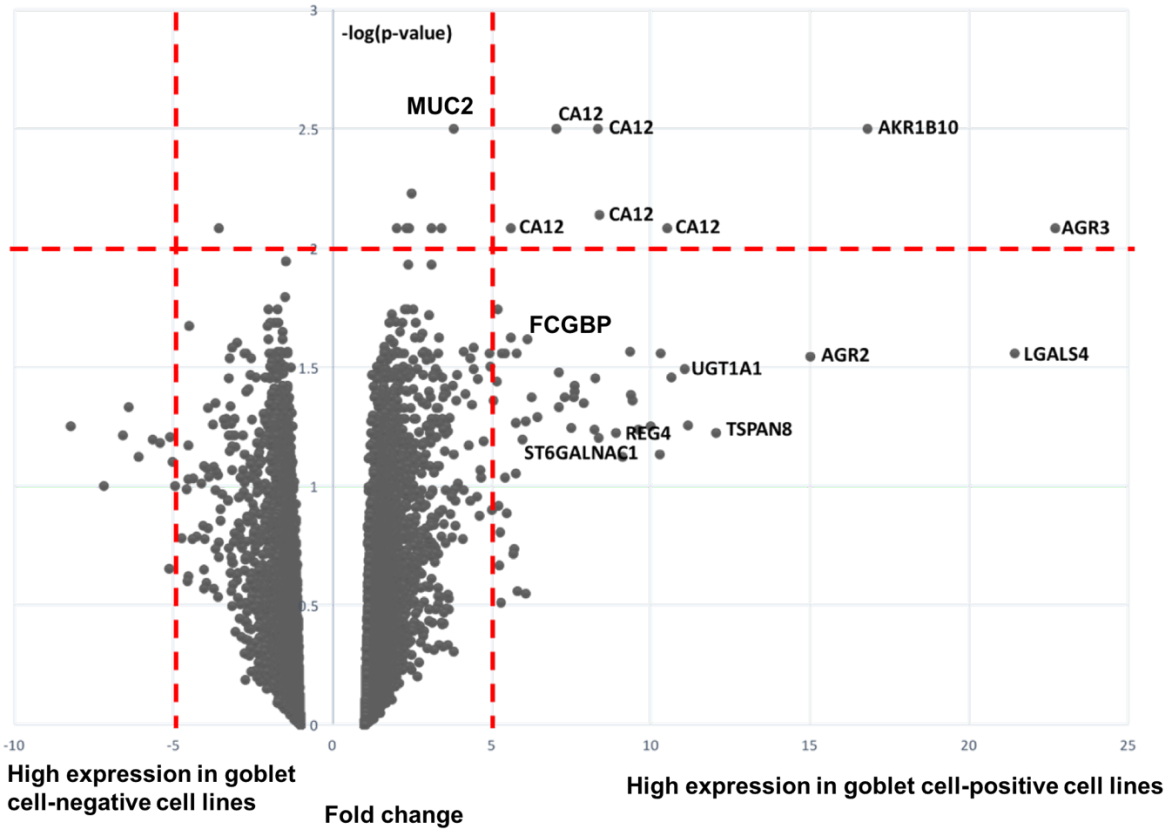


Figure 3.7 Volcano plot representation of microarray expression analysis between goblet cell-positive and -negative cell lines

Microarray expression profiles of high goblet cell group (positive fold-change) and low goblet cell group (negative fold-change) were plotted according to the absolute fold change (x-axis) and log10 p-value (corrected for multiple comparisons) (y-axis). Genes were identified as significantly differential expressed if the corrected p-value was less than 0.01 and fold change was greater than 5.

A

Gene Symbol	Gene Title	p-value	fold-change	Description
MUC2	Mucin 2	0.0317	3.80601	goblet cell pos vs neg
AKR1B10	Aldo-Keto Reductase Family 1 Member B10	0.0031	16.8033	goblet cell pos vs neg
AGR3	Anterior Gradient 3	0.0083	22.7101	goblet cell pos vs neg
CA12	Carbonic Anhydrase 12	0.0083	10.5218	goblet cell pos vs neg

B

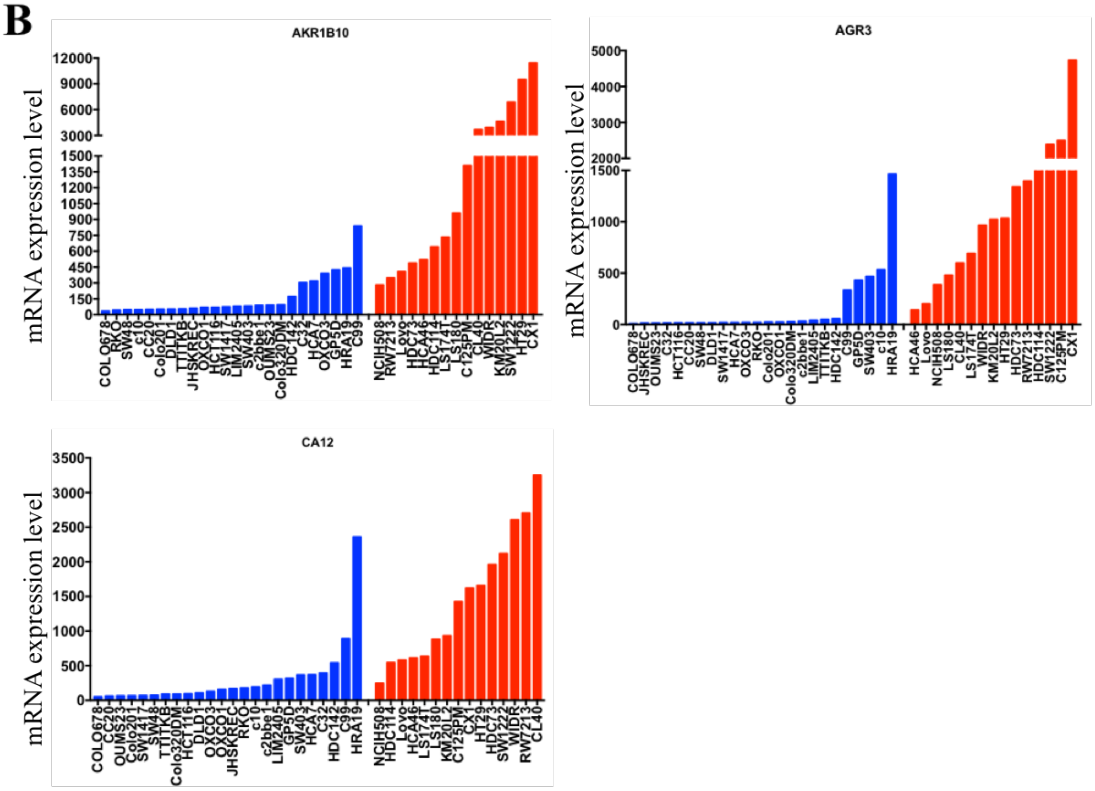


Figure 3.8 Detailed microarray mRNA expression of the differentially expressed genes in goblet cell positive cell lines

(**A**) List of MUC2 and the top three genes that are most significantly differentially expressed between goblet cell-positive and -negative cell lines, ranking by p-value. Fold changes are given relative to goblet cell-negative cell lines; such that positive fold changes reflect genes that are expressed at higher levels in goblet cell-positive compared to -negative cell lines. (**B**) AKR1B10, AGR3 and CA12 mRNA expression in goblet cell-positive (red) and -negative (blue) cell lines measured by Affymetrix Human Genome U133 plus 2.0 microarray.

For AKR1B10 gene, goblet cell-positive cell lines showed 16.8-fold higher expression than goblet cell-negative cell lines ($p = 0.0032$). However, a distinct subset of five goblet cell-negative cell lines: C99, HRA19, GP5D, OXCO3, HCA7 and C32, showed a notably higher expression of AKR1B10 compared to the remaining goblet cell-negative cell lines. Similarly, the expression of AGR3 gene in goblet cell-positive cell lines is 22.7-fold higher than goblet cell-negative cell lines ($p = 0.0083$). The goblet cell-negative cell lines HRA19, C10, SW403, GP5D and C99, showed a clearly higher expression of AGR3 compared to other goblet cell-negative cell lines. For the CA12 gene (fold-change = 10.5, $p = 0.0083$), three cell lines, HRA19, C99 and HDC142 expressed CA12 distinctly from other goblet cell-negative cell lines. These goblet cell-negative cell lines, especially HRA19, C99 and GP5D, represent a subset of cell lines that cannot produce MUC2, but to some extent express the genes that are highly expressed in goblet cell-positive cell lines.

Moreover, compared to MUC2 (3.8-fold higher expression in goblet cell-positive cell lines), AKR1B10, ARG3 and CA12 are highly expressed to a larger extent (more than 10-fold). This is because MUC2 is highly expressed in a small subset of cells and their mRNA can be largely diluted in the bulk population analysis. In contrast, these three genes might be expressed in a larger proportion of the total population. Thus, it is possible that these genes do not directly target goblet cells, but have potential regulatory functions in goblet cell differentiation.

CA12 is a gene that is highly differentially expressed between goblet cell-positive and -negative cell lines, and thus presents the potential to be a novel goblet cell marker. Notably, CA12 is a trans-membrane protein that is reported to be highly expressed in normal colonic tissues. All five probes of CA12 in the microarray data have shown to be highly expressed in goblet cell-positive cell lines (**Figure 3.8**). This suggests the possibility of CA12 as a surface marker for goblet cells or goblet cell precursors. In addition, CA12 has been reported to have a highly confined expression in the basolateral plasma membrane of enterocytes only in the normal colon (Kivela et al., 2000). In colorectal cancers, however, anti-CA12 diffusely stained the adenomatous mucosa and the staining increased with the degree of dysplasia (Kivelä, Antti, et al. 2000). This has been a reasonable starting point to understand the role of CA12 on goblet cell differentiation in human colorectal cancers, which will be further discussed in **Chapter 5**.

3.3 Discussion

To characterise the genes that are differentially expressed during goblet cell differentiation, the PR5D5 antibody was used to target goblet cells in the human colon. The reason we used PR5D5 is due to its exceptional specificity, affinity and large amounts of stock. PR5D5 was raised by immunising and boosting BALB/c mice with mucosal scrapings from normal human colons (Richman et al., 1987). The hybridoma supernatants were screened immunohistochemically on frozen sections of normal colorectal epithelium, and PR5D5 was found to specifically target goblet cells in both fresh and fixed tissues (Richman et al., 1987). Although PR5D5 recognised a glycoprotein of ~700 kilo Daltons from Western Blot which is presumably the glycosylated MUC2 protein (Campbell et al., 1994), there is no publication to date that describes unequivocally the specific molecular target of the PR5D5 antibody. The co-staining results illustrate that PR5D5 and another two commercial MUC2 antibodies targeted the same goblet cells. Moreover, the decreased PR5D5 staining after MUC2 gene silencing and the competitive binding between PR5D5 and MUC2D antibodies further confirm MUC2 as the target of PR5D5. These results show, for the first time, that the molecular target of PR5D5 is MUC2 and that PR5D5 can be used as a MUC2-specific antibody for detecting goblet cells in human tissues and cell lines. In defining the MUC2-specificity of PR5D5, it is also worthy using forced expression of MUC2 in the PR5D5-negative cell lines in complementary to the knock down and protein competition. Also the immunoprecipitation and immunoblot of MUC2 and PR5D5 antibodies followed by mass spectrometry and peptide sequencing will provide further

evidence of PR5D5 specificity. These experiments should be conducted in the future research.

Screening of goblet cell differentiation at protein and mRNA levels allowed the microarray expression analysis and identification of the differentially expressed gene between goblet cell-positive and –negative cell lines. Given the microarray screening of these over 100 cell lines were conducted in a high-throughput manner over years ago, the gene-specific qRT-PCR on MUC2 mRNA expression should be conducted in selected cell lines to further validate its expression and avoid false positive results (Morey, et al., 2006).

Three genes, AKR1B10, AGR3 and CA12, showed significantly differential expressions (fold-change > 5 and p-values < 0.01) in goblet cell-positive cells lines. AKR1B10, as a transcriptional target of p53, was suggested to be under-expressed in human colorectal cancers (Ohashi et al., 2013; Laffin and Petrash, 2012). The decreased expression of AKR1B10 is associated with the metastasis and invasion of colorectal cancer cell lines HT29 and KM20 (Tammali et al., 2011) and poor prognosis in colon cancer patients (Ohashi et al., 2013). Thus, AKR1B10 has the potential as a novel therapeutic target or a candidate prognostic marker regarding metastasis, as it was reported in, non-small cell lung carcinomas (Penning, 2005) and gastric cancers (Yao et al., 2014). However, its role in relation to goblet cell differentiation remains to be determined.

AGR3 is known as a cysteine disulphide isomerase to catalyse the disulphide bond formation in the endoplasmic reticulum. However, the functional understanding of AGR3 in the goblet cell differentiation is very limited. Its homolog AGR2, which is also a top gene from the microarray analysis (fold change > 15-fold, p-value < 0.05 corrected for multiple comparisons) (**Figure 3.7**), seems to have a more important role for goblet cells in colorectal cancers. The AGR2 dysfunction can result in diarrhoea and goblet cell dysfunction in mice (European patent WO2004056858). The AGR2 promoter was also regulated by FOXA1 and FOXA2, two goblet cell-related regulatory factors (W Zheng et al., 2006). These results indicate the important functions of the protein disulphide isomerases in forming the functional net-like mucus during goblet cell maturation. It remains to be determined what the role of these protein disulphide isomerases may be in goblet cell differentiation in the human colon, and more specifically in colorectal cancers.

CA12 is a member of the carbonic anhydrase family, which can catalyse the reversible hydration of carbon dioxide. Despite the physiologically important roles of the entire carbonic anhydrase family, e.g. respiration, gluconeogenesis and signalling transduction (WS Sly et al., 1995), the specific roles of CA12 in goblet cell differentiation in colorectal cancers have been poorly characterised. Given the fact that CA12 is a Type I transmembrane protein and highly expressed in the goblet cell-positive human colorectal cancer cell lines, it is reasonable to hypothesise that CA12

might serve as a potential marker for goblet cells or goblet cell progenitors. This hypothesis will be further investigated in **Chapter 5**.

In addition, ST6GALNAC1 was identified to be expressed more than 5-fold differential expression, indicating ST6GALNAC1 is the major glycotransferase to catalyse silyl-Tn in goblet cell-positive cell lines. This is consistent with the observation that MUC2 is the major carrier of the cancer-associated sialyl-Tn antigen (Conze et al., 2010) and the roles of ST6GALNAC1 in the synthesis of the cancer-associated sialyl-Tn antigen.

Collectively, the results presented in this chapter confirmed the molecular target of the antibody PR5D5. Using PR5D5 for the screening of 64 human colorectal cancer cell lines, several genes were identified that are differentially expressed genes that might serve as the molecule markers of goblet cells or progenitors.

CHAPTER 4

CHARACTERIZATION OF GOBLET CELL TRANSCRIPTOME

4.1 Introduction

Results on gene expression in the previous chapter are all based on studying cell lines in which there are varying proportions of differentiated goblet cells and other cell types. In this chapter, a procedure is developed for isolating goblet cells using the PR5D5 antibody which enables a direct assessment of gene expression in purified goblet cells. However, this is largely restricted by the lack of surface markers for FACS based isolation of goblet cells, and hence it needs to develop a procedure using PR5D5 which only work on fixed cells.

Several publications have described the attempts to address this problem. Saadi Khochbin and colleagues fixed cells with 70% ethanol and then re-suspended the pellet in 5x Saline Sodium Citrate (SSC) buffer before staining and sorting (Khchbin et al., 1990). They eliminated RNases in the flow cytometers with 10% H_2O_2, followed by rinsing with 5x SSC as sheath fluid for 10 minutes (Khchbin et al., 1990). Charlotte Esser and colleagues fixed cells with ethanol/acetic acid (95:5) or 70% methanol with 20mM vanadyl ribonucleoside complexes. They used 2% tryptone as protein additive to replace calf serum or albumin that could not be autoclaved at that time (Esser et al., 1995). Cells were stained each time with newly-opened commercial antibodies that were believed to be RNase-free. Different parts of the flow cytometry sorter were either baked at 180°C for 3 hours or washed in 0.1% sodium dodecyl-sulphate and rinsed with diethyl pyrocarbonate (DEPC)-treated water (Esser et al., 1995). These complex steps

set up barriers to its application. In both protocols, either 1 million or 5 million sorted cells were collected for RNA extraction, which requires a long sorting period sorting for goblet cells which accounts for only 1% of the total FACS events. This might put additional risk to include exogenous RNase contamination. Furthermore, neither paper described the application of their protocols in global expression analysis but only in Northern Blot.

Pan and colleagues confirmed the poor RNA quality presumably of fixed cells given by the standard FACS staining protocol before and after sorting (Pan et al., 2011). They reported that fixation using 4% paraformaldehyde was not the major cause for RNA degradation by varying the duration of fixation. They optimised the RNA integrity by using a washing buffer that consisted of 100ug/mL BSA, 100U/mL RNase inhibitor and 5mM DTT (Pan et al., 2011). Similarly, Hrvatin and colleagues fixed and permeabilised cells with 4% PFA and 0.1% saponin, and included 1:100 or 1:25 RNasin Plus RNase Inhibitor in the washing/staining buffer and sorting buffer respectively (Hrvatin et al., 2014). The total RNA from sorted cells in both protocols were extracted using the RecoverAll Total Nucleic Acid Isolation kit that can de-cross the links between RNA and PFA before downstream sequencing. Despite the application of both protocols in global gene expression analysis via microarray or RNA sequencing, the large volume of RNase inhibitors limits its reproducibility for economic reasons. In addition, there was still controversy about the effects of different fixatives on staining patterns and RNA preservation (Pan et al., 2011; Medeiros et al., 2007; Cox et al., 2006;

Goldsworthy et al., 1999). Thus, it is essential to examine fixatives and permeabilisation agents, as well as efficient RNase inhibitory reagents that can be used in an economic amount to successfully extract RNA from fixed and permeabilised cells.

Chapter 4 describes a novel protocol to overcome these barriers. RNA degradation was systematically examined during each step of the whole procedure. The fixative and permeabilising reagents were evaluated for minimised staining disturbance and RNA degradation. In order to eliminate the influence of RNases, several RNase inhibitory reagents were compared for their ability to preserve RNA. Based on the optimised protocol, RNA isolated from the FACS-enriched fixed cells showed the same quality with unfixed cells. The RNA-seq data from the fixed goblet cells and non-goblet cells were then analysed using edgeR, with proper quality control by using the fastQC package and checking the key goblet cell-related genes. This has given the transcriptomic characterisation of goblet cells in human colorectal cancers. By comparing the transcriptomic profile of goblet cells versus non-goblet cells, we extended the identification of goblet cell-specific gene expression including in particular the search for GC specific surface cell markers.

4.2 Results

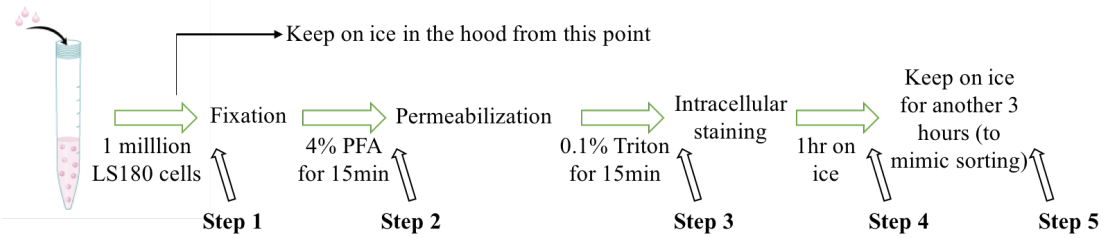
4.2.1 Optimisation of RNA extraction from fixed and permeablised cells

4.2.1.1 Quantification of RNA degradation during fixation, permeabilisation, intracellular staining and FACS sorting

RNA degradation can potentially occur at any point during RNA extraction. We identified four major experimental steps that might represent an increased RNase contamination risk: fixation, permeabilisation, intracellular staining and FACS sorting. Understanding the RNA degradation process in the methodology will help evaluate the time point to include RNase inhibitory reagents and further optimise the procedure.

After trypsinisation, washing and counting, 1 million LS180 cells were fixed with 4% PFA and permeabilised with 1% Triton before intracellular staining with PR5D5 and anti-mouse APC secondary antibody. Cells were then incubated on ice for another 3 hours to mimic the sorting period. Total RNA was isolated at different steps (**Figure 4.1A**) using RecoverAll™ Total Nucleic Acid Isolation Kit for FFPE (Ambion, UK). One optimisation technique has been derived from the literature (Hrvatin, S etc al, 2011) which outlined that for protease digestion and RNA isolation, instead of incubating for 15min at 50 °C and another 15min at 80 °C, cell lysates should be incubated for 3 hours at 50 °C.

A



B

	Step 1	Step 2	Step 3	Step 4	Step 5
A260/280	2.03	1.93	1.72	0.47	0.21
A260/230	2.06	1.55	0.35	0.12	0.27
RNA concentration (ng/uL)	124.5	113.2	30.7	10.3	4.2

Figure 4.1 Identification of the key steps of RNA degradation
(A) Schematic workflow of key steps in which RNA was extracted for quantification. (B) List of the values of A260/280, A260/230 and RNA concentrations before each step during the FACS staining protocol.

The RNA concentration and purity were measured using NanoDrop (**Figure 4.1B**). Before fixation, RNA concentration is 124.5ng/uL in 40uL of elution reagents, and the 260nm/280nm absorbance ratio is 2.03, which is generally considered as pure RNA. After fixation with 4% PFA, the RNA concentration and 260/280 absorbance ratio decreased to 113.2ng/uL and 1.93 respectively. Meanwhile, the 260/230 absorbance ratio, a secondary measurement of RNA purity, decreased from 2.06 to 1.55, indicating existence of contaminants with high absorbance at 230nm.

The significant decrease of RNA concentration occurred during step 3, after cell wash and permeabilisation with 0.1% Triton. This step led to the decrease of RNA concentration to 30.7ng/uL and the value of 260/230 ratio to 0.35, while a moderate decrease of A260/280 value was observed to 1.72. This indicates that the permeabilisation was the key step in introducing exogenous RNase contamination decreasing RNA concentration and purity, and would further interfere with downstream sequencing. All three parameters: RNA concentration, 260/230 and 260/230 ratios, further decreased and indicated that RNA after step 3 would be not suitable for transcriptomic characterisation.

The quantification of RNA degradation after each step identified permeabilisation as the key step to introduce contaminants and external RNases. These results highlight the importance of the addition of suitable RNase inhibitory reagents after the fixation step.

4.2.1.2 Selection of fixation and permeabilisation methods

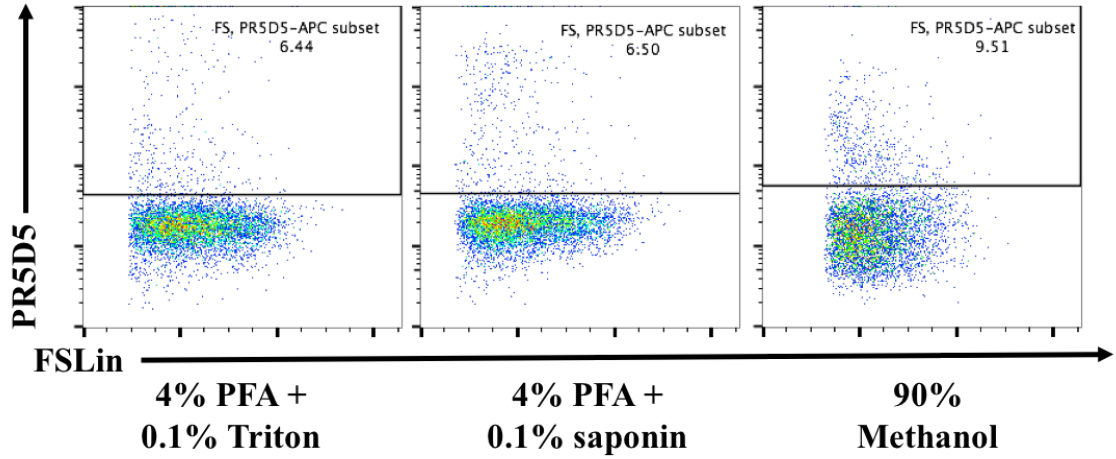
It is critical to choose appropriate fixatives and permeabilisation reagents to preserve i) the FACS staining pattern and ii) RNA integrity. Three common combinations of fixatives and permeabilisation agents were chosen as potential candidates for further evaluation, i.e. 90% methanol, 4% PFA + 0.1% Triton and 4% PFA + 0.1% saponin.

The staining patterns were firstly examined. After labelling with the in-house mouse PR5D5 antibody and anti-mouse APC secondary antibody, 90% methanol gave the highest positive proportions of PR5D5 positive cells (9.51%) compared to 4% PFA with 0.1% Triton (6.44%) or 0.1% saponin (6.50%). Different fixatives, however, gave distinct forward scatters (FS) that correlates with cell sizes. Methanol showed a more uniform distribution of cell sizes compared with PFA-based fixation (**Figure 4.2A**). This might be due to its ability to dissolve cell surface lipids and precipitate proteins, resulting in cellular shrinkage.

The concentration and purity of RNA using different fixatives and permeabilisation agents were also investigated immediately after permeabilisation (**Figure 4.2B**). Compared to unfixed samples (110ng/uL), the samples fixed by PFA showed either reasonably comparable RNA concentration with saponin (100.4ng/uL) or slightly decreased RNA concentration with Triton (83.2ng/uL). The fixation and perimeabilization using methanol, however, showed a significantly lower RNA

concentration of only 30.7ng/uL. Moreover, methanol also gave lower A260/280 values (1.02) than PFA with Triton (1.73) or saponin (1.63) and unfixed samples (1.97). This indicates that more protein contamination occurred with methanol being used as the fixative and permeabilisation agent. All three fixatives and permeabilisation agents showed unsatisfactory A260/230 ratios compared to unfixed samples, which again highlights the significance of choosing suitable RNase inhibitory reagents (see **4.2.1.3**).

A



B

	Unfixed	4% PFA + 0.1% saponin	4% PFA + 0.1% Triton	90% Methanol
A260/280	1.97	1.63	1.73	1.02
A260/230	2.12	0.45	0.36	0.35
RNA concentration (ng/uL)	110	100.4	83.2	30.7

Figure 4.2 The effects of fixatives and permeabilisation agents on staining patterns and RNA degradation

(A) LS180 cells were fixed and permeabilised with methanol or PFA with Triton or saponin, and stained with PR5D5 antibody, before analysed using flow cytometry. The x-axis is forward scatter, and y-axis is PR5D5 labelled with anti-mouse APC. (B) List of the values of A260/280, A260/230 and RNA concentrations using different fixatives and permeabilisation agents.

The high integrity of RNA is important for unbiased sequencing. The assessment of RNA integrity is obtained using the Agilent 2100 Bioanalyser, a commonly used electrophoresis-based platform for RNA quality evaluation. The RNA Integrity Number (RIN) is calculated in the software based on the whole electrophoretic trace to evaluate the RNA quality with the classification numbering system from 1 (the most degraded) to 10 (the most intact). Compared to traditional methods for RNA integrity assessment that compares the ratios of 28s and 18s rRNAs (usually a 2:1 ratio is considered as intact RNA), the RIN shows advantages of standardising the subjective measurement and computation of RNA integrity using a fairly small amount of RNA. Usually, an RIN number greater than 7 is considered to be of reasonable quality for microarray and RNA-seq analysis.

To investigate the influence of different fixation and permeabilisation reagents, RNA was extracted after intracellular staining and compared to the unfixed samples (**Figure 4.3**). For unfixed cells, the RIN value was 9.50 with clear sharp peaks of 18s and 28s rRNA. For the PFA fixation, RIN numbers were 8.60 and 8.30 with saponin and Triton respectively. However, the 28s band of the Triton-treated sample indicated a slight shift towards shorter fragments, indicating greater RNA degradation compared to unfixed and saponin-treated samples. For methanol-treated samples, the RIN value decreased to 7.20, indicating the degradation of RNA using this reagent. In addition, the peak shapes of 18s and 28s ribosomes and the accumulated RNA fragments along the baseline suggested the unsatisfactory preservation of RNA in this fixative. It is

reasonable to deduce that RNA will further degrade during staining and sorting when using methanol. Thus, based on the RNA extraction results presented here, PFA with saponin is the most suitable fixative and permeabilisation agent to preserve RNA integrity among the tested combinations.

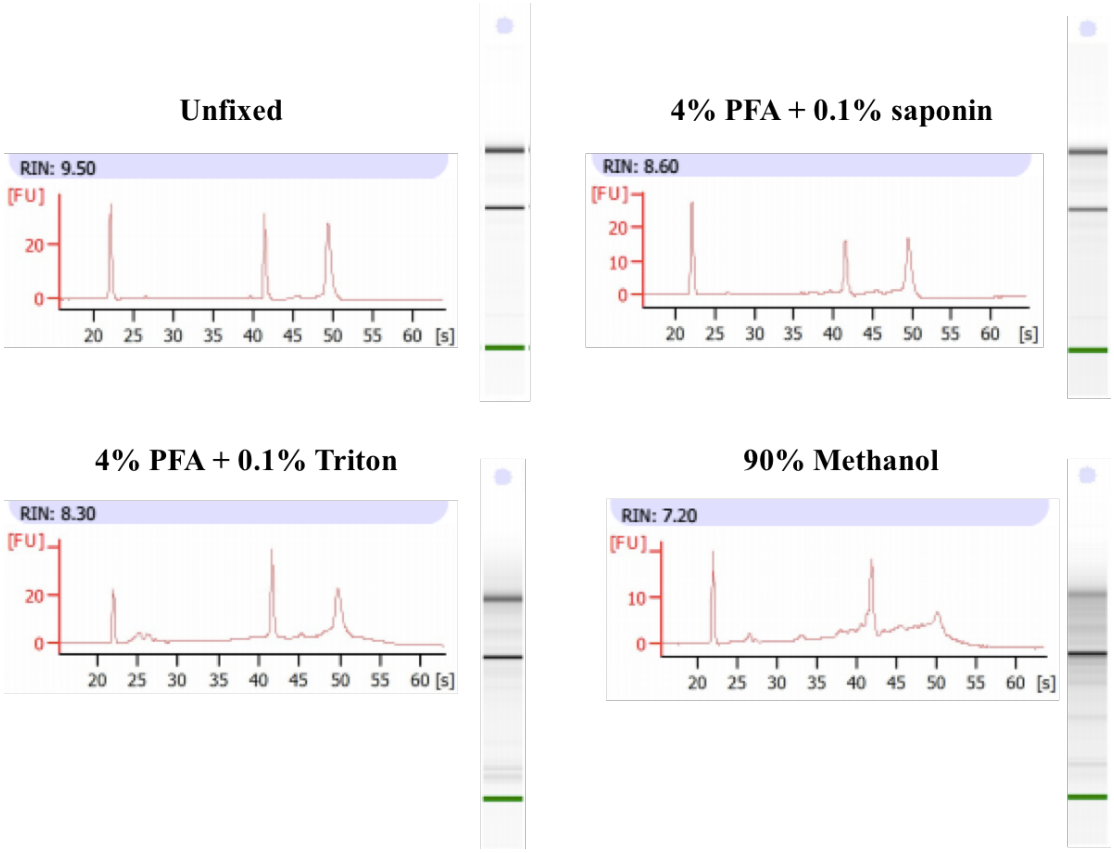


Figure 4.3 Assessment of RNA integrity using different fixatives and permeabilisation agents

The unfixed samples showed clear peaks at 18s and 28s ribosomal band with RIN value of 9.50 (top left). The 4% PFA fixed samples with 0.1% saponin (top right) and 0.1% Triton (bottom left) showed decreased RIN values of 8.60 and 8.30 respectively. The RIN value of the sample fixed using 90% methanol (bottom right) is 7.20.

4.2.1.3 Selection of RNase inhibitory reagents

A panel of RNase inhibitory reagents that are commonly used in molecular and biochemical experiments were selected to test their influences on staining patterns, RNA quantity and RNA quality (**Figure 4.4A**).

RNAlater (Invitrogen, UK) was used before permeabilisation to preserve RNA but it was found to interfere with the intracellular staining and no PR5D5-positive signals could be detected by flow cytometry (**Figure 4.4B**). Another RNase inactivation reagent, RNAsecure, requires incubation at 60°C for 10 minutes. This disrupts cellular and protein structures, and disrupted the staining pattern with altered forward scatters and a largely decreased PR5D5-positive proportion of LS180 cells (**Figure 4.4C)**. By interfering the staining patterns, these two reagents are not suitable for further investigation.

The RNase Inhibitor (TF) (Thermo Fisher Scientifics, UK), part of the High-Capacity cDNA Archive, was also tested regarding its capability to preserve RNA without affecting FACS staining. Its key component is a 50kDa recombinant protein that enzymatically inhibit RNase activity. Though this reagent did not interfere with the staining (**Figure 4.4D)**, it could not perverse RNA at a reasonable concentration for sequencing (**Figure 4.4A**). Thus, we did not investigate the RNA integrity.

Two RNAse inhibitors, RNaseOUT (Thermo Fisher Scientific) and RNasin Plus (Promega), showed the capability of keeping the PR5D5 staining patterns (**Figure 4.4E, 4.4F)** and maintaining RNA quantity (**Figure 4.4A**). These two inhibitors were further investigated for their capability to preserve RNA integrity. **Figure 4.4G** presents an obvious RNA degradation in the sample treated with RNAseOUT, with an RIN value of only 3.10, suggesting the lack of effective RNA protection. RNasin Plus inhibitor, on the other hand, showed clear peaks at 18s and 28s bands with RIN value of 8.30. Thus, the RNAsin was selected as the RNase inhibitory reagent for the fixed cell staining and sorting.

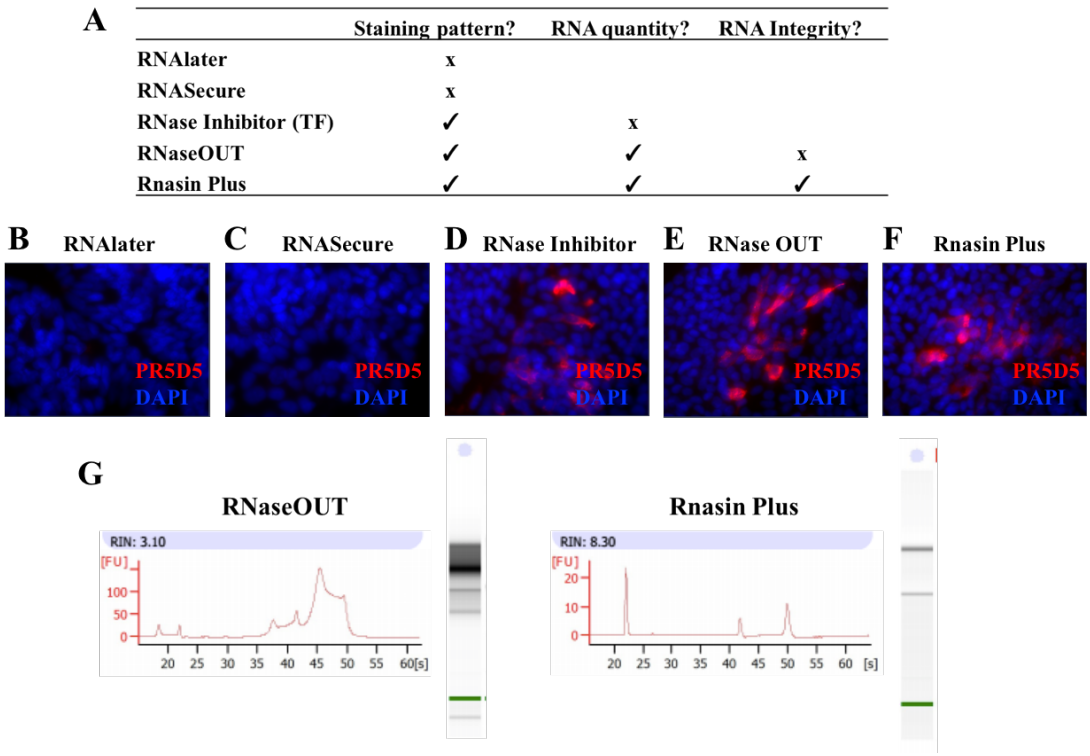


Figure 4.4 Effects of different RNase inhibitory reagents on RNA preservation
(**A**) Summary of different RNase inhibitory reagents on staining patterns, RNA quantity and RNA integrity. RNaseOUT and RNasin Plus, the two RNase inhibitors that do not disturb staining pattern and maintain RNA quantity, were assessed for the capability of preserving RNA integrity. (**B-F**) Immunofluorescent staining examples of LS180 after treated with different RNase inhibitory reagents. RNAlater (B) and RNASecure (C) largely reduced the PR5D5-positive cell number, while RNase Inhibitor (D), RNase Out (E) and Rnasin Plus (F) can maintain the PR5D5 staining patterns for goblet cells. (**G**) The RIN value of RNaseOUT is only 3.10, while RNasin Plus is 8.30.

4.2.2 Transcriptomic characterisation of fixed goblet cells

4.2.2.1 Optimised experimental workflow

Optimisations were then included into the intracellular staining and FACS sorting (full details as described in **Chapter 2, Section 2.6.3**). In brief, **Figure 4.5** demonstrates the optimised workflow. All the staining and sorting steps were carried out at 4°C. The cells were fixed with 4% PFA and permeabilised using 0.1% saponin with 1% RNase-free BSA (Gemini, UK) and 0.005% RNasin Plus RNase inhibitor (Promega, UK). After permeabilising, cells were stained with 1:200 diluted PR5D5 antibody and washed with the RNase-free staining buffer that contains 1x PBS, 1% RNase-free BSA and 0.01%. Instead of one million cells, 100 cells were sorted into each of the strip tubes that contained protein lysis buffer (5uL PKD buffer, 1:4 Proteinase K and ERCC RNA Spike-In Mix). The small cell number sorted in each tube largely reduced the sorting time and minimise RNA degradation. Using this method, 7 tubes of PR5D5-positive cells and 7 tubes of PR5D5-negative cells were prepared for RNA isolation. The miRNeasy FFPE Kit was used to lyse cells and purify RNA. To reverse formalin crosslinking and prevent RNA degradation at high temperature, cells were incubated for 30mins at 50°C instead of 15mins at 50°C and 15mins at 80 °C.

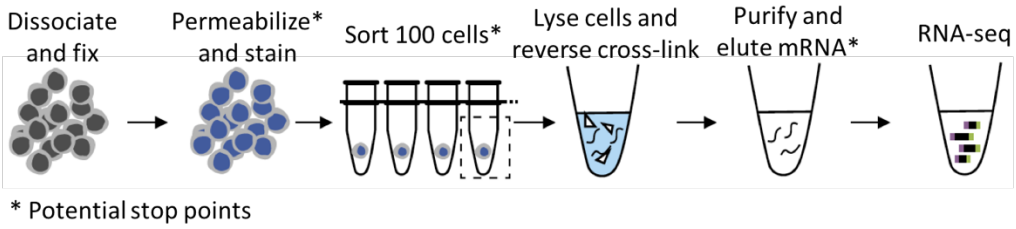


Figure 4.5 Experimental workflow for the optimised protocol. Asterisks represent potential stop points when samples can be stored in -80°C for weeks (Modified from Thomsen et al., 2016).

4.2.2.2 RNA extraction and sequencing from fixed and sorted goblet cells

After RNA isolation and reverse transcription, the cDNA libraries of fixed samples were sent for evaluation and quantification on a Bioanalyser 2100, and compared with the unfixed samples (**Figure 4.6**). Using the optimised workflow, the cDNA libraries of both fixed goblet cells and non-goblet cells showed the similar cDNA size distribution compared to the unfixed samples. The library profile shows a peak of ~1-1.5kb and a small amount of cDNA fragments of ~500-750bp, as well as a moderate peak of primer dimer at 35bp. There was no obvious degradation which would result in the shift of peaks. These results indicate the successful preamplification of cDNA with comparable quality of both fixed and unfixed samples for RNA sequencing.

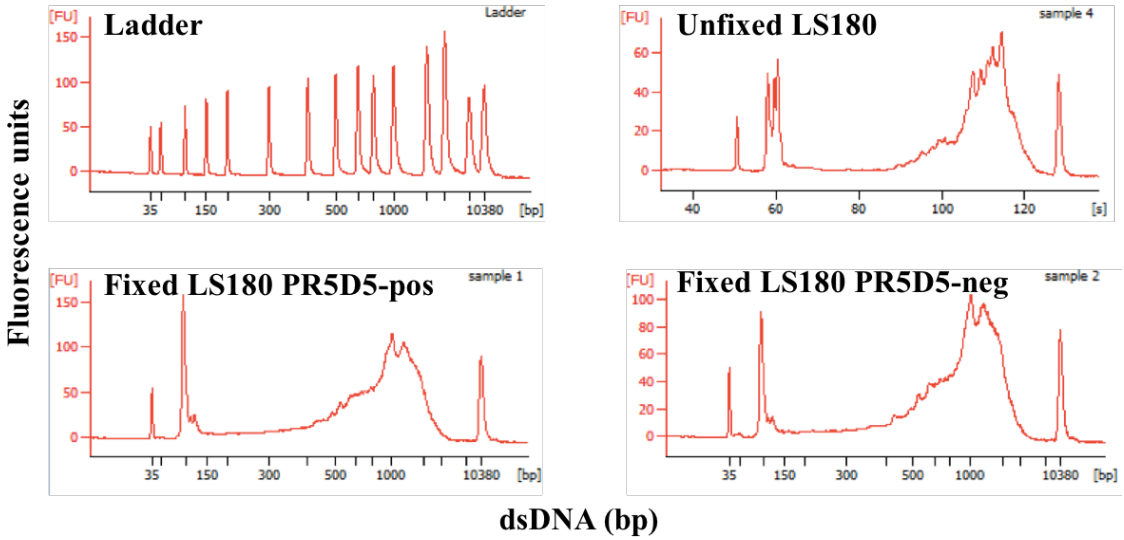


Figure 4.6 The optimised protocol enables mRNA isolation from fixed samples of comparable quality with unfixed ones
Representative example of the Bioanalyser analysis of the amplified cDNA for the unfixed, fixed PR5D5-positive and fixed PR5D5-negative samples, suggesting the comparable integrity of mRNA.

4.2.2.3 RNA-sequencing analysis identifies goblet cell specific genes

14 cDNA libraries from LS180, 7 PR5D5-positive and 7 PR5D5-negative samples, were sent to the NextGen Sequencing Facility at Weatherall Institute of Molecular Medicine for next generation sequencing. All subsequent RNA-seq analysis was conducted by Haoyu. **Figure 4.7** describes the key steps of quality check, read mapping, feature counting and differential expression analysis. The detailed codes of the whole procedure are shown in **Code Appendix.1**. After a quality check using fastQC, the raw data in fastQ format were mapped with STAR and was converted into BAM or SAM format, which can be viewed at UCSC Genome Browser before differential expression analysis. Using the count summary via FeatureCounts function as input into the edgeR package, the lowly expressed genes were filtered, and the sequencing depth and RNA composition were normalised before differential expression analysis.

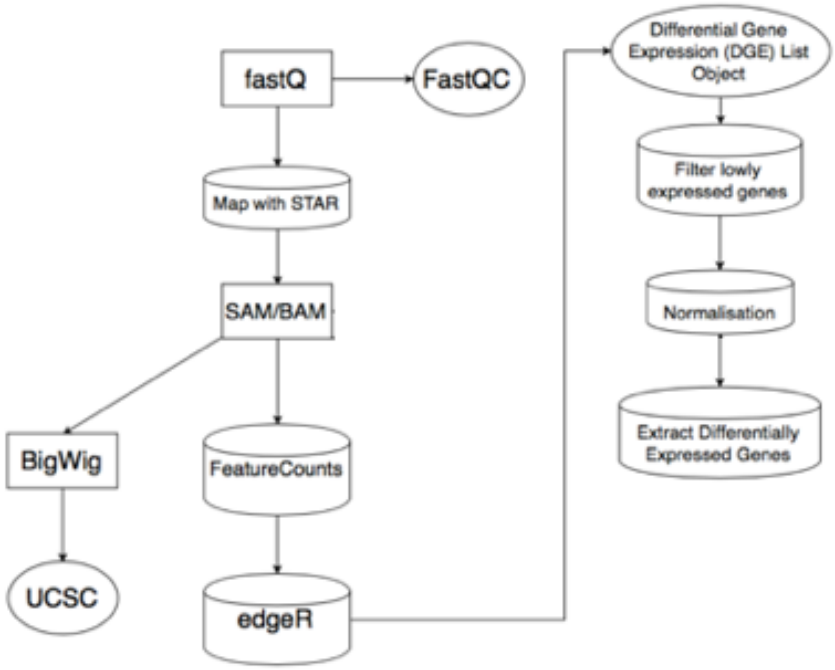


Figure 4.7 Workflow of RNA sequencing differential expression analysis

4.2.2.3.1 Quality evaluation of fastq files

All the detailed codes for RNA-seq analysis can be seen in **Code Appendix.1**. The quality of raw fastQ files was checked using FastQC (Andrews, 2010). Here we are using one of the sequencing samples as an example to illustrate the quality evaluation of fastq files.

The Phred quality score is a widely accepted measurement that is assigned to each nucleotide base call in the automated sequencer tracers to examine the nucleotide quality (Ewing, et al., 1998; Dear and Staden, 1992). The Phred score is defined to be logarithmically related to the base-calling error probabilities. For example, the Phred quality score 20 indicates the 1% probability of incorrect base call, i.e. 99% accuracy; the Phred quality score 30 indicates the 0.1% probability of incorrect base call, i.e. 99.9% accuracy. The Phred quality across all bases from the beginning to the end of each read were examined. As shown in **Figure 4.8**, most parts of the sequence (68 out of total 75 bases), from position 5 to 72, showed reasonable or good quality. The guanine-cytosine content (GC content) was measured through all sequences and compared with the theoretical distribution of GC content that is supposed to exist in a normal random library. The GC distribution across all sequences largely overlapped with the theoretical one that is expected to exist in a normal random library (**Figure 4.9**), suggesting an acceptable GC content distribution of this library.

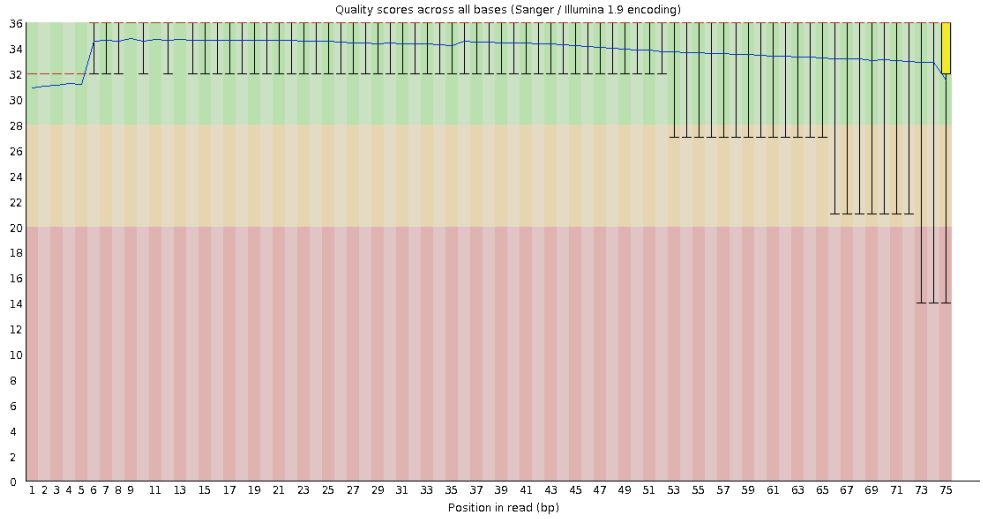


Figure 4.8 Quality scores across all bases of reads

The x-axis shows the position of each base, and the y-axis indicates the quality scores of each position. Background colours represent different base qualities, good quality (green), reasonable quality (yellow) and poor quality (red). In this sample, the length of each sequence is 75bp. The bases from the position 5 to 72 have good or reasonable quality.

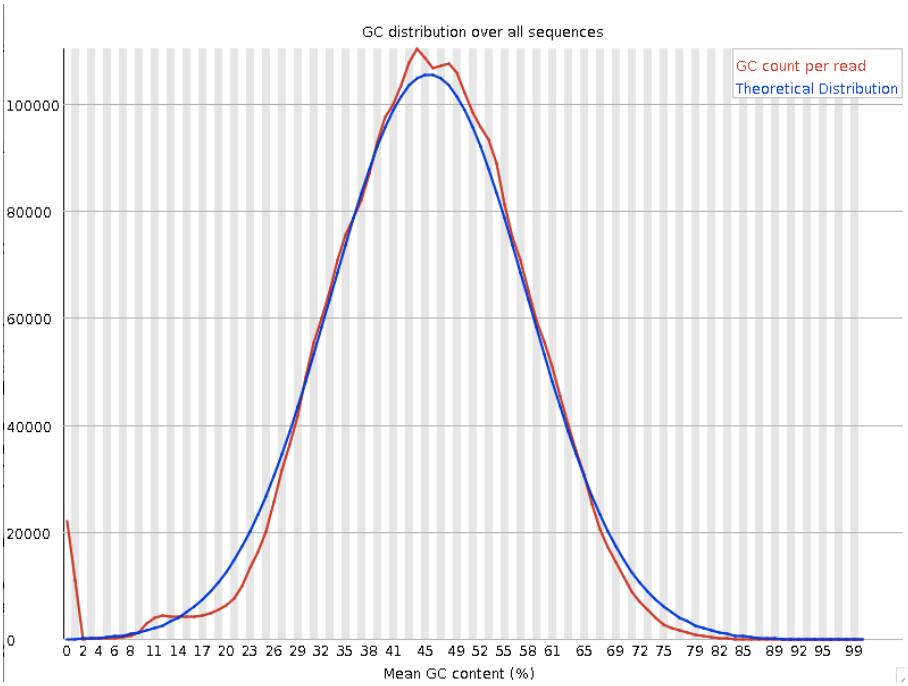


Figure 4.9 GC distribution over all sequences

The red line represents the GC distribution of the existing library. The blue line represents the theoretical distribution of a random library of similar size (Andrews, 2010). The x-axis is the mean GC content (%), and the y-axis is the counts of sequences. A mis-distributed GC content demonstrates the potential contamination of libraries, while a shifted peak indicates systematic bias.

The GC content across each base position was also examined (**Figure 4.10**) and demonstrated little difference from position 23 to 74 (~45%) except at the beginning and end of sequences. This indicates no bias from the original library or systematic issue during sequencing (Andrews, 2010). In order to identify potential overrepresented sequences and libraries with biased composition, the sequence composition was checked across the whole reads (**Figure 4.11**).

Consistent with the GC content, the percentage of each nucleotide base remain the same from the positions 22 to 73. Notably, the first 12bp of each sequence showed significant nucleotide diversity. This is due to the unavoidable selection bias in the libraries generated from random hexamer ligation or tagmentation, but represent no individual biased sequences and will not affect downstream differential analysis expression (Andrews, 2010).

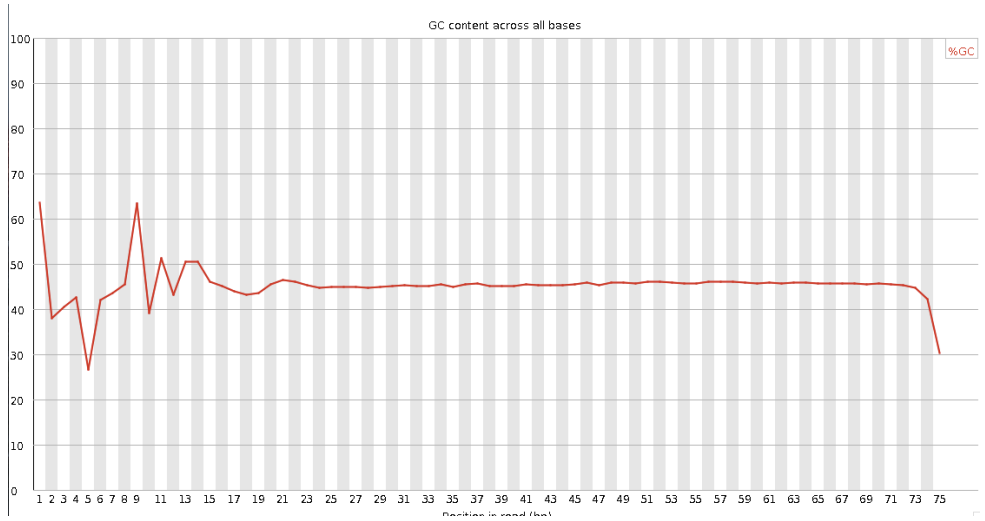


Figure 4.10 GC content across all bases

The x-axis is the base position of each sequence, and the y-axis is the GC content percentage. In a random library, GC percentage is maintained the same across whole sequences, and a horizontal line should be expected in this figure (Andrews, 2010). The consistent GC contents across all bases indicate the unbiased libraries and no systematic problem during sequencing.

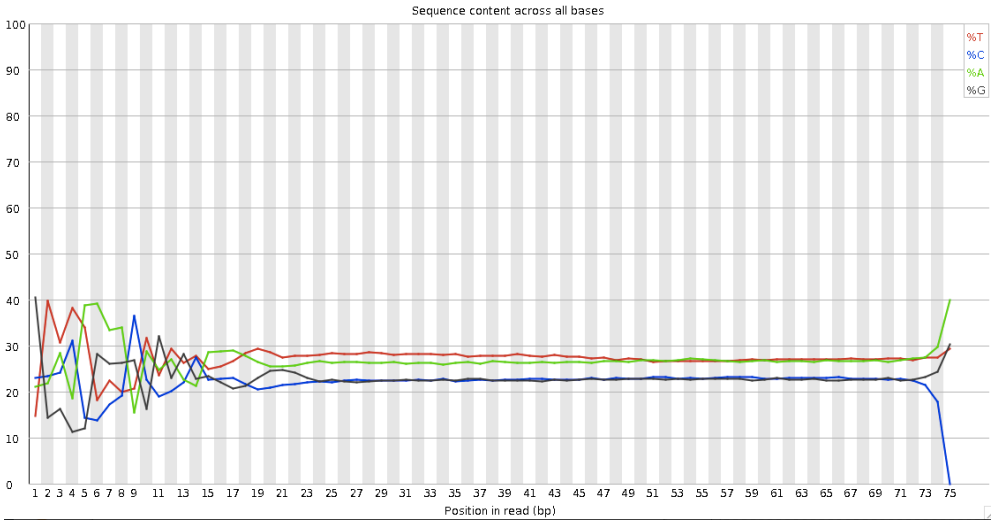


Figure 4.11 Sequence contents across all bases

The x-axis is the base position of each sequence, and the y-axis is the percentage of the four nucleoid bases, i.e. T (red), C (blue), A (green) and G (black). A random library is expected to show little to no differences between different sequences, and the line of each nucleoid base is expected to be parallel to each other (Andrews, 2010).

In order to identify potential fragmentation during sequencing, the sequence length distribution was examined across all the reads in **Figure 4.12**. As described in **Chapter 2**, RNA-seq libraries in this project were sent to the Next Generation Seq Facility at the WIMM, and HiSeq was carried out using 75-base pair-end reads. Consistent with this setting, the library showed highly uniform length of the fragments (~75bp), indicating little degradation and fragmentation.

The information of Phred score distribution across all sequences was extracted and checked in order to examine whether there is a subset of sequences with poor qualities across all samples (**Figure 4.13**). The majority of reads showed high Phred scores of more than 30 (indicating 99.9% base call accuracy) (Ewing and Green, 1998), without specific low-quality subset, confirming no systematic problem in the library preparation and sequencing.

In brief, the samples sent for sequencing passed the quality control, FastQC, and were considered to be of reasonably good quality for downstream analysis.

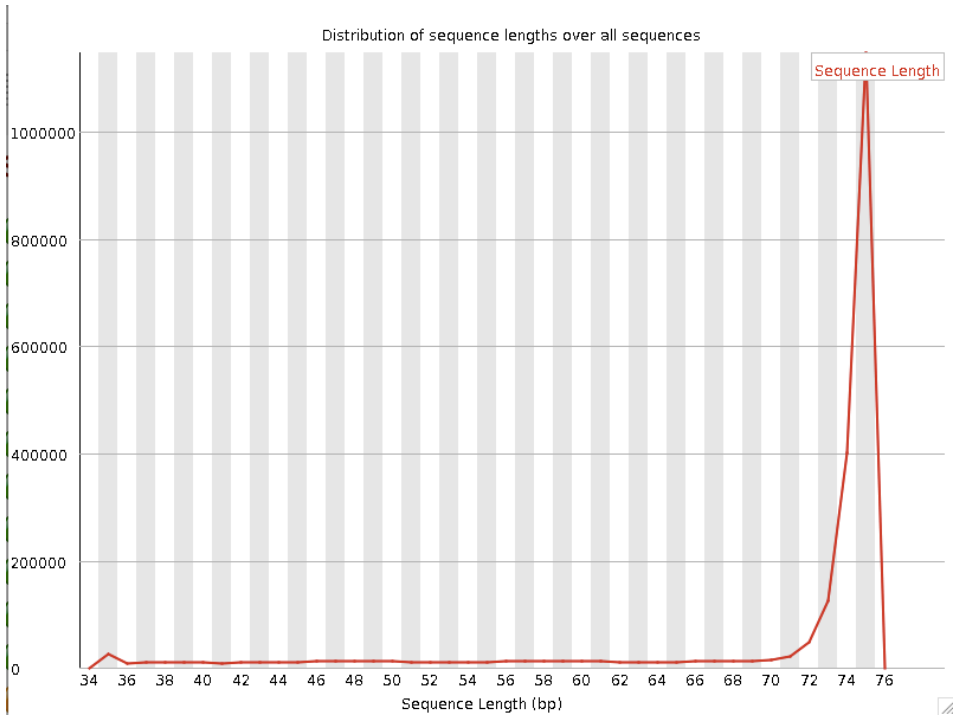


Figure 4.12 Distribution of sequence lengths over all sequences

The x-axis is the length of each sequence, and the y-axis is the counts of sequences. A library with degradation or fragmentation will contain sequences of widely varying lengths.

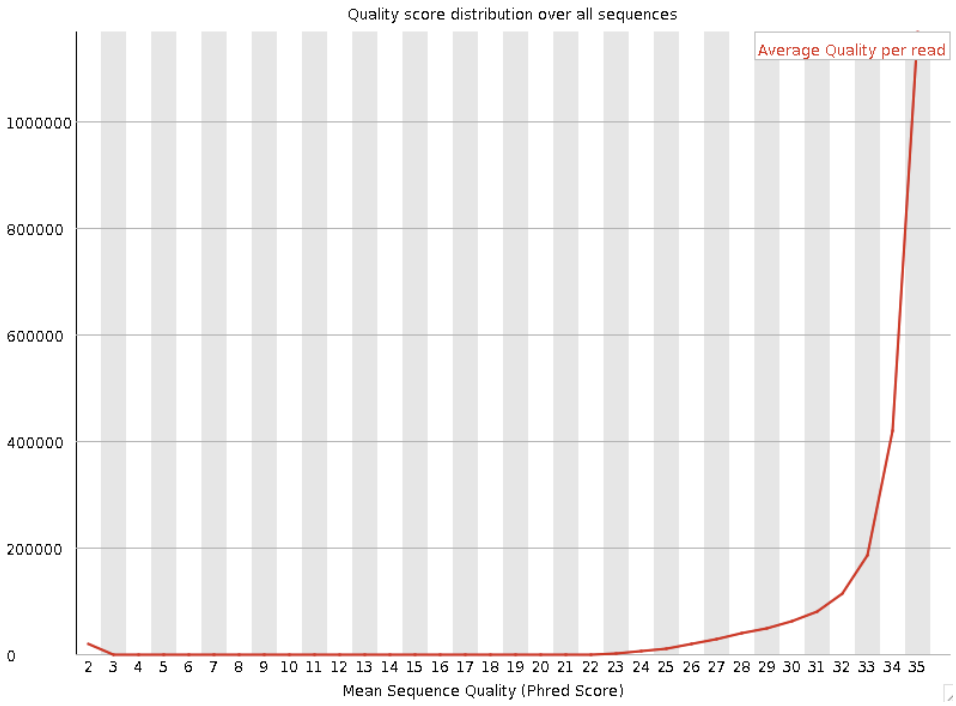


Figure 4.13 Quality score distribution over all sequences

The a-axis is the Phred score (mean sequence quality), and the y-axis is the counts of sequences.

4.2.2.3.2 Mapping with STAR and visualising in the UCSC Genome Browser

The Spliced Transcripts Alignment to a Reference (STAR) software was used to map sequences to the reference genomes (GRCh37, hg19). Compared to other aligners, STAR performs mapping with high accuracy and speed, and is capable of detecting non-contiguous, short-length transcripts. The output files from STAR mapping were in SAM form, and were transformed into BAM files with samtools. The unique mapping reads percentage for all 14 samples was > 80% out of all the reads (**Figure 4.14**), indicating the high sequencing quality and mapping precision.

Before proceeding the analysis further, the expression of several key genes was confirmed by visualising at the UCSC Genome Browser. The sam files were converted into BigWig format using the samtools command before uploading and visualising. In the Genome Browser, the key markers for goblet cells, MUC2, FCGBP and TFF3, along with CA12, the differentially expressed genes between goblet cell-positive and –negative cell lines from microarray analysis, were checked.

Here the representative goblet cell sequencing (STAR_1_S1_Aligned, bottom panel) and non-goblet cell sequencing (STAR_8_S11_Aligned, top panel) were used as an example to show the expression difference (**Figure 4.15**). The expression of MUC2 and FCGBP is high in goblet cells and low in non-goblet cells. On the other hand, TFF3, despite its high expression in goblet cells, has some extent of expression in non-goblet

cells. However, the goblet cells and non-goblet cells have a similarly low expression of CA12, which is consistent with the co-staining pattern of CA12 with PR5D5 (see **Chapter 5**). The expression of these key genes confirms the reliable sequencing quality of the fixed goblet cells and non-goblet cells before further differentially expression analysis.

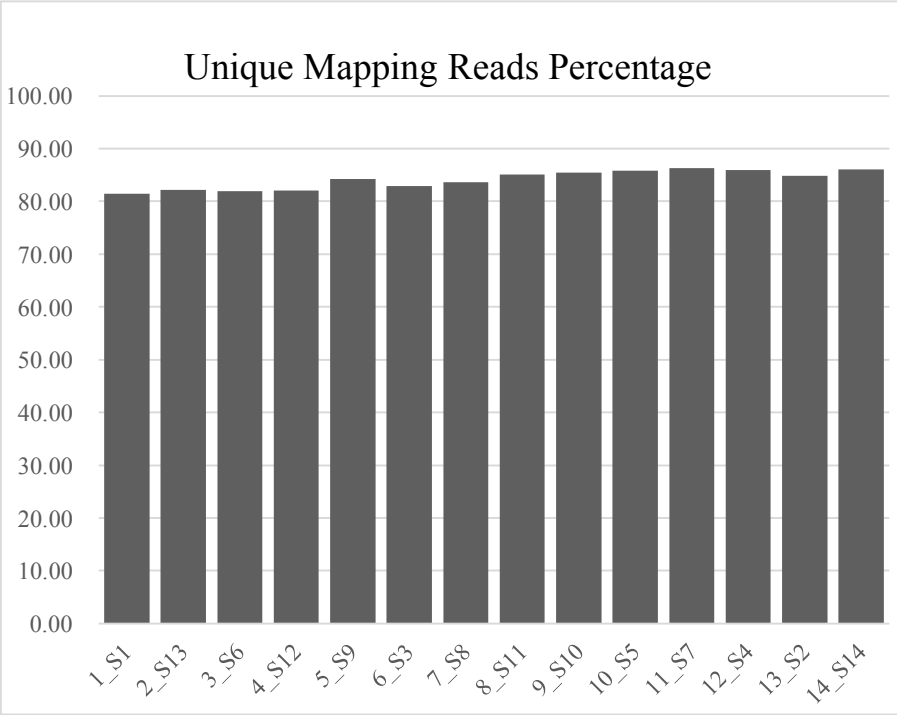
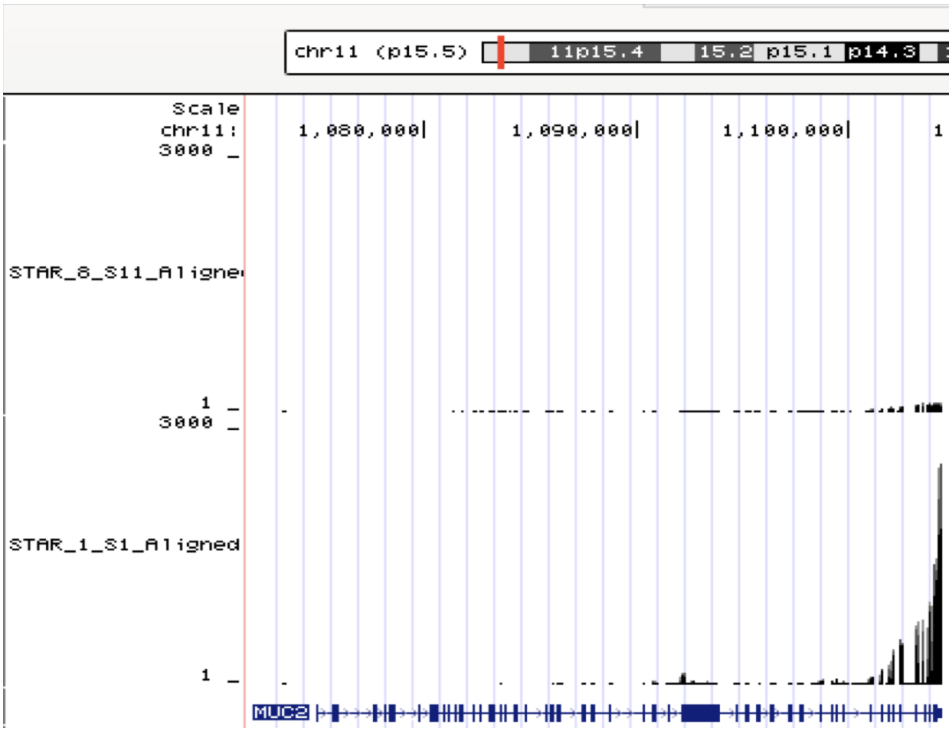
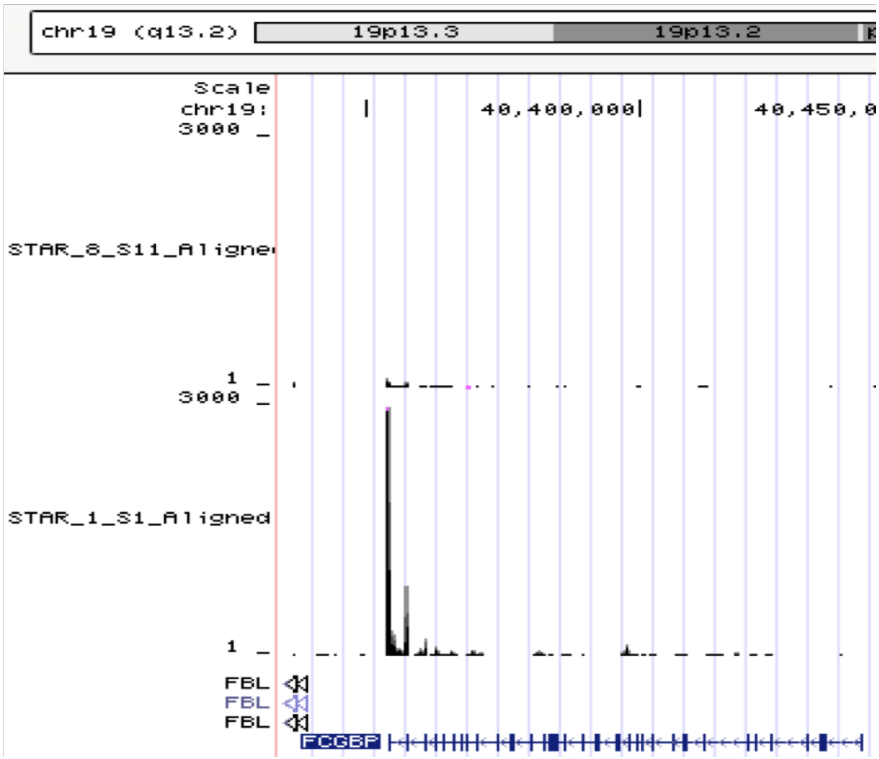


Figure 4.14 Reads are highly uniquely mapped to the reference genome
The percentage was calculated by the number of uniquely mapped reads divided by the number of total input reads.

MUC2


chr11 (p15.5)
11p15.4
15.2
p15.1
p14.3
Scale
chr11:
1,080,000
1,090,000
1,100,000
3000
STAR_8_S11_Aligne
1
3000
STAR_1_S1_Aligned
1
MUC2

FCGBP


chr19 (q13.2)
19p13.3
19p13.2
Scale
chr19:
40,400,000
40,450,0
3000
STAR_8_S11_Aligne
1
3000
STAR_1_S1_Aligned
1
FBL
FBL
FBL
FCGBP

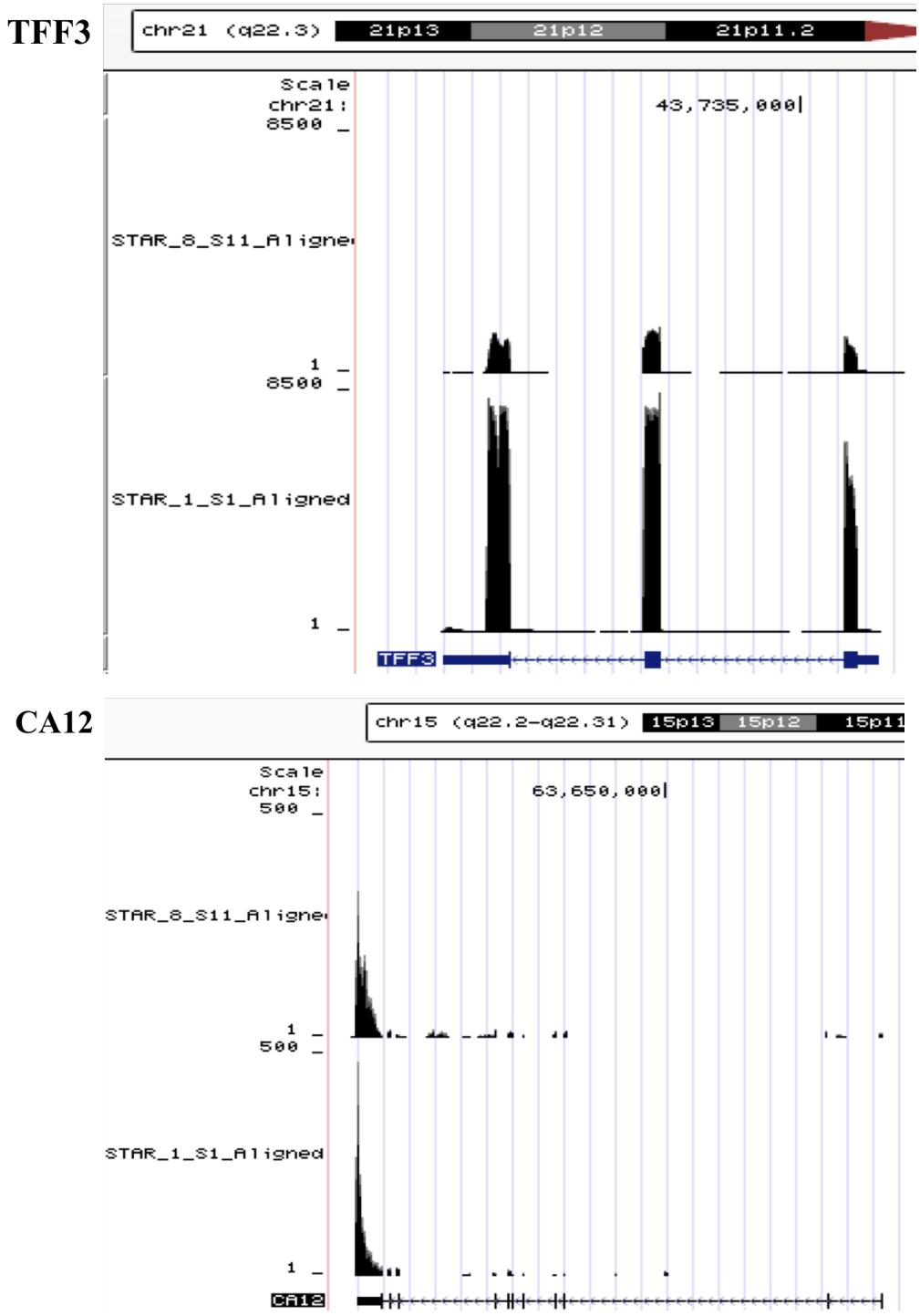


Figure 4.15 Expression of MUC2, FCGBP, TFF3 and CA12 in goblet cells and non-goblet cells. Representative RNA-seq data was uploaded to UCSC Genome Browser for visualization. Top panel is non-goblet cell sample, and the bottom panel is goblet cell sample.

4.2.2.3.3 Feature counting, filtering lowly expressed genes and normalisation

One of the primary tasks in the RNA-seq analysis is to count the number of reads that can map to genes or features. The featureCounts, a read summarization function, was used with Human Genome 19 (GRCh37, hg19) as the reference genome. It took BAM files generated by STAR as input and outputted a text (.txt) file that contained the read numbers that were successfully assigned. The R package edgeR was then imported for the differential gene expression analysis. The assigned count table by featureCounts was used as input, and the DGEList function was used to create the objective that contained read counts, grouping factors and gene annotation.

Genes should be expressed at minimal levels to be translated into proteins or address biological importance. The lowly expressed genes were filtered out in this analysis before further analysis. Specifically, the cpm function (count per million) was used as the threshold to measure gene expression levels. Only the genes with at least 5 cpm in at least 7 out of 14 samples were considered to be expressed and kept for further analysis. The remaining genes were considered to be lowly expressed and filtered out.

Two types of normalisation were conducted in the analysis using edgeR. The first one is the sequencing depth represented by the varying library sizes. This is part of edgeR's basic modelling and can automatically be normalised (Robinson et al., 2010). The second normalisation is the RNA composition. It is a common problem when a subset

of highly expressed genes occupies an extensive proportion of one or a few samples, leaving other genes under-sampled. In this analysis, the calNormFactors function was used for RNA composition normalisation. It normalised the RNA composition by producing the ‘effective library size’ that consists of the original library size and the scaling factor using the trimmed mean values between samples. The scaling factors minimise the inter-sample fold-change for most genes.

4.2.2.3.4 Differential expression analysis

After filtration of lowly expressed genes, the topTags function was used for differentially expression analysis (Mark Robinson; Biostatistics; 2008). This function outputs the data frame that contains log-fold change (logFC), log-average abundance (logCPM), exact p-value for differential expression (PValue) and corrected p-value for multiple comparison (FDR) for each gene in the goblet cell samples and non-goblet cell samples. The differentially expressed genes represent a spectrum of candidate biomarkers for goblet cells.

In order to identify potential variation between samples, the Multi-dimensional scaling (MDS) analysis was conducted for the whole transcriptomes of 14 samples (**Figure 4.16**). The MDS analysis has clearly identified the perfectly clustered non-goblet cell samples. The seven goblet cell samples (blue) are separated from the seven non-goblet cell samples (red) along Leading logFC (log-fold change) dim1. Leading logFC dim2, on the other hand, illustrates the variance within the goblet cell samples. Specifically, the goblet cell sample 2 shows the largest variance compared to others, while the goblet cell samples 4-6 and all non-goblet cells showed little variance (Kruskal and Wish, 1978) (**Figure 4.16**). Thus, to examine the potential influence by the variance within goblet cells, the goblet cell samples 4-6 were selected to compare with the non-goblet cell samples (namely "3vs7") in addition to the analysis of all goblet cell samples against non-goblet cell samples (namely "7vs7").

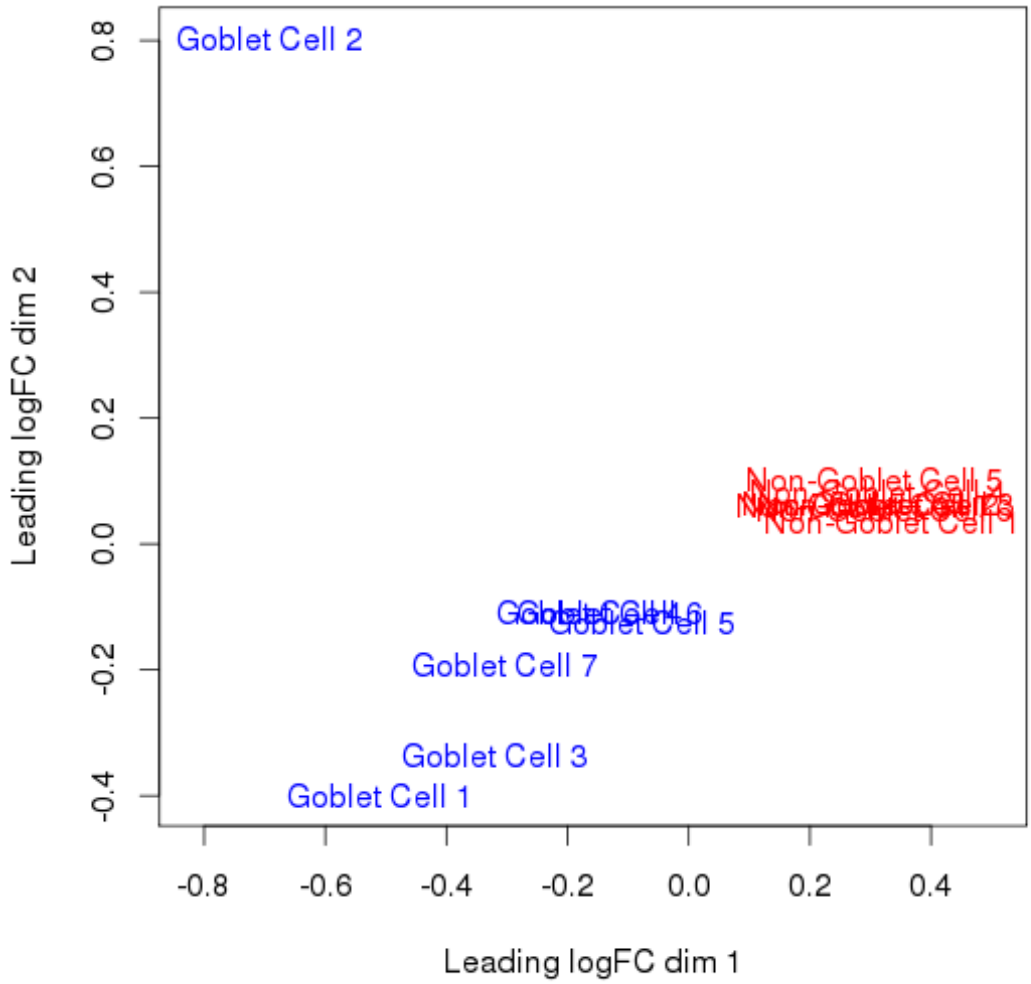


Figure 4.16 Multi-dimensional scale plot of goblet cell and non-goblet cells
The MDS scatter plot compares thousands of variables in the whole transcriptomes of Sample 1-14, and evaluates and represents the observed similarities (distances) between them with two dimensions. Leading logFC dim1 and Leading logFC dim2 separate the goblet cells versus non-goblet cells based on the expressions 12,380 genes after filtration. The RNA from goblet cells (Goblet Cell 1-7, blue) cluster and are separated from the RNA from non-goblet cells (Non-Goblet Cell 1-7). The close distances between Non-Goblet Cells represent the high similarities between the samples.

In order to evaluate the outlier effect, the differentially expressed genes were cross-checked between the 3vs7 and 7vs7 analyses. Among the top 200 highly expressed genes in goblet cells, 160 genes (80%) overlapped between the 3vs7 and 7vs7 analyses (data not shown). Three out of the four top genes with a log fold-change of more than 2 (i.e. CXCL11, DKK1 and APCDD1) in 7vs7 analysis also appeared in the list of the top 22 genes in 3vs7 analysis (data not shown). Taken together, these results indicate the little effect on differential gene expression analysis by the outlier.

The details of top 50 genes in goblet cells and top 15 genes in non-goblet cells are listed in **Figure 4.17** (Detailed up-regulated genes in goblet cells and non-goblet cells can be seen in **Table Appendix.3** and **Table Appendix.4** respectively). Among these genes, all top 50 goblet cell-specific genes showed more than 6.40-time fold-change, while only the top 8 non-goblet cell-specific genes showed more than 4-time fold change. All the differentially expressed genes showed significantly low p-values (smaller than 0.05). In the genes specifically expressed in goblet cells, 49 out of 50 top genes showed p-values smaller than 0.001. In the genes specifically expressed in non-goblet cells, 8 out of 15 top genes showed p-values smaller than 0.001.

	Log2(FC)*	Log2(CPM)	P-Value	FDR
FCGBP	7.43	9.71	3.53E-82	8.38E-79
ATP2A3	7.01	3.71	1.32E-47	6.28E-45
MUC2	6.76	9.60	1.26E-49	7.06E-47
ANO7	6.54	6.31	6.85E-109	3.26E-105
HEPACAM2	5.96	5.62	1.17E-25	2.47E-23
ZG16	5.79	6.69	1.89E-19	2.68E-17
SPDEF	5.76	6.63	1.39E-39	5.08E-37
KLK3	5.65	6.25	1.33E-30	3.60E-28
TRPA1	5.50	4.26	5.28E-20	8.09E-18
GPR153	5.42	4.43	4.10E-46	1.86E-43
LINC00261	5.30	5.99	7.94E-52	6.29E-49
SPINK4	5.15	9.55	3.98E-81	7.57E-78
DEFA6	5.07	6.14	1.04E-05	0.000211767
TFF3	5.04	11.44	5.15E-185	4.90E-181
NEURL	4.78	5.88	3.18E-39	1.12E-36
TMEM61	4.70	4.31	2.83E-27	6.12E-25
TFF2	4.58	6.45	9.78E-16	9.12E-14
KLK15	4.56	4.33	3.60E-16	3.60E-14
ENG	4.29	3.41	9.50E-20	1.39E-17
CBFA2T3	4.22	3.94	2.15E-32	6.39E-30
NXPE1	4.14	3.98	1.80E-15	1.63E-13
SERPINA1	4.05	5.90	1.12E-40	4.24E-38
RAB3B	3.87	4.68	5.46E-18	6.57E-16
TFF1	3.82	8.69	2.29E-70	3.11E-67
RASD1	3.69	4.37	6.55E-20	9.88E-18
RETNLB	3.62	5.50	2.40E-27	5.31E-25
FAM83E	3.59	3.29	3.30E-16	3.33E-14
KLK1	3.54	6.90	6.94E-104	2.20E-100
GSN	3.53	7.28	9.44E-72	1.50E-68
L1TD1	3.46	3.88	4.07E-07	1.22E-05
REG4	3.33	9.39	1.48E-51	9.36E-49
KLK12	3.33	6.26	4.17E-25	8.43E-23
RAB26	3.31	6.10	6.03E-53	5.21E-50
CAPN9	3.31	3.99	2.45E-17	2.80E-15
ANXA13	3.29	4.33	6.16E-10	3.31E-08
MST1	3.27	4.78	1.10E-16	1.18E-14
SELM	3.26	4.74	3.15E-29	8.10E-27
DLL1	3.17	5.13	2.02E-21	3.42E-19
RAP1GAP	3.15	6.97	4.77E-45	2.06E-42
CACNA2D2	3.12	3.83	2.37E-19	3.31E-17
CA8	3.05	3.67	5.37E-09	2.57E-07

WFDC2	3.01	4.90	2.64E-25	5.45E-23
RASD2	3.00	3.88	0.000474374	0.005279216
DLL4	2.96	5.82	1.18E-20	1.91E-18
FOSB	2.94	5.04	2.93E-23	5.26E-21
MLPH	2.85	6.59	4.92E-48	2.46E-45
ALDH1A1	2.80	5.43	1.52E-11	1.06E-09
NTN4	2.79	4.21	1.45E-10	8.61E-09
TP53INP2	2.68	4.27	4.22E-08	1.62E-06
KLF4	2.68	6.95	6.80E-50	4.04E-47
CXCL11	-2.58	5.09	2.61E-15	3.68E-13
LCN2	-2.32	3.35	3.26E-06	9.96E-05
RAC2	-2.24	3.26	3.64E-07	1.52E-05
DKK1	-2.21	5.83	3.66E-06	0.000109511
CDADC1	-2.18	3.31	3.03E-06	9.47E-05
BIRC3	-2.13	3.07	9.38E-05	0.001638434
STOM	-2.09	3.48	0.001387565	0.012887651
LDLRAD3	-2.04	3.38	5.10E-06	0.000144936
CORO1A	-1.89	2.84	0.002092093	0.017372342
C7orf25	-1.74	4.30	0.000112064	0.001898841
RUNDC3B	-1.72	3.13	0.001916391	0.01631861
CYR61	-1.69	3.67	0.000195449	0.002887403
ANLN	-1.69	5.28	3.49E-11	2.87E-09
APCDD1	-1.68	4.04	0.000933871	0.009658422
ARSK	-1.65	4.32	1.33E-05	0.000323325

* Positive values mean highly expressed in goblet cells; Negative values mean highly expressed in non-goblet cells.

Table 4.1 List of differentially expressed genes in goblet cells and non-goblet cells. Genes were sorted based on Log2(FC) from high to low in goblet cell samples. Top 50 goblet-specific genes and top 15 non-goblet cell-specific genes were listed based on expression fold change.

After correction for multiple tests, 39 goblet cell-specific genes showed significant differential expressions, as labelled in the Volcano Plot (**Figure 4.17**). The Volcano Plots presents an asymmetric distribution with a number of genes highly expressed in goblet cells. The red line is drawn at (-log10(FDR)) * logFC = 60 in order to distinguish the genes that are asymmetric in Volcano Plot and highly expressed in goblet cells. Among these genes there are as might be expected, several identified goblet cell markers or mucus components. Importantly, MUC2, as a positive control, showed a108-fold higher expression in goblet cells. FCGBP, TFF3, Zymogen Granule Protein 16 (ZG16) and Serine protease inhibitor Kazal-type 4 (SPINK4) ranked in the top of the list of differentially expressed genes with more than 32-fold higher expression. SPDEF, a transcriptional factor that is proposed to be involved in goblet cell maturation (Gregorieff et al., 2009; Noah et al., 2010), was expressed 32 times higher in goblet cells, indicating its important role in goblet cell differentiation.

The rpkm (Reads Per Kilobase of transcript per Million mapped reads) values of the top genes were listed in **Table Appendix.5** and **Table Appendix.6**. For most goblet cell-specific genes, e.g. ANO7, WFDC2, MUC2, DLL4, ZG16, SELM, RAB26, SERPINA1, NEURL, FCGBP, RETNLB, KLK3, KLK12, MLPH, HEPACAM2, DEFA6, KLF4, KLK15 and SPDEF, the expression level differences between goblet cells and non-goblet cells are quite obvious. The rpkm values are very small, even nearly 0, in the non-goblet cell samples, while generally more than 40 in the goblet cell samples. For some genes, e.g. TFF1, REG4, TFF3, SPINK4 and KLK1, they are expressed at some extent in non-goblet cells (with rkpm values more than 40), while

more than 6-fold higher expression in the goblet cells. For the remaining goblet cell-specific genes, e.g. ATP2A3, CA8, L1TD1 and CBFA2T3, despite the differential expression, their rpkm values are small in both goblet cells and non-goblet cells.

For the top 15 genes in non-goblet cells, in spite of the significant fold-change between both groups, almost all genes are expressed at low levels across all samples. The only exception is DKK1, with rpkm values near 0 in goblet cells and generally more than 40 in non-goblet cells.

This transcritptomic characterisation presents a panel of genes that are selective for goblet cell differentiation, which can potentially act as novel markers or be involved in regulation of goblet cell differentiation.

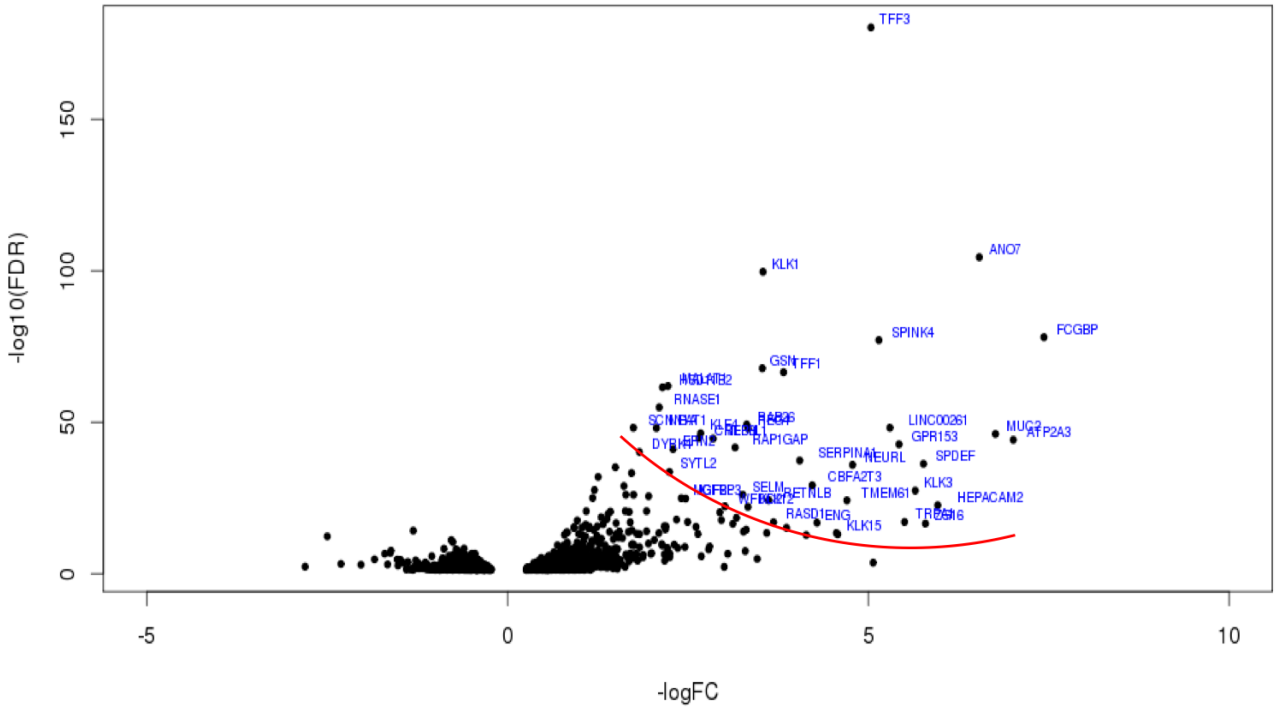


Figure 4.17 Volcano plot of RNA-seq data between goblet cells and non-goblet cells. RNA-seq data of goblet cells and non-goblet cells were plotted based on log2FC (x-axis) and log10 FDR (y-axis). In the x-axis, the right side means the fold change highly expressed in goblet cells, while the left side means the fold change highly expressed in non-goblet cells. The red line is drawn to distinguish these genes, based on the -log10(FDR) (y-axis) and log2FC (x-axis), and represents y-axis * log2FC equals 60. Genes were identified with significant differential expression if -log10(FDR) * log2FC >60.

4.3 Discussion

In this chapter, we first identified permeabilisation as the key RNase contamination step that leads to the decreased RNA concentration during RNA extraction from fixed cells. This highlights the critical time point of applying RNase inhibitors to preserve RNA, which is consistent with previous research (Pan et al., 2011).

Comparing different fixatives, permeabilisation agents and RNase inhibitors, a combination of reagents was chosen to minimise the RNA degradation. In respect to fixing intracellular antigens and preserving RNA integrity, it has been reported that methanol, ethanol and acetone are more efficient than PFA (Medeiros et al., 2007; Cox et al., 2006; Goldsworthy et al., 1999). However, the research of Pan and colleagues suggests that 4% PFA is a better choice regarding RNA preservation (Pan et al., 2011). In this study, cells fixed with 90% methanol demonstrated a higher proportion of positive PR5D5 staining, while 4% PFA with 0.1% saponin or 0.1% Triton gave higher RNA yields with improved purity.

Based on the evaluation, this chapter demonstrates the establishment of a novel protocol for high quality RNA isolation for sequencing after fixation, permeabilisation, intracellular staining and FACS sorting. This presents several important improvements compared to current methods. Firstly, it advances FACS application to investigate the transcriptome of a specific subset of cell lineages without the priori requirement of

surface markers. Secondly, compared to previous protocols, the optimised procedure requires a small amount of RNase inhibitor, therefore is more cost efficient. The volume of RNase inhibitor used in this protocol is even lower compared to a recent protocol for fixed single cell RNA extraction (Thomsen et al., 2016). Thirdly, the optimised protocol requires only 100 cells to extract RNA, decreasing the sorting time and potential risk of RNA degradation. This will enhance its application in characterising a specific subpopulation using only intracellular markers, or combining with genetic reporters.

Utilising this protocol, the total RNA was successfully isolated from fixed goblet cells of comparable quality with unfixed cells. As an important factor in defining a reliable sequencing, a reasonable sequencing depth is considered to lead to an appropritate number of detected transcripts (Mortazavi, et al., 2008). Although there has not yet been a conclusive number of mapped reads required for the sequencing reliability, some researchers suggest that accurate quantification of the medium to significantly expressed genes should be achieved using 5 million reads (Sims, et al., 2014; Mortazavi, 2008), and in single cell study, even as few as 20,000 reads were used for cell type differentiation in splenic tissue (Jaitin, et al., 2014). In this study, due to the limited cell type complexity in the highly enriched goblet cells, in total 14 cDNA libraries (7 goblet cells and 7 non-goblet cells) were sent for RNA sequencing, with the sequencing depth of 5-20 million mapped reads for each sample. After quality control and differential expression analysis, the first transcriptomic profile of goblet cells was characterised, and the differentially expressed genes between goblet cells and non-goblet cells in the human colorectal cancer cell line LS180 were identified.

This analysis highlights the goblet cell markers and potential transcriptional regulators that are involved in goblet cell differentiation and maturation.

Three differential expression patterns were identified in the goblet cell and non-goblet cell populations (see rkpm values in **Table Appendix.5** and **Table Appendix.6**). Firstly, genes including MUC2, FCGBP, DEFA6 and SPDEF, were exclusively expressed in goblet cells with little or no expression in non-goblet cells. MUC2 and FCGBP are known to be well-characterised goblet cell markers (KMAJ Tytgat et al., 1994; DK Podolsky et al., 1984) and their expression in goblet cells served as the positive control of the sequencing.

Interestingly, DEFA6, a Paneth cell marker (Bevins et al., 2011; Shi, 2007), is highly expressed in the goblet cells in the cell line LS180. Paneth cells are usually differentiated in the normal small intestine – the extent varies between individuals, but are not found beyond the transverse colon (Porter, 2002; Ayabe, 2004). Paneth cell metaplasia in human colons can be observed following longstanding inflammation with crypt architectural distortion (Surawicz et al. 1994). The high expression of DEFA6 in the PR5D5-positive goblet cells may indicate an interesting idea that Paneth cells can derive from, presumably, other epithelial cells. This raises the immediate question of its differentiation in colorectal cancers, and how it is induced by the inflammatory effects.

SPDEF was reported to regulate goblet cell hyperplasia in lung epithelium (Park et al., 2007). It is regulated by another important transcriptional factor ATOH1 (Lo et al., 2017), and the inhibited SPDEF expression can repress the expression of goblet cell specific genes including AGR2, MUC2 and SPINK4 (Noah et al., 2010; Gregorieff et al., 2010; Aronson et al., 2014). The fact that SPEDF is highly expressed in the purified goblet cells is consistent with its previously described roles in goblet cells (Noah et al., 2010), and will be further investigated in **Chapter 5**.

The second pattern of gene expression, such as for TFF3, is that genes are expressed in both goblet cells and non-goblet cells, but to a much higher extent in goblet cells. TFF3, a 59-amino-acid intestinal trefoil factor as part of secreted mucus, was demonstrated as a goblet cell marker, (Kim and Ho, 2010; Durual et al., 2005). It was described to have an initiative role of mucosal healing to maintain the gastrointestinal integrity (Taupin, et al. 2003) and its synthesis was stimulated by TLR2 (Podolsky, et al. 2009). Its expression patterns and regulation by other key transcriptional factors will be discussed in **Chapter 6**.

The third expression pattern lies in the genes that are highly expressed in non-goblet cells. Although the differentially expressed genes in non-goblet cells present high fold-changes and low p-values, the absolute levels of most gene expressions are very low. The only exception is DKK1, with the rpkm values more than 40 in non-goblet cells. DKK1 is described as a Wnt antagonist which negatively regulates the Wnt pathway

by inhibiting the interaction between LRP5/6 and Wnt (Kuhnert, et al. 2004). It is also a target of the beta-catenin/TCF pathway (Niida, et al. 2004). There is no published evidence suggesting its function in goblet cell differentiation. Its main functions may just be in WNT activation if not compatible with goblet cells.

To summarise, this characterisation of goblet cell transcriptional profile described in this chapter serves as a foundation to discover potential goblet cell markers, and to define transcriptional network of goblet cell differentiation and maturation.

CHAPTER 5

INVESTIGATION OF KEY GENES IN GOBLET CELL DIFFERENTIATION

5.1 Introduction

The capacity of single cell derived colonies to form lumens can be used as an indicator to characterise cancer stem cells in the lumen-forming cell lines (Ashley et al., 2013; Yeung et al., 2010). The association between lumen formation and goblet cells, if there is any, has yet not been characterised. It will help to understand the multipotency and the lineage commitment of cancer stem cells by assessing goblet cells in lumens.

In the previous chapters, goblet cell differentiation at the protein level was screened in a panel of 64 human colorectal cancer cell lines. Based on the screening, microarray analysis of bulk populations and the transcriptomic profile of enriched goblet cells highlight the important genes as potential goblet cell markers or regulators, including in particular TFF3, SPDEF and CA12.

TFF3 serves as a specific marker for goblet cells in human colons, playing an important role in the mucosal repair and regeneration processes (Hoffman et al., 2001). Together with MUC2 and FCGBP, TFF3 is secreted into the lumen as a constituent of the mucus layers (Johansson et al., 2009). Increased expression of TFF3 was observed after the injury in colonic tracts (Mashimo et al., 1996). Mice deficient in TFF3 presented the destructed mucosal healing and extensive colitis after epithelial injury (Mashimo et al., 1996). The expression of TFF3 and MUC2 is not regulated co-ordinately in rat

intestines (Matsuoka et al., 1999), while the co-staining patterns and exact mechanisms of their co-regulation in human colorectal cancer cell lines remain elusive.

Another top gene from the RNA-seq data is SPDEF, a critical transcriptional factor in goblet cell maturation. In the mouse small intestine, SPDEF depletion results in the increased number of secretory progenitors at the cost of goblet and Paneth cells (Gregorieff et al., 2009). A recent ChIP-Seq analysis in mouse small intestinal crypts identified the core promoter of SPDEF is directly bound by ATOH1 (Lo et al., 2017) which is a key mediator in secretory lineage differentiation downstream Notch pathway (Yang et al., 2001; Shroyer et al., 2007; Shroyer et al., 2013). In mouse small intestines, ATOH1 is essential for all secretory cell lineage commitment (Yang et al., 2001). Thus, it is rational to hypothesise that ATOH1 and SPDEF co-operatively regulate goblet cell differentiation in human colorectal cancer cell lines.

CA12, a transmembrane carbonic anhydrase, was identified to be differentially expressed in goblet cell-positive cell lines from the bulk microarray analysis. In the normal colon, prominent polarised CA12 staining was restricted to the basolateral plasma membrane of enterocytes, especially at the surface cuff areas, while it is absent in small bowels (Kivelä et al., 2000). In colorectal cancers, however, increasing staining of CA12 was detected in the deeper part of the lesion (Kivelä et al., 2000; Kivelã et al., 2001). It is interesting to investigate how CA12 is involved in colorectal carcinogenesis or goblet cell differentiation.

Chapter 5 investigates the potential association between lumen formation and goblet cell differentiation under 3D conditions. In addition to the microarray analysis and RNA-seq differential expression analysis, the expression patterns of TFF3 in human colorectal cancer cell lines were characterised. The co-operative transcriptional regulation on goblet cell-specific genes via key transcriptional factors, ATOH1 and SPDEF was also outlined. Finally, CA12, one of the top genes from the microarray analysis, was characterised and may serve as a novel marker for goblet cell precursors.

5.2 Results

5.2.1 Association between goblet cell differentiation and lumen formation

Based on the screening results presented in **Chapter 3**, goblet cell differentiation of a panel of 64 cell lines was compared to the preliminary lumen formation data that were previously obtained in the Cancer and Immunogenetics Laboratory, University of Oxford (**Table 5.1A**). In 3D culture of goblet cell positive cell lines, 5 were identified as lumen forming cell lines and 10 were unable to develop lumens. Notably, CX1, HT29, KM2012 and WIDR are replicated cell lines with the same origin. Four goblet cell-negative cell lines show the lumen formation, while 20 others cannot.

The number of cell lines of each categorisation was summarised ('Observed cell line counts' in (**Table 5.1B**). The replicated cell lines: CX1, HT29, KM2012 and WIDR, were counted as one to avoid over-representation in the association analysis, thus only 7 instead of 10 goblet cell+/lumen formation- cell lines. Chi-squared analysis was used to test the association between goblet cell differentiation and lumen formation in a 2-way table. The results were visualised across all the cell lines in the 2x2 table. With 1 degree of freedom, the Pearson's Chi square is 2.667 (p-value = 0.10247), smaller than the critical value (for 95% significance) 3.84. And the Fisher exact p-value is 0.22. These analyses do not support the existence of any association between lumen formation and goblet cell differentiation.

A

	Lumen +	Lumen -		
Goblet cell +	CL40 HCA46 LS180 RW7213 SW1222	C125PM CX1 HDC114 HDC73 HT29	KM2012 LOVO LS174T NCIH508 WIDR	
Goblet cell +/-	C106 C80 Caco2 HT55 LS1034	C84 CAR1 GP2D HDC111 HDC57 HDC82 JHCOLOY1	LIM1215 LIM1863 LS123 PMFKO14 RCM1 RW2982 SKCO1	SNUC2B SW480 SW837 SW948 T84 VACO10MS
Goblet cell -	C10 HRA19 OXCO1 OXCO3	C2BBe1 C32 C99 CC20 COLO201 COLO320DM COLO678	DLD1 GP5D HCA7 HCT116 HDC142 JHSKREC LIM2405	OUMS23 RKO SW1417 SW403 SW48 TITTKB

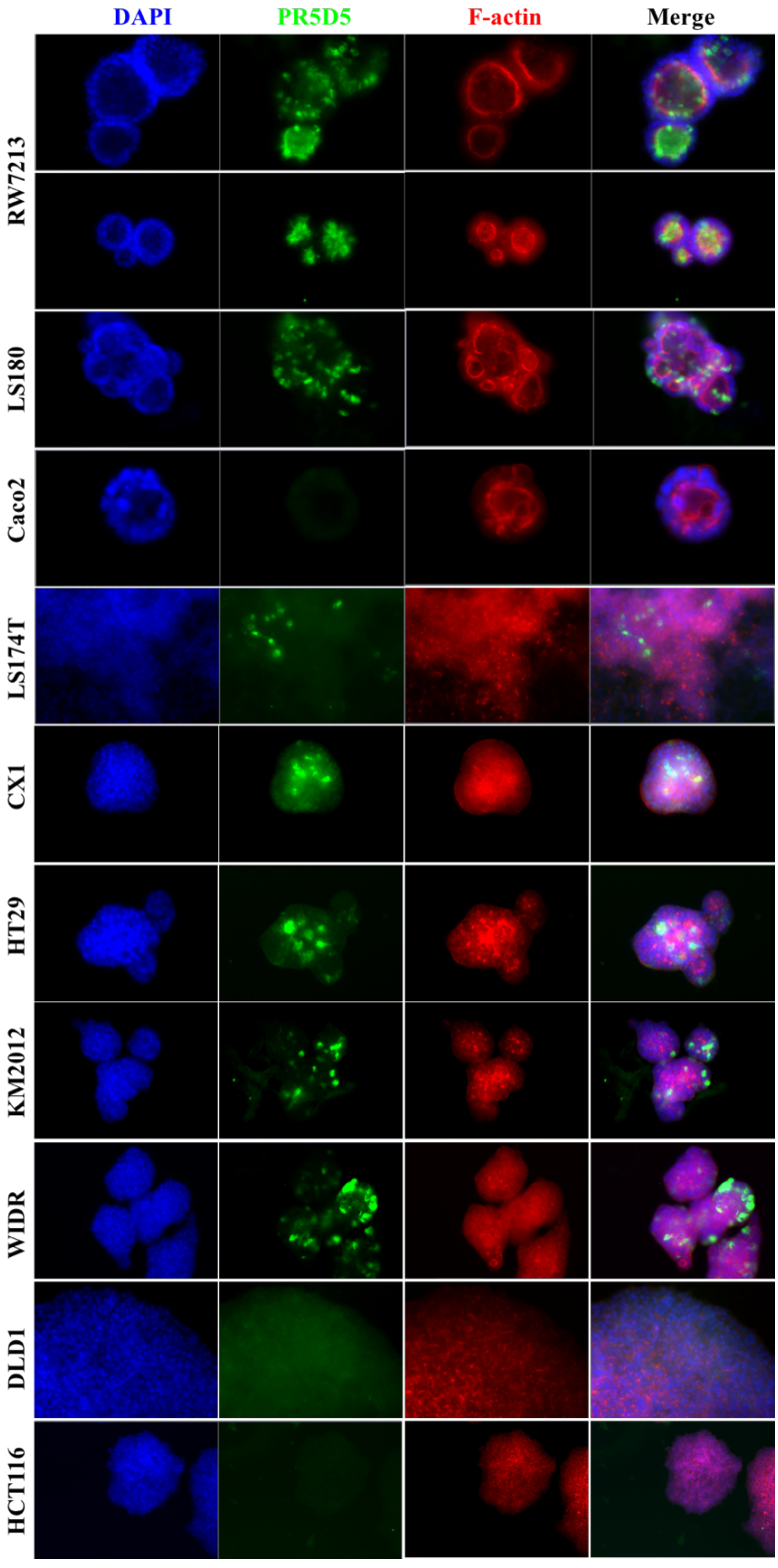
B Observed cell line counts

	Lumen formation +	Lumen formation -	Total
Goblet cell +	5	7	12
Goblet cell -	4	20	24
Total	9	27	36

Table 5.1 Association analysis of goblet cell differentiation and lumen formation
(**A**) Summary of goblet cell differentiation and lumen formation in 64 human colorectal cancer cell lines. (**B**) Observed cell line counts of goblet cell differentiation and lumen formation were summarised.

To further confirm the lumen formation and goblet cell differentiation, single cell suspensions of representative cell lines were seeded in Matrigel and cultivated for two weeks to allow 3D growth and lumen formation. F-actin was used to label membranes of the polarised cells towards the central lumen (**Figure 5.1**). For RW7213 and LS180, lumen-forming colonies were observed. The expression of MUC2 was identified to be within the apical front of goblet cells and accumulate in lumens through both PR5D5 and MUC2-D staining. Lumen formation was observed in Caco2 but MUC2 production was not evident on immunohistochemistry. In contrast, in the cell lines LS174T, CX1, HT29, KM2012 and WIDR, MUC2 production was observed within goblet cells but no visible lumens had formed. In DLD1 and HCT116, there were neither lumen nor goblet cells. These four staining patterns confirm the existence of lumen forming cell lines with no goblet cell differentiation, and the goblet cell-positive cell lines with no lumens. Thus, further gene candidates from microarray and RNA-seq analysis are required to reveal the transcriptional regulation of the differentiation processes towards goblet cells.

A


DAPI
PR5D5
F-actin
Merge
RW7213
LS180
Caco2
LS174T
CX1
HT29
KM2012
WIDR
DLD1
HCT116

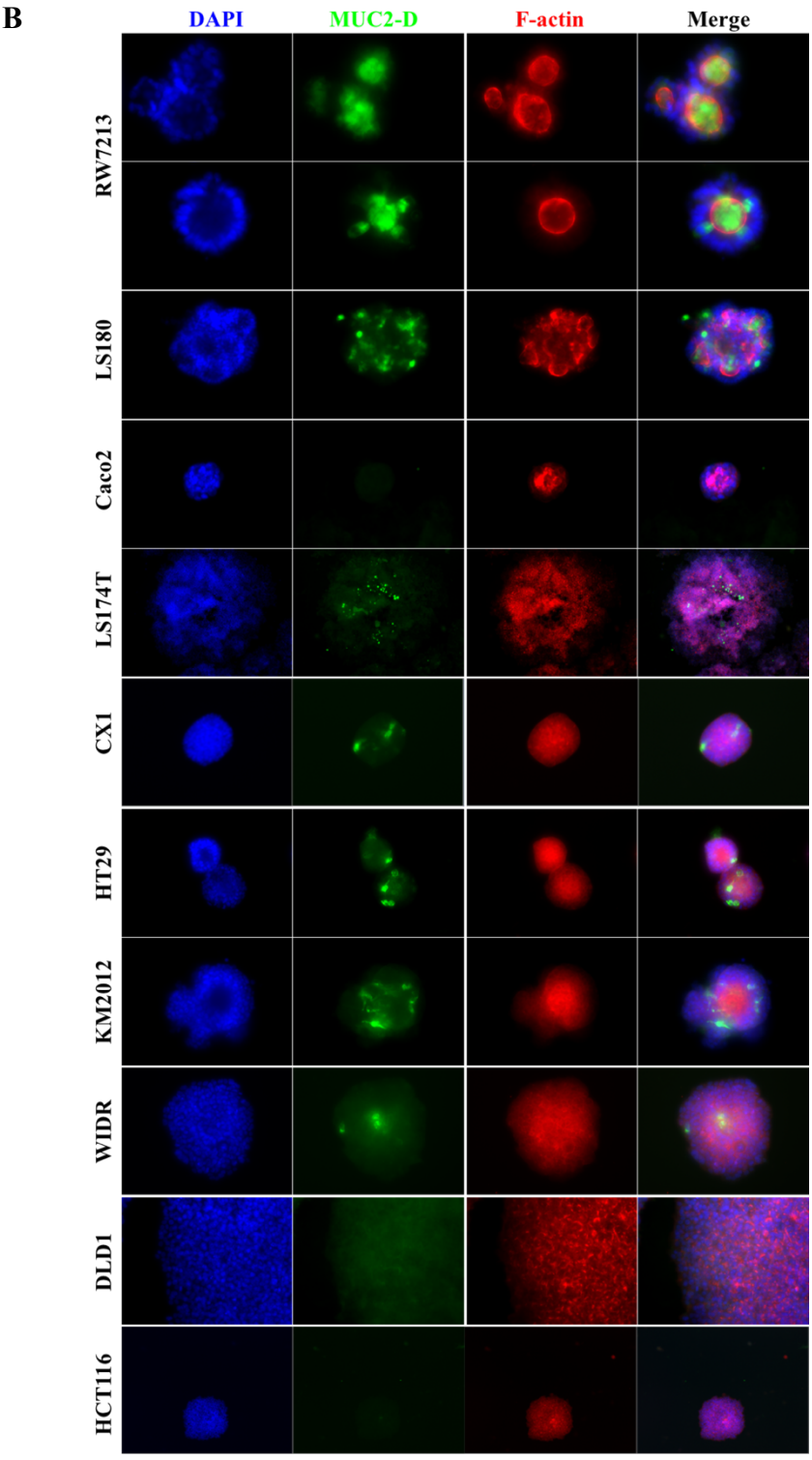


Figure 5.1 Representative staining with PR5D5 and MUC2D under 3D culture
Single cell suspensions (500 cells per well) of human colorectal cancer cell lines were embedded in Matrigel and grown for 14 days before being fixed and labelled with DAPI (blue, 1:5000), PR5D5 (**A**) or MUC2D (**B**) (green, 1:200) and Phalloidin (red, 1:10000) for F-actin staining.

5.2.2 TFF3 identifies the goblet cells that cannot produce MUC2 in human colorectal cancer cell lines

5.2.2.1 Expression of TFF3 in colorectal goblet cells

In **Figure 5.2**, mRNA expression of TFF3 was detected at high levels, with rpkm values more than 4000, in all goblet cells (sample 1-7), whereas, in non-goblet cells (sample 8-14), TFF3 was expressed to a moderate extent with rkpm values between 100 to 200 (**Table Appendix.5**). This suggests that although TFF3 is highly accumulated in goblet cells, the remaining population in colorectal cancer could also express the gene to some extent, indicating possible functions not restricted to goblet cells.

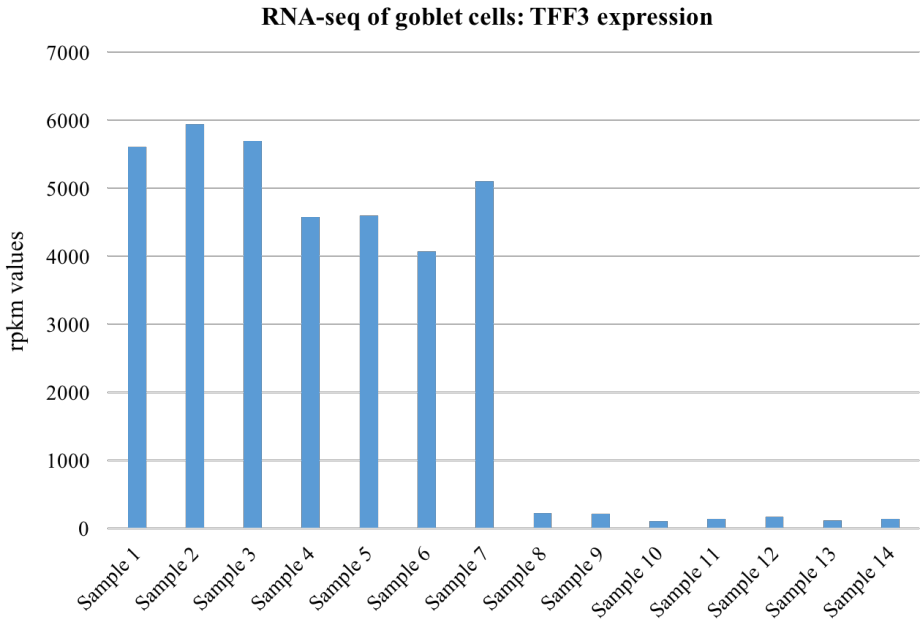


Figure 5.2 RNA-seq rpkm values of TFF3 in goblet cells and non-goblet cells
Expression levels of TFF3 were represented by the rpkm values of goblet cells (Sample 1-7) and non-goblet cells (Sample 8-14). In goblet cells, TFF3 is highly expressed with rpkm values > 4,000. In non-goblet cells, TFF3 is moderately expressed with rpkm values 100-200.

5.2.2.2 TFF3 co-stains the same goblet cells identified by PR5D5 antibody

TFF3 and PR5D5 were co-stained in the human colorectal cancer cell lines LS180 and SW1222, with or without the DBZ treatment for Notch blockade (**Figure 5.3**). At the basal level, TFF3 (red) and PR5D5 (green) stained the same cells in LS180 and SW1222, indicating specific expression of TFF3 within goblet cells. When treated with DMSO as control, the proportion of both PR5D5- and TFF3- positive cells remained the same, and the co-staining patterns did not change. When cells were treated with DBZ to block the Notch pathway, both LS180 and SW1222 showed a significant increase of PR5D5 and TFF3 staining, reflecting the up-regulated goblet cell differentiation in response to the Notch blockade. This is consistent with previous studies describing that Notch inhibition can turn the proliferative cells from normal crypts or adenomas into goblet cells (van Es et al., 2005; Milano et al., 2004). With DBZ treatment, PR5D5 and TFF3 still target the same cells in LS180 and SW1222.

Collectively, co-staining with PR5D5 confirmed the expression of TFF3 in goblet cells at the protein level. The expression of TFF3 is up-regulated co-ordinately with PR5D5 when the Notch pathway was blocked.

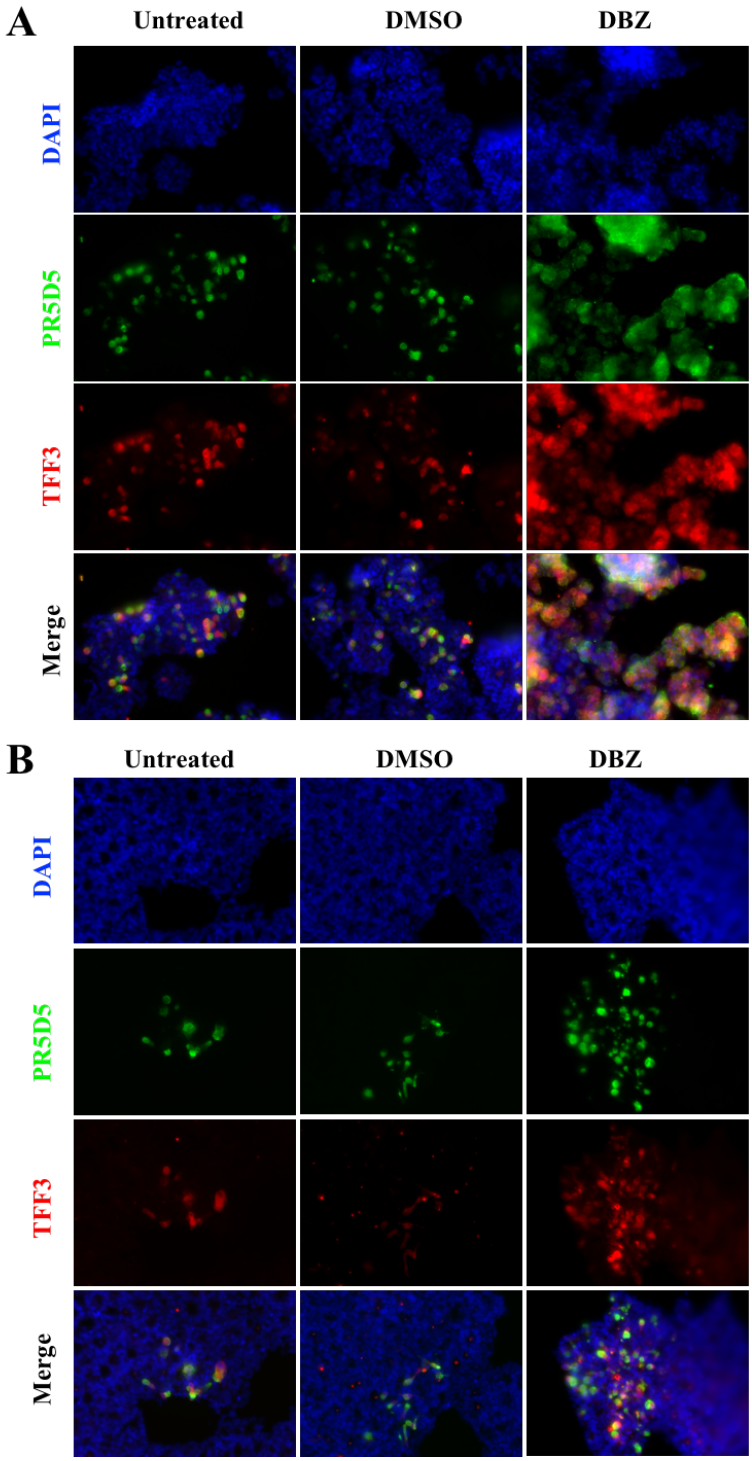


Figure 5.3 TFF3 and PR5D5 co-staining in LS180 and SW1222 under Notch blockade

5,000 cells of human colorectal cancer cell line LS180 (**A**) and SW1222 (**B**) were seeded into each well of 96-well plates. The next day, medium was changed in the untreated sample, and cells were treated with 200nM DMSO or DBZ in corresponding wells. Cells were allowed to grow for another five days before fixation and intracellular co-staining with PR5D5 (green, 1:200), TFF3 (red, 1:200) and DAPI (blue).

5.2.2.3 TFF3 identifies presumed goblet cells or goblet cell precursors that produce no MUC2 protein

The expressions of TFF3 and MUC2 were examined in the normal colonic crypts in a panel of eight colorectal carcinoma cell lines co-stainined with PR5D5 and TFF3 antibodies (**Figure 5.4**). In the normal crypt, PR5D5 and TFF3 stained exactly the same goblet cells along the epithelial layer. In human colorectal cancer cell lines, four co-staining patterns of PR5D5 and TFF3 were observed. In the first pattern, represented by LS180 and SW1222, TFF3 and PR5D5 co-stained the same goblet cells, but the proportion of goblet cells was largely reduced compared to normal colonic crypts. In the second pattern, e.g. RW7213, TFF3 was widely expressed in almost all cells, while stronger staining intensity was observed in the cells targeted by PR5D5. The third pattern, represented by the three colorectal cancer cell lines SW403, JHCOLOY1 and PMFKO14, showed high expression of TFF3 but no reactivity with PR5D5 antibody. The fourth pattern showed double negative immune-reactivity with TFF3 and PR5D5, e.g. HCT116 and SKCO1.

The four co-staining patterns indicate that TFF3 can serve as a marker to identify goblet cells, or goblet cell precursors that produce no MUC2 in human colorectal cancer cell lines, and so presumably also in tumours in patients.

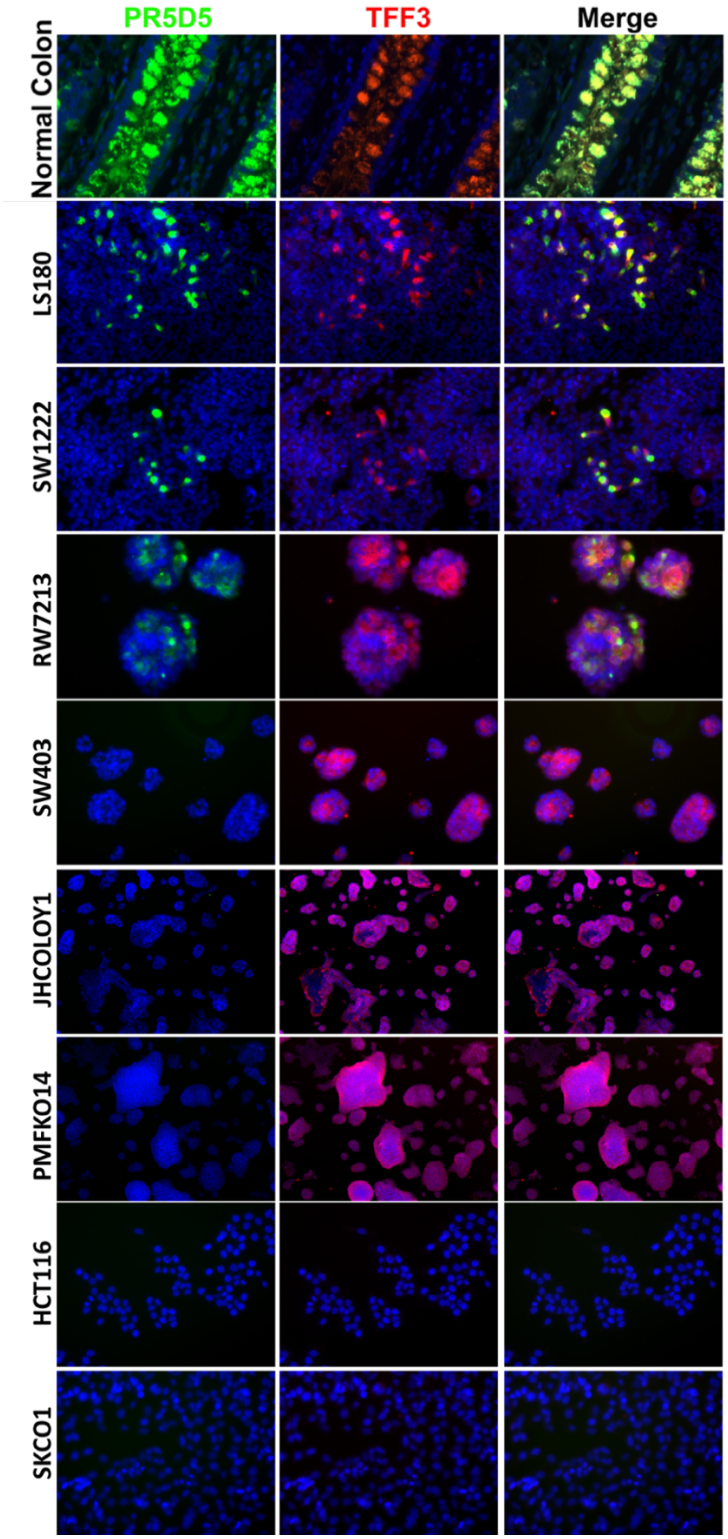


Figure 5.4 TFF3 and PR5D5 co-staining in normal crypts and colorectal cancer cell lines

Human colonic crypts and colorectal carcinoma cell lines were co-stained with PR5D5 (green, 1:200), TFF3 (red, 1:200) and DAPI (blue). The picture of normal colonic crypt was taken by Dr Neil Ashley (Weatherall Institute of Molecular Medicine, University of Oxford).

5.2.3 Investigation of ATOH1 and SPDEF in regulating goblet cell differentiation

ATOH1 is suggested to play a central role to link the Notch pathway to the downstream differentiation of all secretory lineages in the mouse small intestine (Lo et al., 2017). SPDEF, on the other hand, promotes goblet cell maturation in mouse small intestines. Loss of SPDEF leads to the accumulation of secretory progenitors (Gregorieff et al., 2009). Despite their important functions, the expression of the genes *ATOH1* and *SPDEF* have not yet been evaluated in the specific subset of goblet cells in human colorectal carcinomas. In addition, their mechanism of how they work together in controlling expression of MUC2 and other goblet cell-specific genes remains unclear. Thus, it is interesting to further investigate the potential co-operative regulation of this transcriptional triangle on goblet cell differentiation in colorectal cancer cell lines.

5.2.3.1 SPDEF, but not ATOH1, co-expresses with MUC2 in colonic goblet cells

The expression of SPDEF and ATOH1 was examined in the RNA-seq data of goblet cells and non-goblet cells (**Figure 5.5A**). SPDEF, as already mentioned, is one of the top genes with significant differential expression between goblet cells and non-goblet cells. It is highly expressed in goblet cells with rpkm values between 70 to 150 (sample 1-7), while it showed almost no expression in non-goblet cells with rpkm values smaller than 1 (sample 8-14). On the other hand, the expression of ATOH1 is generally low in both goblet cell samples 1-7 (rpkm values 1 – 20) and non-goblet cell samples 8-14 (rpkm values 0.5 – 10). The differential expression of ATOH1 between goblet cells and non-goblet cells is not as clear as SPDEF.

The expression of SPDEF and ATOH1 was further confirmed at the protein level by co-staining with PR5D5 in the human colorectal cancer cell line SW1222 - If time had permitted, qRT-PCR would have been performed to further confirm ATOH1 and SPDEF expression. Staining in the nuclear compartment, SPDEF labelled the same goblet cells as PR5D5 (**Figure 5.5B**). However, the proportion of the cells that positively react with ATOH1 is much higher than with PR5D5, and the two antibodies did not necessarily target the same cells (**Figure 5.5C**). The potential difference regarding protein and mRNA levels for ATOH1 in **Figure 5.5C** might result from the nature of the differences between ATOH1 mRNA and protein turnover. The staining for SPDEF is consistent with the RNA-seq data, and confirms the specific expression of *SPDEF* in goblet cells. The expression and staining patterns indicate that ATOH1 functions upstream during secretory lineage commitment, while SPDEF regulates goblet cell differentiation at a later stage.

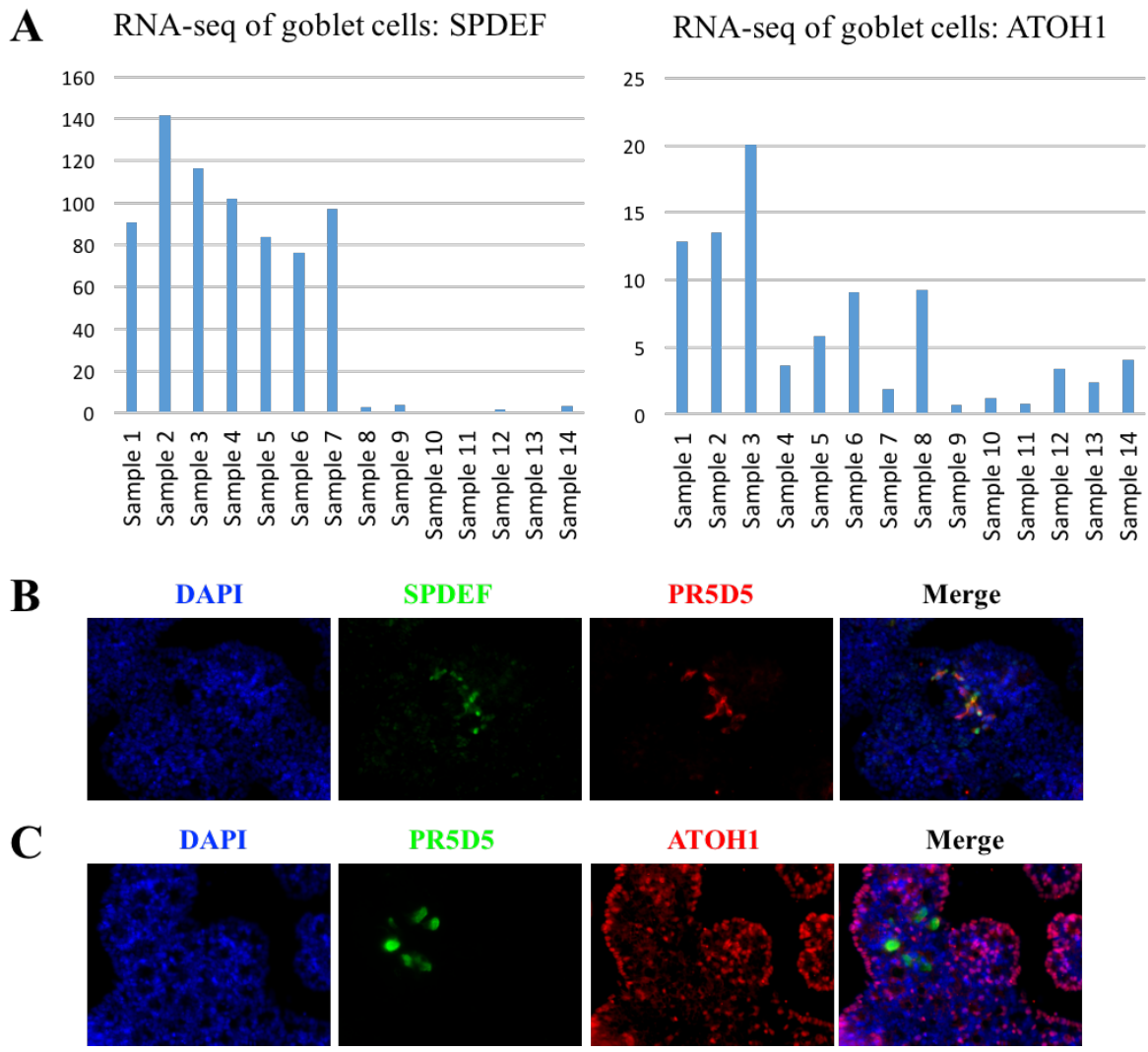


Figure 5.5 Expression and staining of SPDEF and ATOH1

(**A**) Expression levels of SPDEF and ATOH1 were represented by the rpkm values of goblet cells (Sample 1-7) and non-goblet cells (Sample 8-14). SPDEF is highly expressed with rpkm values > 4,000 in goblet cells and lowly expressed with rpkm values < 1 in non-goblet cells. The expression of ATOH1 is low in both goblet cells (rpkm values 1-20) and non-goblet cells (rpkm values 0.5-10). (**B-C**) Immunostaining of SPDEF and ATOH1 with PR5D5 in human colorectal cancer cell line SW1222.

5.2.3.2 SPDEF is downstream regulated by ATOH1

LS180 cells were chosen to study the effects of ATOH1 and SPDEF knock down because of its high levels of SPDEF and ATOH1, as well as the high proportion of goblet cells and high growth rate. Cell viability was analysed by flow cytometry using the LIVE/DEAD® Fixable Dead Cell Stain Kits after RNA silencing (**Figure 5.6A**). Little difference regarding viability was observed when the cells were treated with siRNA at a concentration of 50nM or smaller. When cells were treated with siRNA at the concentration of 100nM, the viability of cells started to decrease. Notably, the decreased viability is likely to be a non-specific effect with the increased siRNA concentration as also shown in the scrambled siRNA. Thus, 50nM siRNA was used to knock down ATOH1 and SPDEF in the further experiments.

The knock-down kinetics were conducted over a period of 120 hours using siRNA against ATOH1 and SPDEF (**Figure 5.6B**). The proportion of PR5D5-positive cells started to decrease after siRNA treatment, and a 48-hour treatment was determined to be the time point giving the maximum knock-down effects, although the ATOH1 knock-down efficiency is clearly better than that for SPDEF. From 72 hours, the proportion of PR5D5-positive cells increased along with time. Thus, the time point of 48 hours was used to culture the cells after siRNA knock down against ATOH1 and SPDEF in our experiments.

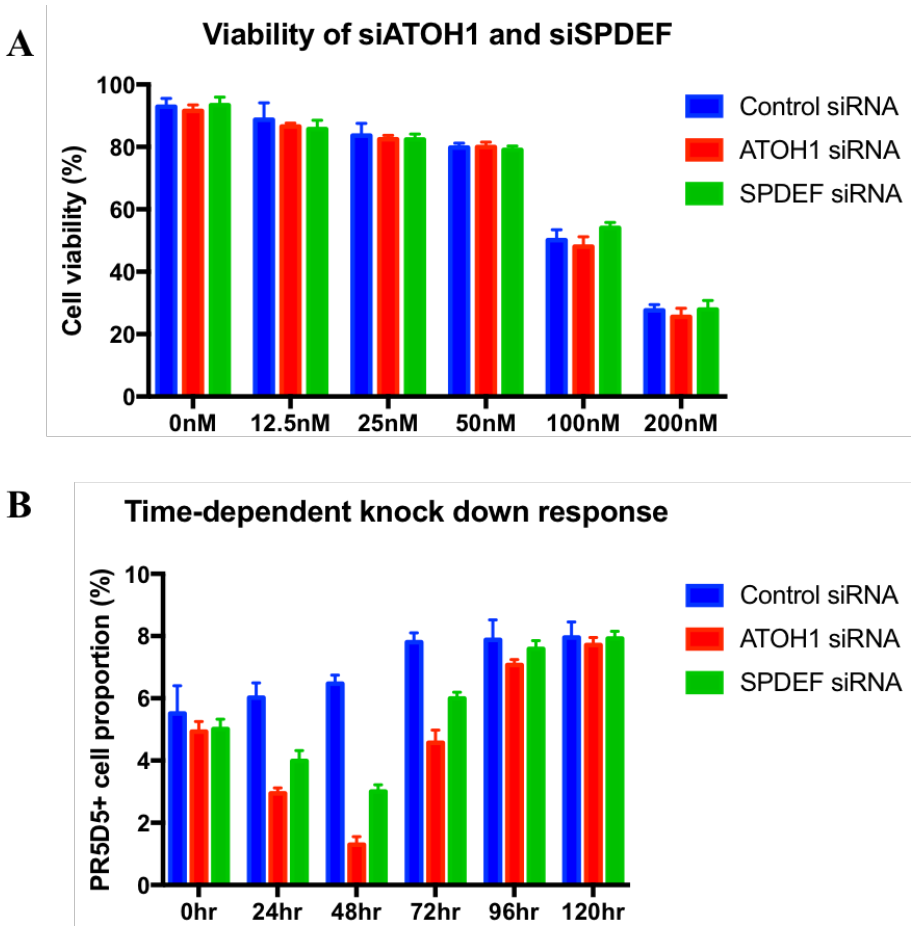


Figure 5.6 Cellular viability and knock down kinetics with siRNA transfection against ATOH1 and SPDEF

(**A**) 30,000 LS180 cells were seeded into each well of 6-well plates, and treated with scrambled siRNA and siRNA against ATOH1 (red) and SPDEF (green) at varying concentrations (0nM, 12.5nM, 25nM, 50nM, 100nM and 200nM) for 24 hours. Scrambled siRNA (blue) was used as control. Medium was changed on the next day, and cells were allowed to further grow for 72 hours before staining with LIVE/DEAD® Fixable Dead Cell Stain Kits and viability FACS analysis. (**B**) LS180 cells were fixed and stained with PR5D5 for FACS analysis at varying time points after treatment with siRNA. Error bar: mean ± SD, n=3.

Using the optimised experimental conditions, LS180 cells were transfected with the siRNA sequences specifically for ATOH1 and SPDEF at 50nM (the optimised experimental condition), with scrambled sequences as a control. After changing medium on the next day and letting the cells grow for another 48 hours, the expression of ATOH1 and SPDEF were assessed at protein level by immunostaining (**Figure 5.7**). Compared to the scrambled siRNA control, the proportion of ATOH1-positive cells in LS180 was largely down-regulated after the siRNA treatment against ATOH1. ATOH1 expression remained at similar level to the scrambled treatment in respond to the knock-down of SPDEF (**Figure 5.7A**). This indicates the efficient knock-down of ATOH1 and that SPDEF does not regulate the ATOH1 expression.

Similarly, the staining of SPDEF decreased when LS180 cells were transfected with siRNAs against SPDEF. However, the proportion of SPDEF-positive cells also decreased when the ATOH1 gene was knocked down. The staining for SPDEF did not show any observable change when transfected with scrambled siRNA (**Figure 5.7B**).

Taken together, the immunostaining of ATOH1 and SPDEF confirms the efficient knock down of their corresponding genes. Also, the decreased staining of SPDEF after ATOH1 knock-down suggests that SPDEF is a downstream gene of ATOH1, which is consistent with previous ChIP-seq data in mice small intestine (Lo et al., 2017).

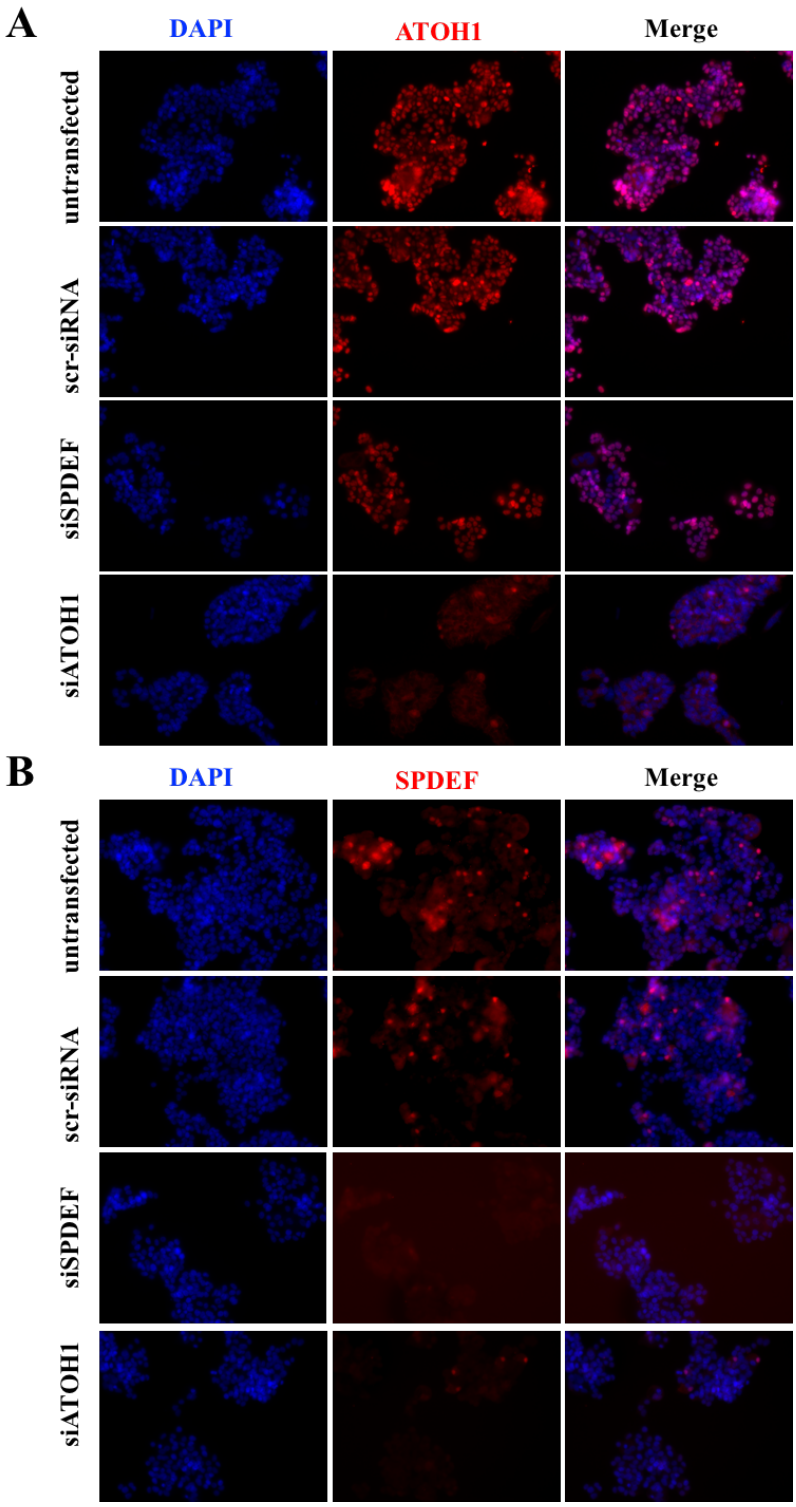


Figure 5.7 SPDEF and AOTH1 protein levels decreased after siRNA knock down 5,000 LS180 cells were seeded into each well of 96-well plates, and treated with scrambled siRNA and siRNA against ATOH1 and SPDEF at 50nM for 24 hours. Medium was changed on the next day, and cells were further cultured for 48 hours before fixation and immunostaining with DAPI (blue) and (**A**) ATOH1 (red, 1:500) or (**B**) SPDEF (red, 1:100).

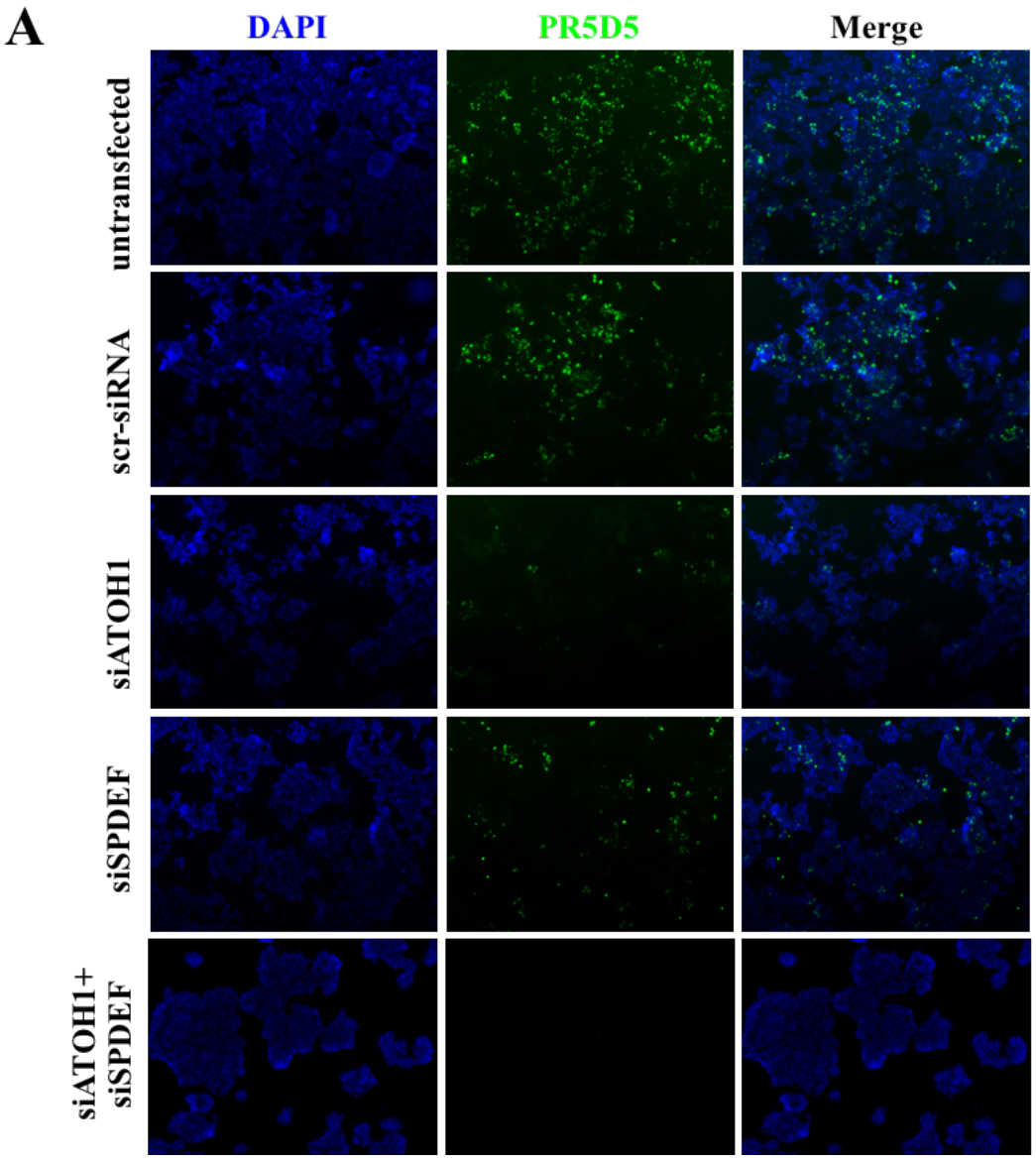
5.2.3.3 ATOH1 and SPDEF co-operatively regulate the expression of goblet cell-specific genes

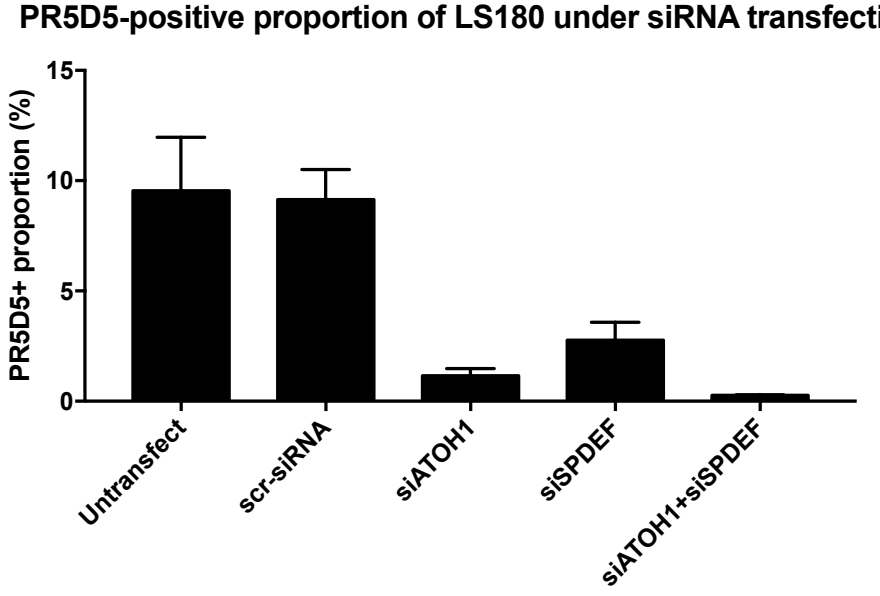
Given the significant roles of ATOH1 and SPDEF on goblet cell differentiation on its own (Noah, et al., 2010; Gregorieff, et al., 2009; Yang, et al., 2001) and the binding of ATOH1 directly on SPDEF (Lo, et al., 2017), it is rational to hypothesize the cooperative regulation of both factors in goblet cell differentiation. The expression of PR5D5 and TFF3 was assessed at protein level by immunostaining after the ATOH1 and SPDEF double knock-down. As shown in **Figure 5.8A**, when the expression of ATOH1 was knocked down, PR5D5 staining largely decreased to 1.16% from the basal level (9.55%), suggesting the positive regulation of ATOH1 on MUC2 expression. Knock down of SPDEF also resulted in the down-regulation of MUC2 expression with decreased PR5D5 staining, which is consistent with previous research of SPDEF knock down in the mouse model (Noah, et al., 2010; Gregorieff, et al., 2009). However, the proportion of positive MUC2 staining was higher (2.76%) compared to siRNA against ATOH1. Importantly, the double knock down of ATOH1 and SPDEF led to the absence of PR5D5 staining (0.25%). This suggested the co-operative effect of ATOH1 and SPDEF in regulating goblet cell differentiation and MUC2 expression.

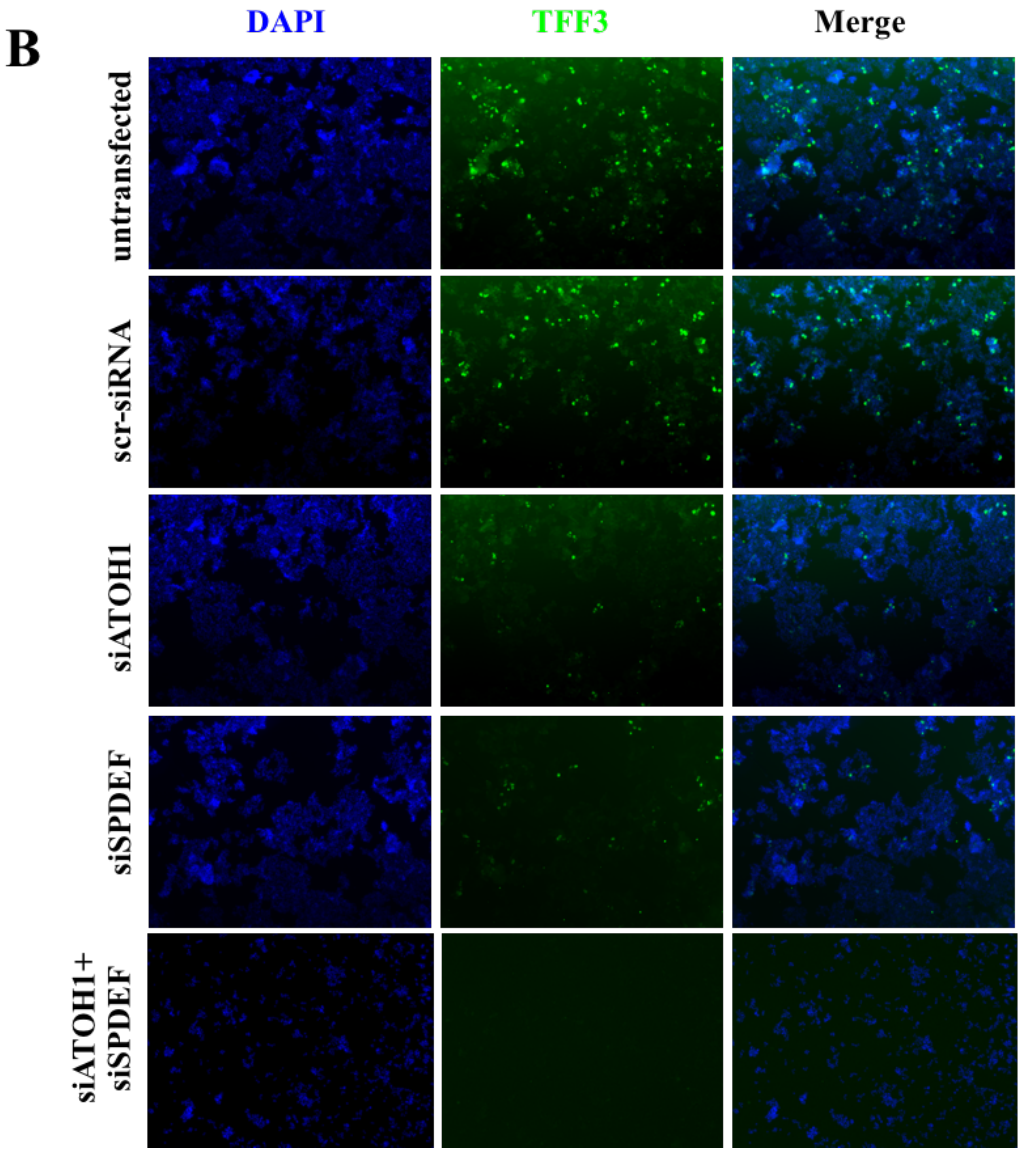
Figure 5.8B illustrates the decrease of TFF3 staining to 1.19% and 1.75% from the basal level (9.54%) after transfection with siRNA against ATOH1 and SPDEF separately. However, similar to PR5D5 staining, the double knock down of ATOH1 and SPDEF led to the complete absence of TFF3 staining (0.16%), confirming the

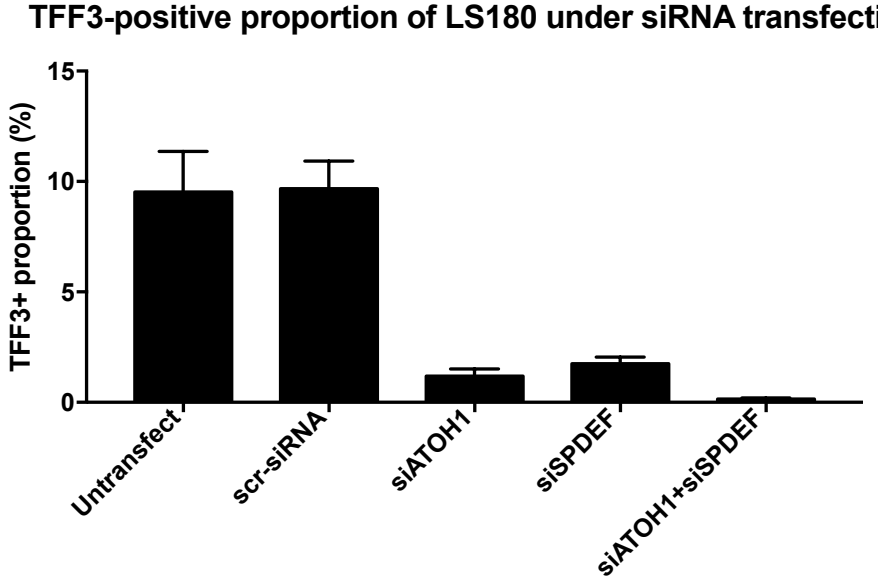
critical roles of the two genes *ATOH1* and *SPDEF* on regulating the expression of goblet cell-specific genes. This provides the evidence of the regulatory model based on the key transcriptional factors ATOH1 and SPDEF.

Figure 5.8 (the following page) Double knock-down of SPDEF and AOTH1 depletes MUC2 and TFF3 expression
5,000 LS180 cells were seeded into each well of 96-well plates, and treated with scrambled siRNA, siATOH1, siSPDEF and siATOH1+siSPDEF at 50nM for 24 hours. Medium was changed on the next day, and cells were further cultured for 48 hours before fixation and immunostaining with DAPI (blue) and (**A**) PR5D5 (green, 1:200) or (**B**) TFF3 (green, 1:200). Corresponding quantification of (**A**) PR5D5 and (**B**) TFF3 staining were conducted by counting the positive cells in at least three independent, random pictures in each well under different siRNA treatment.


A
DAPI
PR5D5
Merge
untransfected
scr-siRNA
siATOH1
siSPDEF
siATOH1+
siSPDEF


PR5D5-positive proportion of LS180 under siRNA transfection
PR5D5+ proportion (%)
15
10
5
0
Untransfect
scr-siRNA
siATOH1
siSPDEF
siATOH1+siSPDEF


B
DAPI
TFF3
Merge
untransfected
scr-siRNA
siATOH1
siSPDEF
siATOH1+
siSPDEF


TFF3-positive proportion of LS180 under siRNA transfection
TFF3+ proportion (%)
15
10
5
0
Untransfect
scr-siRNA
siATOH1
siSPDEF
siATOH1+siSPDEF

5.2.4CA12, a potential cellular surface marker to identify goblet cell progenitors

Based on the microarray analysis, *CA12* is a significantly differentially expressed gene in goblet cell-positive cell lines. CA12 is expressed at the surface cuff areas of basolateral plasma membrane of enterocytes in normal colon tissues (Kivelä et al., 2000). Thus, CA12 was considered as a potential surface marker for goblet cells, and chosen for further investigation. In colorectal cancers, *CA12* is expressed highly at the bottom parts of adenomatous mucosa, and its expression levels increase along with the dysplasia grades (Kivelä et al., 2000). However, the functions of CA12 during the tumour development and progression, and its relation to goblet cell differentiation are not clear.

5.2.4.1 Expression of CA12 in colorectal goblet cells

In microarray analysis, CA12 expression is 5-fold higher in goblet cell-positive versus negative cell lines (p-value < 0.01). In the RNA-seq data, mRNA expression of CA12 in goblet cells (sample 1-7), with rpkm values between 30 and 70, was slightly higher than non-goblet cells (sample 8-14), with rpkm values between 20 and 40 (**Figure 5.9A**). However, with a fold-change smaller than 2, CA12 did not rank as one of the top differentially expressed genes in the RNA-seq of goblet cells.

Therefore, CA12 was co-stained with PR5D5 in a panel of colorectal cancer cell lines with the highest and lowest CA12 expression from microarray data for further investigation. Most CA12-positive cells did not co-stain with PR5D5. In LS180, 6.00%

and 4.44% showed positive reactivity with CA12 and PR5D5 respectively, while only 0.14% cells were positive for both. In cell lines RW7213, CL40 and HCA46, the double positive staining of CA12 and PR5D5 was 2.07%, 2.39% and 1.72% respectively, and the CA12-positive population clearly separated from the main population. This might suggest an intermediate transition state from CA12-positive to PR5D5-positive populations.

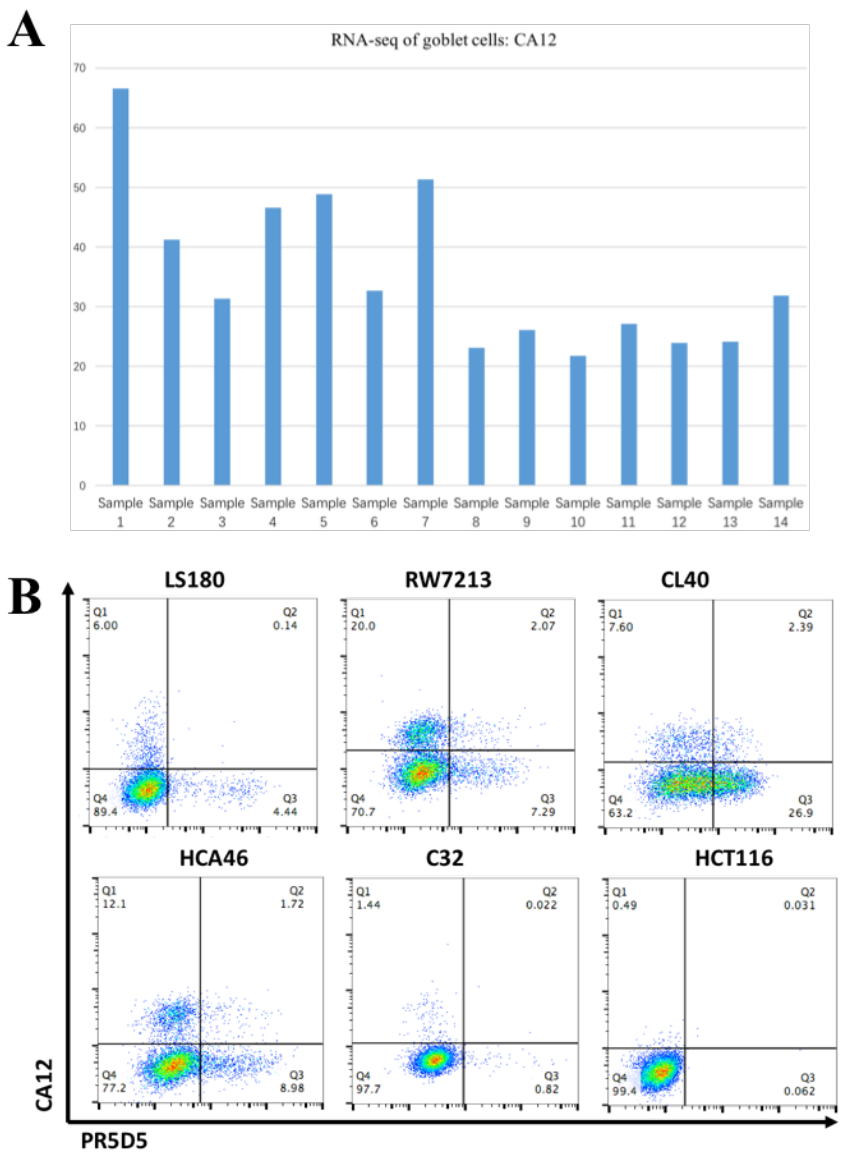


Figure 5.9 CA12 expression in colorectal goblet cells and co-staining with PR5D5
(**A**) Expression levels of CA12 were represented by the rpkm values of goblet cells (Sample 1-7) and non-goblet cells (Sample 8-14). (**B**) Flow cytometry analysis of CA12 co-staining with PR5D5.

5.2.4.2 CA12 may serve as a potential marker for goblet cell progenitors

As a transmembrane protein, CA12 could serve as a useful marker to investigate its association with goblet cells without permeabilising plasma membranes. Through enriching for CA12-positive and –negative cells from the SW1222 cell line by FACS sorting (**Figure 5.10A**), we showed a morphological difference between CA12-positive and –negative cells under 2D culture - CA12-positive cells showed larger size and faster growth rate along the 5-day culture (**Figure 5.10B**).

After sorting and culturing for five days, the CA12-positive and –negative cells were fixed and stained with different antibodies. **Figure 5.11A** shows CA12 reactivity in the CA12-positive sorted cells, while a few CA12-negative sorted cells started to express CA12 from after five days of culture. **Figure 5.11 B-D** shows the staining with PR5D5, TFF3 and SPDEF in the CA12-positive sorted cells. There was no reactivity against these three antibodies in the CA12-negative sorted cells. CA12-positive and -negative sorted cells showed little difference in the staining pattern of ATOH1 (**Figure 5.11E**).

Collectively, these data suggest that CA12, a cellular surface protein, might serve as a novel marker for goblet cell progenitors in human colorectal cancers. This observation should be repeated and further validated, especially for the CA12 selection. It could be further tested in primary tissues with varying grades of differentiation. The expression of CA12 can be also checked in the ATOH1 and SPDEF knocked-down cell lines.

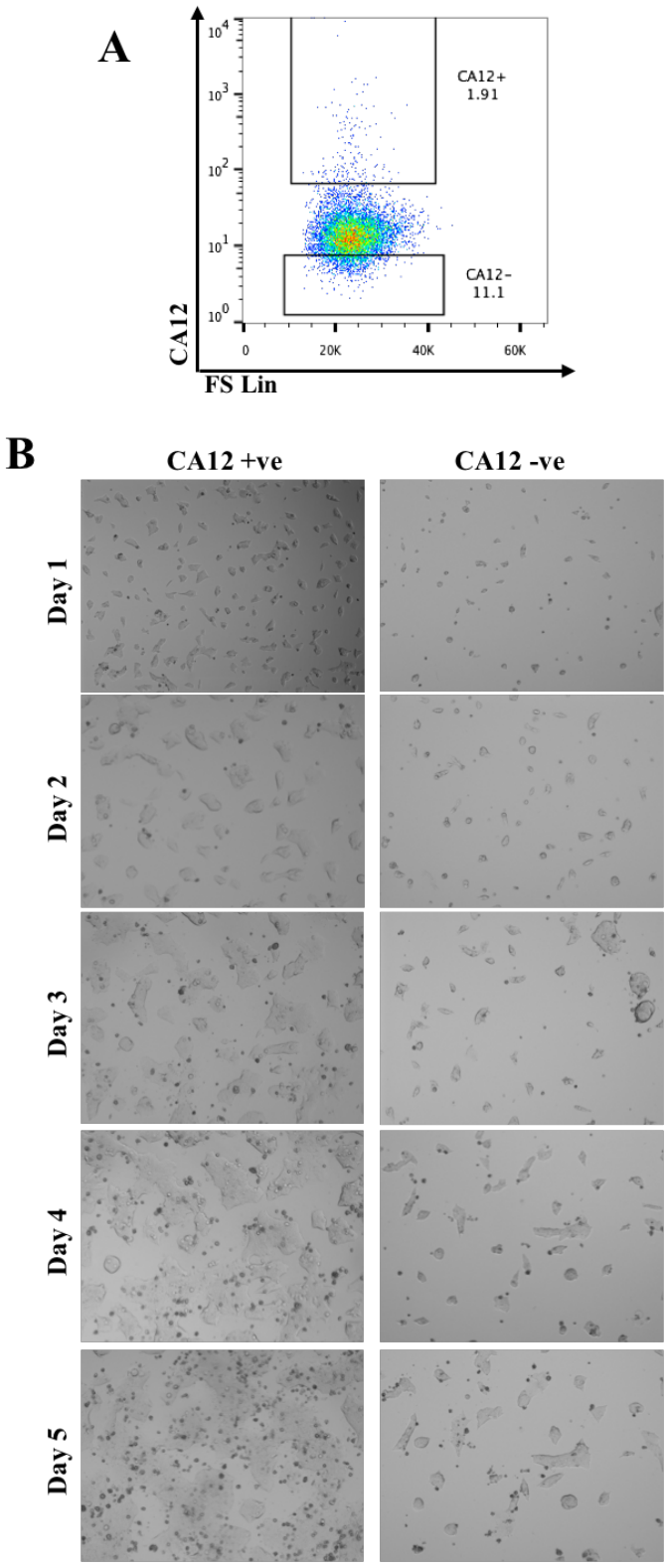


Figure 5.10 CA12 staining and FACS sorting on LS180 cells

(**A**) SW1222 cells were stained with CA12 (1:200) and DAPI. After gating out dead cells, the isotype control was used to set up the sorting gate. 5000 cells of the extreme 1-2% CA12-positive and ~10% CA12-negative subgroup were sorted into each well for further experiments. (**B**) Sorted cells were grown and pictures were taken for 5 days.

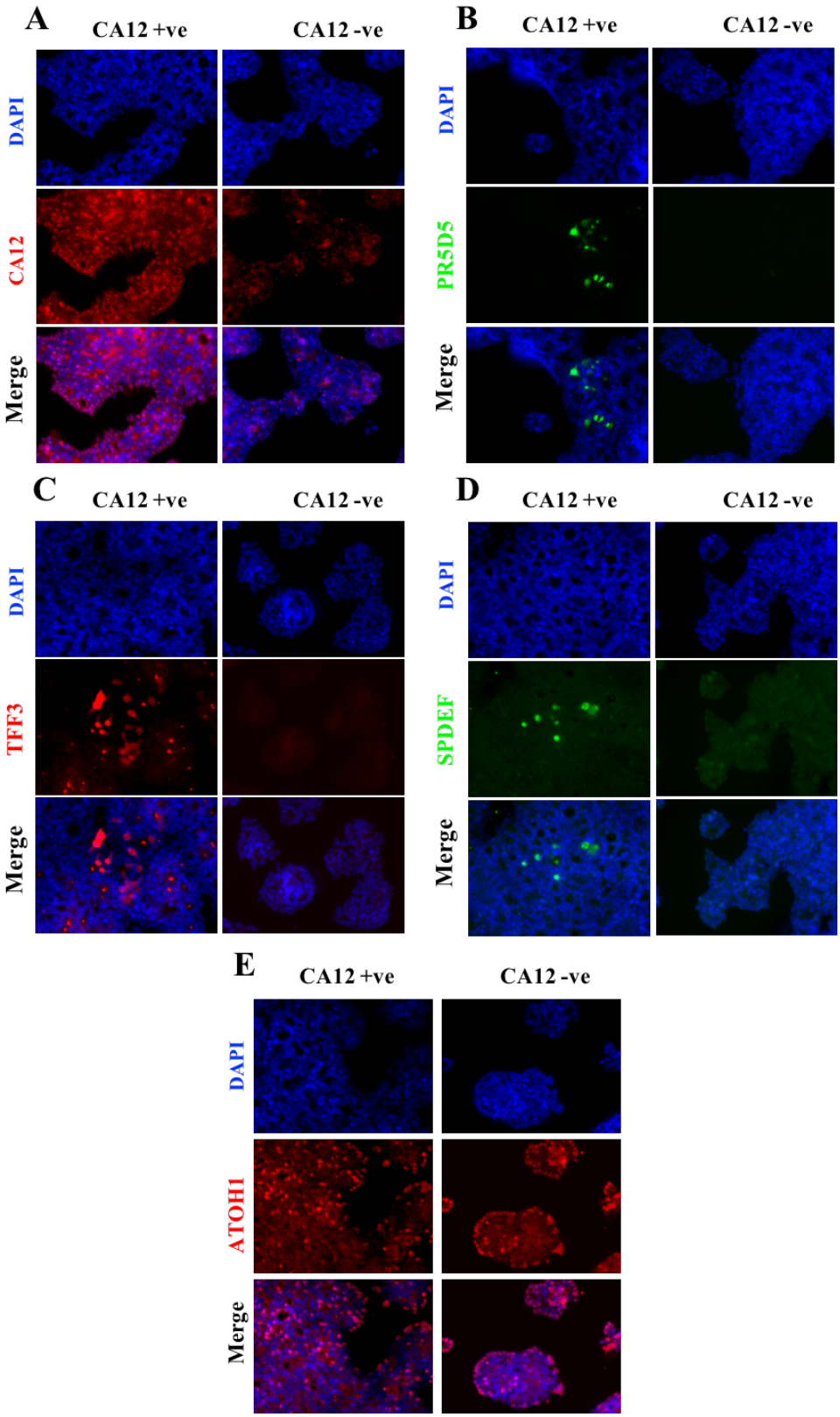


Figure 5.11 Immunostaining with CA12, PR5D5, TFF3, SPDEF and ATOH1 in the CA12-positive and -negative sorted cells

After sorting and culturing for five days, SW1222 cells were fixed and stained with CA12 (1:200), PR5D5 (1:200), TFF3 (1:200), SPDEF (1:100), ATOH1 (1:500) and DAPI.

5.3 Discussion

Lumen formation is a critical feature to determine differentiation of cancer stem cells under the 3D culture (Ashley et al., 2013; Yeung et al., 2010). This raised the question whether goblet cell differentiation is associated with lumen formation in colorectal cancer cell lines. Based on the goblet cell screening in **Chapter 3**, and preliminary lumen formation characterisation, no statistically significant association between goblet cell differentiation and lumen was identified. This finding is consistent with the previous observation that knockdown of CDX1, a key regulator in columnar cell differentiation and lumen formation, disrupted lumen formation without affecting goblet cells (Yeung et al., 2011), suggesting that additional regulation is required other than those involved in lumen formation.

The microarray expression analysis and RNA-seq of goblet cells outlined the genes that might serve as goblet cell markers or transcriptional regulators. TFF3, as the top gene from goblet cell-specific RNA-seq, is already known for its critical role in mucosal repair and regeneration processes after gastrointestinal injury (Mashimo et al., 1996). This might be achieved by heterodimer formation with FCGBP that is attached to MUC2 (Johansson et al., 2009). These findings raised interesting questions as to whether TFF3 is solely expressed and secreted within goblet cells, and how the dysregulation of TFF3 is different from MUC2 in colorectal cancers.

To address these questions, the mRNA expression of TFF3 in the RNA-seq data of goblet cells and non-goblet cells was examined. The high expression of TFF3 in goblet cells is consistent with previous studies (Wong et al., 1999; Hoffman et al., 2001; Taupin et al., 2003), while moderate TFF3 expression was also observed in non-goblet cells. The TFF3 expression in goblet cells is confirmed at protein level by the co-localization of TFF3 with PR5D5, confirming TFF3 as a goblet cell marker in human colonic crypts. This observation was strengthened by showing that TFF3 expression increased, in a similar pattern to MUC2, by Notch blockade via DBZ treatment. The immunostaining of TFF3 was further extended to investigate the expression patterns of MUC2 and TFF3 across human colorectal cancer cell lines. Notably, TFF3 was highly expressed in SW403, JHCOLOY1 and PMFKO14, showing no reactivity with PR5D5, suggesting that TFF3 identifies goblet cells or goblet cell precursors that lack in MUC2 production in human colorectal cancer cell lines. This is consistent with the previous observation of continuous TFF3 expression in the intestines of MUC2-deficient rat (Matsuoka et al., 1999). These findings suggest that MUC2 and TFF3 are both regulated by Notch pathway, whereas different mechanisms regulate their expression.

SPDEF, as the top transcriptional factor in the goblet cell RNA-seq data, is suggested to promote goblet cell maturation (Gregorieff et al., 2009), and its core promoter region is directly bound and regulated by ATOH1 (Lo et al., 2017). ATOH1 serves as a key mediator in secretory lineage commitment and is negatively regulated by the Notch pathway target gene *Hes1* (Yang et al., 2001; Shroyer et al., 2007). It was hypothesised

that ATOH1 and SPDEF co-operatively regulate the expression of goblet cell-specific genes in human colorectal cancer cell lines. This hypothesis was tested by checking their expressions in the RNA-seq data of goblet cell and non-goblet cells. Even though ATOH1 is expressed at a slightly higher level in goblet cells, its absolute rpkm values are low. The ATOH1 staining in nucleus of colorectal cancer cell lines might result from the distinct turnovertime. SPDEF is significantly expressed in goblet cells and shows almost no expression in non-goblet cells. The mRNA expression was confirmed in the immunostaining that SPDEF co-labelled the cells with PR5D5, confirming the specific expression of SPDEF in goblet cells. The functional significance of ATOH1 and SPDEF was addressed by siRNA-mediated gene silencing. The observation that SPDEF expression was down-regulated after knocking down ATOH1 but not vice versa, indicates that SPDEF is the downstream target of ATOH1 consistent with the recent published ChIP data (Lo et al., 2017). PR5D5 and TFF3 staining decreased after knocking down of ATOH1 or SPDEF, and the double knockdown resulted in the absence of both staining. This has illustrated the co-operative regulation of ATOH1 and SPDEF on the goblet cell-specific gene expression.

From the microarray analysis, CA12 is expressed more than 5-fold higher in the goblet cell-positive than -negative cell lines (p-value < 0.01). In the goblet cell RNA-seq data, however, CA12 did not present much differential expression between goblet and non-goblet cells. This was confirmed by the co-staining with PR5D5 and CA12 resulting in only a small subset of cells that are double positive. As a transmembrane protein, CA12

allows surface staining, sorting and further culturing without permeabilising the plasma membranes. The sorted CA12-positive cells were morphologically larger than CA12-negative cells. In addition, a subset of CA12-positive cells showed positive immuno-reactivity with PR5D5, TFF3 and SPDEF. This indicates that CA12 may be a novel marker for goblet cell progenitors. However, this experiment has been only conducted once and further investigations are required to confirm this statement.

To summarise, results in this chapter characterised the key genes in the goblet cell differentiation in human colorectal cancer cell lines. TFF3, a mucus constituent, can act as a marker to identify goblet cells without producing MUC2 in colorectal cancer cell lines. The two transcriptional factors, SPDEF and ATOH1, regulate expression of MUC2 and TFF3 in a co-operative way. CA12 may serve as a novel surface marker for goblet cell progenitors. These results have deepened the current understanding of the dysregulated gene expression during goblet cell differentiation, with the potential for more precise classification of colorectal cancers.

CHAPTER 6
DISCUSSION AND FUTURE DIRECTIONS

6.1 Goblet cell differentiation in colorectal cancer cell lines

6.1.1 PR5D5, an in-house antibody targets MUC2

PR5D5, an in-house monoclonal antibody with resistance of antigenic determinants to formalin fixation, shows highly restricted positive staining within goblet cells (Richman and Bodmer, 1987). In this thesis, MUC2 has been further identified as the target protein of PR5D5 using a combination of immune-based and molecular techniques. Commercial anti-MUC2 (Clone CCP58, Dako) was used for immunofluorescence analysis to show that the regions of positive reactivity within MUC2 were substantially recognised by PR5D5. The specificity of PR5D5 in goblet cells was further validated by the absence of staining when siRNA was used to knock-down MUC2 in the colorectal cancer cell line LS180. PR5D5 specificity was further confirmed by the competitive binding assay against anti-MUC2. The staining pattern of PR5D5 is highly consistent with a collection of previous studies (Richman and Bodmer, 1987; Albaugh et al., 1992; Campbell et al., 1994; Nair et al., 2003; Ashley et al., 2013; Yeung, Patent WO 2015/198065 A1). These data further confirm the Western Blot result that PR5D5 targets a ~70kDa protein, approximate size of glycosylated MUC2 protein (Campbell et al., 1994).

6.1.2 Classification of colorectal cancer cell lines regarding goblet cell differentiation

Although the significance of goblet cell functions has been highly acknowledged for long, their transcriptional regulation has only started to be understood in the past decade. It is not completely clear how the goblet cell differentiation is dysregulated in human colorectal cancer. Thus, characterisation of goblet cell and its transcriptomic profile is a critical step to solve these questions. The CIL has access to a panel of more than 100 human colorectal cancer cell lines and their expression microarray data, which serve as a valuable resource to assess their differentiation potential *in vitro* (Ashley, 2013; Yeung, 2011). These cell lines not only reflect a wide range of primary human colorectal cancers with varying degrees of differentiation, and also represent different proportions and growth characteristics of goblet cells in colorectal cancers.

The mRNA expression of MUC2 was examined. **Figure 6.1** presents the distribution of MUC2 mRNA expression across 143 cell lines from the microarray data. A narrow distribution is at the low end that is contributed by the cell lines with log2(MUC2 expression) at 5-7, i.e. MUC2 expression between 32AU to 128AU. This reflects the goblet cell-negative lines. A heavy tail can be found at the high end with log2(MUC2 expression) at 9-13, i.e. MUC2 expression more than 512AU. This corresponds to the goblet cell-positive cell lines. The board shallow distribution characteristic is expected from the MUC2 expression of only a subset of cells within the line.

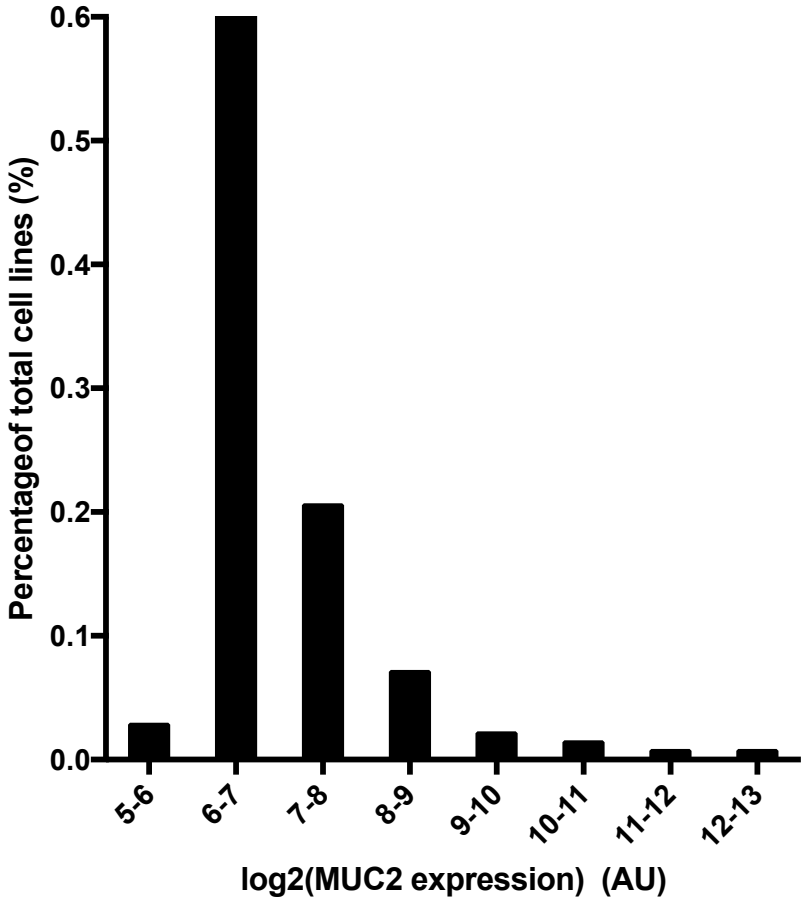


Figure 6.1 MUC2 expression distribution across 143 colorectal cancer cell lines
The MUC2 expression across 143 human colorectal cancer cell lines. The x-axis is the log2(MUC2 expression). The y-axis is the percentage of cell line counts compared to the total 143 cell lines.

Goblet cell differentiation was also screened at protein levels using PR5D5 and anti-MUC2 mAb (Dako, UK) in a panel of 64 human colorectal cancer cell lines. The screening results of these 64 cell lines were compared with their microarray data of MUC2 expression. Together, 64 colorectal cancer cell lines were classified according to their goblet cell differentiation. The proportion of goblet cell-positive, -intermediate and -negative cell lines are 23.4% (15/64), 39.1% (25/64) and 37.5% (24/64) respectively. It is worth noteworthy that some goblet cell-intermediate cell lines are classified based on weak mRNA expression but with no sign of any MUC2 protein expression. This is comparable to one previous immunohistochemical MUC2 staining in colorectal tumours from 381 patients, where high, low and loss of MUC2 expression is observed in 61 (16%), 225 (61%) and 85 (23%) patients (Betge et al., 2016). It should be noted a high proportion of colorectal cancers express only quite low levels of MUC2.

For each new antibody that was introduced in this thesis, the cell lines with the highest and lowest expressions of corresponding genes from the microarray mRNA profiles were used as positive and negative controls respectively to confirm the antibody specificity. It should be noted here as well as in the later sections where a new antibody is used that the lack of antibody validation using Western Blot might give rise to unspecific antigen identification (Bordeaux, et al., 2010; Signore and Reeder, 2012). Thus Western Blot should be included in order to evaluate the molecular weight of the targeted protein and further validate the specificity of antibodies in the future study.

6.1.3 CA12, a potential novel marker for goblet cell progenitors

Studies of CA12 have traditionally focused on it catalytic functions as a carbonic anhydrase (Pastorek, et al., 1994; Türeci, et al., 1998) and expression patterns in normal tissue and cancers (Parkkila, et al., 2000; Kivelã et al., 2000), with less attention to its roles in cellular differentiation. In this thesis, CA12 is identified as one of the significantly differentially expressed genes in goblet cell-positive cell lines through a global gene expression analysis by comparing microarray expression profiles of goblet cell-positive versus -negative cell lines. The co-staining of CA12 and PR5D5 identified distinct subpopulations with only a small subset of co-stained cells. This means the expression of CA12 is high in goblet cell-positive cell lines, but not specifically targets goblet cells, indicating the functions of CA12 is not restricted within goblet cells, and might functionally serve as a goblet cell precursor.

Thus, the function of CA12 in differentiation was further defined. The CA12-positive sorted cells did present higher expression of MUC2, TFF3 and SPDEF than the CA12-negative sorted cells in the human colorectal cancer cell line SW1222. The ATOH1 expression did not show difference between CA12-positive and -negative cells. This indicates that goblet cells might be derived from the CA12-expressing precursors. A high possibility might be that the CA12 expression is switched off in goblet cell differentiation, and so would still not be a surface marker for mature goblet cells.

For now, there has been no well-established marker for goblet cell precursors described. In mouse small intestines, secretory progenitors are usually defined morphologically, and the maturation of goblet cells is assessed by the expression of goblet cell genes, predominantly MUC2 and TFF3 (Van der Sluis et al., 2006; Velcich et al., 2002; Wenzel et al., 2014; Katz et al., 2002). Clevers and colleagues suggested the transcriptional factor SPDEF as a goblet cell precursor marker in mouse intestines (Gregorieff et al., 2009). Data in this thesis defined CA12 as a surface marker for goblet cell precursors, possibly upstream to SPDEF but downstream to ATOH1, adding additional layers of hierarchical understanding of goblet cell differentiation and maturation. Further experiments would be to see if it is possible to isolate CA12-positive cells that are not MUC2 positive from normal colorectal tissue - ideally then they would grow for short term and see if they turn into goblet cells producing MUC2.

6.2 Development of a novel method for RNA isolation from fixed goblet cells

Despite the power of microarray expression analysis in a large panel of human colorectal cancer cell lines, it must be noted that the proportion of goblet cells is relatively small in colorectal cancer cell lines (usually less than 10% based on the classification in **Section 3.2.3**). In this case, the mRNA from goblet cells is largely diluted when the expression profiles from the bulk population were compared. This situation can be aggravated by the key transcriptional factors expressed at a low level. For example, SPDEF is only expressed 1.62-fold higher in goblet cell-positive cell lines (ranking 1192). Thus, the transcriptomic profile of goblet cells, a specific subset of the

total epithelial population, should be characterised. To date, however, this attempt has largely been restricted by the lack of available surface markers. Therefore, this project explored using PR5D5 on fixed cells for FACS isolation of goblet cells and the first step was to be able to get good mRNA from fixed cells

A systematic characterisation of RNA degradation was pursued after each step of fixation, permeabilisation, intracellular staining and mimicked FACS sorting. Permeablisation was identified as the key step that results in RNA degradation. Further, we have also evaluated the effects of different fixatives, permeabilising reagents and RNase inhibitory reagents on preserving RNA quantity, purity and integrity. 4% PFA and 0.1% saponin gave the reasonable PR5D5 staining and showed the best performance in RNA preservation. Several RNase inhibitors were also optimised, and RNasin Plus (Promega) was selected for its ability to maintain PR5D5 staining pattern, RNA purity and integrity. The optimisations based on the evaluation showed the advantages over previous research by being easier, less time-consuming and more economic-friendly (Khchbin et al., 1990; Esser et al., 1995; Pan et al., 2011; Hrvatin et al., 2014). Subsequent to the optimisation presented here, Thomsen and colleagues (January 2016) reported an optimised protocol for RNA isolation from the fixed single cells from radial glia (Thomsen et al., 2016). Their optimisation, including the choice of RNase inhibitor and RNA-preserving buffer in the sorting strip tubes, were in agreement with the results in this thesis (Thomsen et al., 2016).

Using the optimised protocol, the goblet cell transcriptomic profile in human colorectal cancer cell line LS180 was characterised. This helped to identify the genes that are overwhelmingly expressed in the cancerous goblet cells, which could serve as potential markers for goblet cells, or key regulators that are highly specific within goblet cells. These genes were further discussed and analysed in **Section 6.3-6.5**.

6.3 Three categories of genes in goblet cell differentiation

The key genes that were identified from the literature, the mRNA expression microarray analysis and the goblet cell-specific RNA-seq were categorised into three groups: the 'designers'; the 'bricks'; and the 'glue', based on their putative functions in the differentiation and maturation of goblet cells.

Group 1, i.e. the 'designer', refers to the goblet cell-fate decision makers, which are mainly transcriptional factors such as ATOH1 (Gersemann et al., 2009), SPDEF (Noah et al., 2010), KLF4 (Katz et al., 2002) and Notch-related genes (van Es et al., 2005; Zheng et al., 2011). These genes usually have long-term significant effects and low-level expressions. In most cases they are only within a small subgroup of cells (e.g. stem cells and/or secretory progenitors), thus it is unlikely to identify them via microarray analysis on the bulk population as described before. For example, the SPDEF expression in goblet cells is >32-fold higher than non-goblet cells from the RNA-seq, while only 1.6-fold higher from the bulk population-based microarray analysis. Not only does SPDEF have a highly restricted expression within goblet cells,

but also it has an important role at a relatively later stage of goblet cell differentiation. From RNA-seq data, ATOH1 and KLF4 showed 2.8- and 6.4-fold higher expression in goblet cells, respectively. Manipulating expression of these transcriptional factors should show clear influences on the downstream genes (Zheng et al., 2011; Gersemann et al., 2009; Leow et al., 2004; Gregorieff et al., 2009; Noah et al., 2010).

Group 2 refers to the mucus components (the 'bricks'), e.g. MUC2 (Velcich et al., 2002; Van der Sluis et al., 2006), FCGBP (Johansson et al., 2011; Johansson et al., 2008; Harada et al., 1997), TFF3 (Fernández-Estívariz et al., 2003) and ZG16 (Pelaseyed et al., 2014). FCGBP is not an immunoglobulin binding protein as its name indicates, instead it covalently binds to MUC2 via disulphide bonds at its vWF D domains (Johansson et al., 2009). TFF3, a trefoil factor, also shows direct or indirect binding with FCGBP and MUC2 (Albert et al., 2010). ZG16 is a lectin-like protein that can bind to Gram-positive bacteria and push away from mucosa (Pelaseyed et al., 2014). These proteins are identified as important constituents of mucus. Their expression in goblet cells, usually in a large amount, can therefore be easily identified via microarray analysis and goblet cell RNA-seq. For example, the RNA-seq data showed that expression levels of FCGBP, MUC2, SPINK4, TFF3 and ZG16 are more than 32-fold higher in goblet cells than non-goblet cells. They all showed similar staining patterns according to the Human Atlas Protein (http://www.proteinatlas.org) and accumulated in the large secretory vesicles within goblet cells where they are packed together to form a net-like structure before secretion (Johansson et al., 2011; Johansson et al.,

2008). Group 2 gene mutations in cancer, especially MUC2, are generally considered as the phenotype of dysregulated differentiation and maturation of goblet cells.

Group 3 refers to the 'decorator', i.e. the genes involved in mucus connection and modification, e.g. ST6GALNAC1 (Marcos et al., 2011; Larsson et al., 2011), AGR2 (Zhao et al., 2010; Park et al., 2009) and AGR3 (Zheng et al., 2006). All these three genes were identified to be expressed more than 5-fold higher in goblet cell-positive cell lines from microarray analysis, and about 4-fold higher in RNA-seq of purified goblet cells. These genes are important in the full maturation of mucus within goblet cells, including the glycosylation and mucus packaging. For example, AGR2 and AGR3, as important protein disulphide isomerase family members, can catalyse disulphide bond formation between cysteine residues of MUC2, TFF3 and FCGBP (Zheng et al., 2006). ST6GALNAC1, on the other hand, is involved in the sTn synthesis during the O-glycosylation at the tandem repeats of the MUC2 protein (Munkley, 2016). These genes also showed tremendous clinical potential for diagnostics and prognostics. For example, by recognising a specific pattern of MUC2 glycosylation alteration, the *Wisteria Floribunda Lectin* (WFA) that binds to the terminal GalNAc residues at the side sugar chain of MUC2, has been patented by our collaborator Mr Yeung to distinguish benign hyperplastic polyps (HPs) from pathologically significant polyps, including sessile serrated polyps (SSPs), traditional serrated adenomas (TSAs) and mucinous cancers (Yeung, Patent WO 2015/198065 A1). The application of WFA in surgery has been under investigation in an ongoing clinical

trial run by our collaborator Mr Barnes and colleagues (Trial number: ISRCTN90128107).

It must be noted that not all genes from the microarray analysis and RNA-seq data have been defined into this classification system, and further genetic manipulation and functional verification are required for the further characterisation of these genes.

6.4 TFF3 in colorectal cancer

TFF3 is abundantly expressed in the theca of mature goblet cells in the form of both monomers and dimers. This work has shown that TFF3 and PR5D5 target the same cells in the cell lines LS180 and SW1222, confirming the expression of TFF3 within goblet cells. Under the Notch pathway blockade via gamma-secretase inhibitor treatment, TFF3 expression increased co-ordinately with MUC2, indicating the regulation of TFF3 through the Notch pathway. The RNA-seq data of goblet cells indicates the significantly high TFF3 expression in goblet cells, while the non-goblet cells also retained moderate TFF3 expression with rpkm values of 100-200.

In human colon, TFF3 plays a central role in the epithelial restitution and mucosal regeneration processes (Taupin and Podolsky, 2003; Kjellev, 2009). TFF3 is important in forming and stabilising the net-like structure of mucus by interacting with MUC2 directly or indirectly. Tomasetto and colleagues identified a role of TFF3 in enhancing mucin viscosity by binding to the vWF region of MUC2 (Tomasetto et al., 2000). Albert

and colleagues report the formation of heterodimer via disulphide-bond between TFF3 and FCGBP that connects with MUC2 covalently or non-covalently (Albert et al., 2010). Thus it is important to express both MUC2 and TFF3 correctly for the functional integrity of goblet cells and mucus.

Thus, the co-staining patterns of TFF3 and PR5D5 were characterised in a panel of colorectal cancer cell lines. The expression of TFF3 in almost all cells without MUC2 expression has been identified in a small subset of cell lines (SW403, JHCOLOY1 and PMFKO14 in **Figure 5.4**). The abundant TFF3 expression in almost all cells, and showed stronger intensity in the cells where it co-labelled with PR5D5 is also demonstrated (RW7213 in **Figure 5.4**). No cell lines with highly expressed MUC2 and lowly expressed TFF3 were found across the cell line panel. These observations suggest that TFF3 can identify the goblet cells without MUC2 production in colorectal cancers. This is consistent with the observation in a mouse model where $MUC2^{-/-}$ mice lacked MUC2 expression and showed no morphologically recognizable goblet cells, while retaining TFF3 (Velcich et al., 2002).

This observation seems to happen only in cancers but not in normal colons (**Figure 5.4**). One possibility is that in cancers there is MUC2 gene promoter methylation (Okudaira et al., 2010; Hanski et al., 1997) which would specifically switch off MUC2 but not TFF3, and have nothing to do with the normal situation. There may be a selection against MUC2 expression for the outgrowth of a colorectal tumour (Mizoshita et al.,

2007) with no further advantage to losing TFF3 expression. This would explain this disparity of expression in some colorectal cancers without any implication for what happens in normal no tissues. Switching of goblet cell differentiation completely would be another way to achieve the same goal.

The co-staining for MUC2 and TFF3 might be applied in primary tissue slides of premalignant colorectal dysplasia and varying stages of cancers. And the methylation of MUC2 and its regulation can also be further investigated.

6.5 A regulatory triangle

Here a regulatory triangle of ATOH1 and SPDEF is presented from the expression analysis and functional validation (**Figure 6.2**). Based on the RNA-seq data, SPDEF was identified to be 32-fold highly expressed in goblet cells compared to non-goblet cells. The specific expression of SPDEF in goblet cells was confirmed at the protein level by co-staining with PR5D5 (**Figure 5.5**). Both MUC2 and TFF3 decreased when SPDEF was knocked down, indicating the direct regulation of SPDEF on the expression of goblet cell genes.

No obvious ATOH1 expression alteration was observed when SPDEF expression is silenced, but SPDEF did decrease when ATOH1 was knocked down. This suggests that SPDEF is regulated downstream by ATOH1, which is consistent with previous ChIP-seq data that the SPDEF core promoter is bound by ATOH1 (Lo et al., 2017). These

data indicate an indirect regulation of ATOH1 on goblet cell differentiation via SPDEF (**Figure 6.2**). ATOH1 can also positively regulate on the expression of the goblet cell genes *MUC2* and *TFF3*. By knocking down ATOH1, the expression of MUC2 and TFF3 decreased in colorectal cancer cell lines. Together with the ChIP-seq data that ATOH1 directs bind to MUC2 core promoter,

The co-operative effects of ATOH1 and SPDEF on goblet cell differentiation were further demonstrated. The expression of MUC2 and TFF3 decreased in response to the single knock down of either ATOH1 or SPDEF, while when they were knocked down together, a complete depletion of MUC2 and TFF3 expression was observed, suggesting the co-operative effects of ATOH1 and SPDEF in regulating goblet cell differentiation. This further strengthened the regulatory triangle of ATOH1 and SPDEF (**Figure 6.2**).

The qPCR experiments will be needed to further assess the decreased gene expression under knock down. The co-operative model can be also tested in regulating other genes that are identified differentially expressed in microarray analysis and the goblet cell transcriptome.

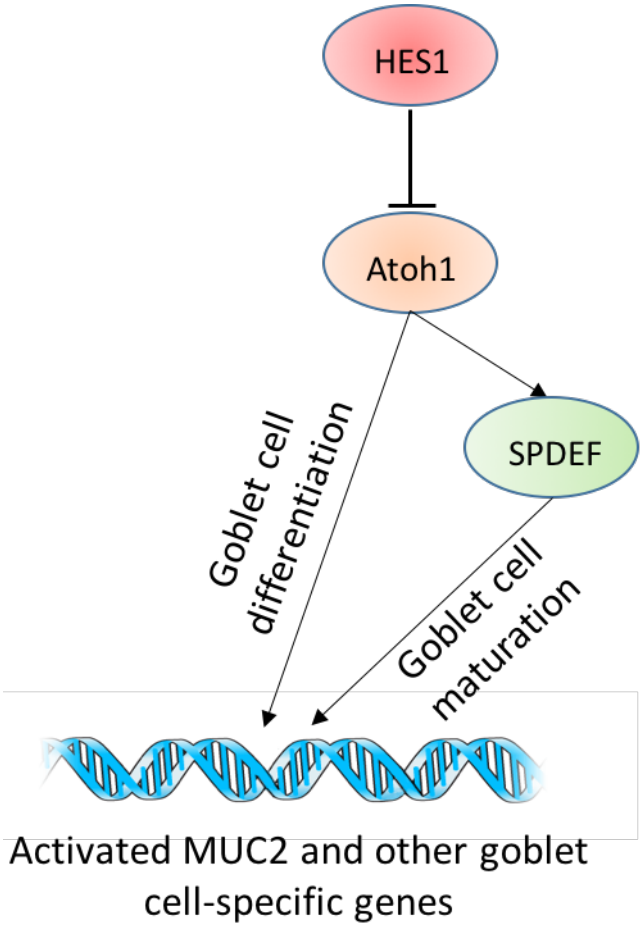


Figure 6.2 Schematic model of ATOH1 and SPDEF on regulating the expression of MUC2 and TFF3

ATOH1 and SPDEF are two key genes in goblet cell differentiation from the RNA-seq data. Based on the work of this thesis, we show a regulatory triangle of ATOH1 and SPDEF on the expression of MUC2 and TFF3. SPDEF is downstream regulated by ATOH1, and both transcriptional factors co-operatively regulate goblet cell differentiation.

6.6 Summary

In conclusion, the goblet cell-related genes have been characterised from the bulk microarray analysis and the transcriptome of purified goblet cells. CA12 is shown as a potential surface marker for goblet cell progenitors, whose expression may be switched off in goblet cell maturation. The intra-cell line variability has been also illustrated regarding MUC2 and TFF3 expression, which can be due to differentiation to goblet cells, possibly including previously undescribed non-MUC2 producing abnormal or precursor goblet cells - their presence may define a new subtype of colorectal cancers. A regulatory triangle also clarifies the pattern of hierarchical gene control of ATOH1 and SPDEF that determines goblet cell differentiation. A better understanding of the genetic nature and how goblet cell differentiation is controlled in colorectal cancers will potentially lead to more precise diagnostics and yield therapeutic targets against cancer.

REFERENCES

Ajioka, Y., H. Watanabe and J. R. Jass (1997). "MUC1 and MUC2 mucins in flat and polypoid colorectal adenomas." J Clin Pathol **50**(5): 417-421.

Albaugh, G. P., V. Iyengar, A. Lohani, M. Malayeri, S. Bala and P. P. Nair (1992). "Isolation of exfoliated colonic epithelial cells, a novel, non-invasive approach to the study of cellular markers." Int J Cancer **52**(3): 347-350.

Albert, T. K., W. Laubinger, S. Muller, F. G. Hanisch, T. Kalinski, F. Meyer and W. Hoffmann (2010). "Human Intestinal TFF3 Forms Disulfide-Linked Heteromers with the Mucus-Associated FCGBP Protein and Is Released by Hydrogen Sulfide." Journal of Proteome Research **9**(6): 3108-3117.

Allen, A., A. Bell, M. Mantle and J. P. Pearson (1982). "The structure and physiology of gastrointestinal mucus." Adv Exp Med Biol **144**: 115-133.

Ambort, D., M. E. Johansson, J. K. Gustafsson, H. E. Nilsson, A. Ermund, B. R. Johansson, P. J. Koeck, H. Hebert and G. C. Hansson (2012). "Calcium and pH-dependent packing and release of the gel-forming MUC2 mucin." Proc Natl Acad Sci U S A **109**(15): 5645-5650.

Andrews, Simon. "FastQC: a quality control tool for high throughput sequence data." (2010): 175-176.

Andrianifahanana, M., N. Moniaux and S. K. Batra (2006). "Regulation of mucin expression: mechanistic aspects and implications for cancer and inflammatory diseases." Biochim Biophys Acta **1765**(2): 189-222.

Aronson, B. E., K. A. Stapleton, L. A. Vissers, E. Stokhuijzen, H. Bruijnzeel and S. D. Krasinski (2014). "Spdef deletion rescues the crypt cell proliferation defect in conditional Gata6 null mouse small intestine." BMC Mol Biol **15**: 3.

Artavanis-Tsakonas, S., M. D. Rand and R. J. Lake (1999). "Notch signaling: cell fate control and signal integration in development." Science **284**(5415): 770-776.

Ashley, N. (2013). "Regulation of intestinal cancer stem cells." Cancer Letters **338**(1): 120-126.

Ashley, N., T. M. Yeung and W. F. Bodmer (2013). "Stem cell differentiation and lumen formation in colorectal cancer cell lines and primary tumors." Cancer Res **73**(18): 5798-5809.

Atuma, C., V. Strugala, A. Allen and L. Holm (2001). "The adherent gastrointestinal mucus gel layer: thickness and physical state in vivo." Am J Physiol Gastrointest Liver Physiol **280**(5): G922-929.

Axelsson, M. A., N. Asker and G. C. Hansson (1998). "O-glycosylated MUC2 monomer and dimer from LS 174T cells are water-soluble, whereas larger MUC2 species formed early during biosynthesis are insoluble and contain nonreducible intermolecular bonds." J Biol Chem **273**(30): 18864-18870.

Ayabe, T., T. Ashida, Y. Kohgo and T. Kono (2004). "The role of Paneth cells and their antimicrobial peptides in innate host defense." Trends Microbiol **12**(8): 394-398.

Belli, S., H. O. Aytac, E. Karagulle, H. Yabanoglu, F. Kayaselcuk and S. Yildirim (2014). "Outcomes of surgical treatment of primary signet ring cell carcinoma of the colon and rectum: 22 cases reviewed with literature." Int Surg **99**(6): 691-698.

Benedix, Frank, et al. "Comparison of 17,641 patients with right-and left-sided colon cancer: differences in epidemiology, perioperative course, histology, and survival." Diseases of the Colon & Rectum 53.1 (2010): 57-64.

Bennett, E. P., U. Mandel, H. Clausen, T. A. Gerken, T. A. Fritz and L. A. Tabak (2012). "Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family." Glycobiology **22**(6): 736-756.

Bergstrom, Kirk SB, et al. "Modulation of intestinal goblet cell function during infection by an attaching and effacing bacterial pathogen." Infection and immunity **76.2** (2008): 796-811.

Betge, J., N. I. Schneider, L. Harbaum, M. J. Pollheimer, R. A. Lindtner, P. Kornprat, M. P. Ebert and C. Langner (2016). "MUC1, MUC2, MUC5AC, and MUC6 in colorectal cancer: expression profiles and clinical significance." Virchows Arch **469**(3): 255-265.

Bevins, C. L. and N. H. Salzman (2011). "Paneth cells, antimicrobial peptides and maintenance of intestinal homeostasis." Nat Rev Microbiol **9**(5): 356-368.

Birchenough, G. M., M. E. Johansson, J. K. Gustafsson, J. H. Bergstrom and G. C. Hansson (2015). "New developments in goblet cell mucus secretion and function." Mucosal Immunol **8**(4): 712-719.

Blatt, E. N., X. H. Yan, M. K. Wuerffel, D. L. Hamilos and S. L. Brody (1999). "Forkhead transcription factor HFH-4 expression is temporally related to ciliogenesis." Am J Respir Cell Mol Biol **21**(2): 168-176.

Bordeaux, Jennifer, et al. "Antibody validation." Biotechniques **48.3** (2010): 197.

Bosman, F. T., Carneiro, F., Hruban, R. H. & Theise, N. D. (Eds) WHO classification of tumours of the digestive system 4th edn (International Agency for Research on Cancer, 2010)

Bozic, I., T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein and M. A. Nowak (2010). "Accumulation of driver and passenger mutations during tumor progression." Proc Natl Acad Sci U S A **107**(43): 18545-18550.

Brownlee, I. A., M. E. Havler, P. W. Dettmar, A. Allen and J. P. Pearson (2003). "Colonic mucus: secretion and turnover in relation to dietary fibre intake." Proc Nutr Soc **62**(1): 245-249.

Buisine, M. P., L. Devisme, T. C. Savidge, C. Gespach, B. Gosselin, N. Porchet and J. P. Aubert (1998). "Mucin gene expression in human embryonic and fetal intestine." Gut **43**(4): 519-524.

Byrd, J. C. and R. S. Bresalier (2004). "Mucins and mucin binding proteins in colorectal cancer." Cancer Metastasis Rev **23**(1-2): 77-99.

Campbell, A. P., M. N. Merrett, M. Kettlewell, N. J. Mortensen and D. P. Jewell (1994). "Expression of colonic antigens by goblet and columnar epithelial cells in ileal pouch mucosa: their association with inflammatory change and faecal stasis." J Clin Pathol **47**(9): 834-838.

Chan, A. T. and E. L. Giovannucci (2010). "Primary prevention of colorectal cancer." Gastroenterology **138**(6): 2029-2043 e2010.

Chand, M. et al. Adjuvant chemotherapy improves overall survival after TME surgery in mucinous carcinoma of the rectum. Eur. J. Surg. Oncol. 40, 240–245 (2014). Nature Reviews Clinical Oncology 13, 361–369 (2016) doi:10.1038/nrclinonc.2015.140

Chang, S. K., A. F. Dohrman, C. B. Basbaum, S. B. Ho, T. Tsuda, N. W. Toribara, J. R. Gum and Y. S. Kim (1994). "Localization of mucin (MUC2 and MUC3) messenger RNA and peptide expression in human normal intestine and colon cancer." Gastroenterology **107**(1): 28-36.

Chen, H., A. K. Nandi, X. Li and C. J. Bieberich (2002). "NKX-3.1 interacts with prostate-derived Ets factor and regulates the activity of the PSA promoter." Cancer Research **62**(2): 338-340.

Chinery, R., P. A. Bates, A. De and P. S. Freemont (1995). "Characterization of the Single-Copy Trefoil Peptides Intestinal Trefoil Factor and Ps2 and Their Ability to Form Covalent Dimers." Febs Letters **357**(1): 50-54.

Chung, Y. T., K. A. Matkowskyj, H. Li, H. Bai, W. Zhang, M. S. Tsao, J. Liao and G. Y. Yang (2012). "Overexpression and oncogenic function of aldo-keto reductase family 1B10 (AKR1B10) in pancreatic carcinoma." Mod Pathol **25**(5): 758-766.

Conesa, Ana, et al. "A survey of best practices for RNA-seq data analysis." Genome biology **17.1** (2016): 13.

Conze, T., A. S. Carvalho, U. Landegren, R. Almeida, C. A. Reis, L. David and O. Soderberg (2010). "MUC2 mucin is a major carrier of the cancer-associated sialyl-Tn antigen in intestinal metaplasia and gastric carcinomas." Glycobiology **20**(2): 199-206.

Cox, M. L., C. L. Schray, C. N. Luster, Z. S. Stewart, P. J. Korytko, M. K. KN, J. D. Paulauskis and R. W. Dunstan (2006). "Assessment of fixatives, fixation, and tissue processing on morphology and RNA integrity." Exp Mol Pathol **80**(2): 183-191.

CRUK. http://info.cancerresearchuk.org/cancerstats/types/bowel/.

Dear, Simon, and Rodger Staden. "A standard file format for data from DNA sequencing instruments." DNA Sequence **3.2** (1992): 107-110.

Delbuono, R., M. Pignatelli, W. F. Bodmer and N. A. Wright (1991). "The Role of the Arginine-Glycine-Aspartic Acid-Directed Cellular-Binding to Type-I Collagen and Rat Mesenchymal Cells in Colorectal Tumor Differentiation." Differentiation **46**(2): 97-103.

Deplancke, B. and H. R. Gaskins (2001). "Microbial modulation of innate defense: goblet cells and the intestinal mucus layer." Am J Clin Nutr **73**(6): 1131S-1141S.

Desbordes, S. C., D. Chandraratna and B. Sanson (2005). "A screen for genes regulating the wingless gradient in Drosophila embryos." Genetics **170**(2): 749-766.

Desbordes S, López-Schier H (2005). "Drosophila Patterning: Delta-Notch Interactions". Encyclopedia of Life Sciences: 4. doi:10.1038/npg.els.0004194.

Dharmani, P., V. Srivastava, V. Kissoon-Singh and K. Chadee (2009). "Role of intestinal mucins in innate host defense mechanisms against pathogens." J Innate Immun **1**(2): 123-135.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Dobin, A. and T. R. Gingeras (2015). "Mapping RNA-seq Reads with STAR." Curr Protoc Bioinformatics **51**: 11 14 11-19.

Dobin, A. and T. R. Gingeras (2016). "Optimizing RNA-Seq Mapping with STAR." Methods Mol Biol **1415**: 245-262.

Dovey, H. F., V. John, J. P. Anderson, L. Z. Chen, P. de Saint Andrieu, L. Y. Fang, S. B. Freedman, B. Folmer, E. Goldbach, E. J. Holsztynska, K. L. Hu, K. L. Johnson-Wood, S. L. Kennedy, D. Kholodenko, J. E. Knops, L. H. Latimer, M. Lee, Z. Liao, I. M. Lieberburg, R. N. Motter, L. C. Mutter, J. Nietz, K. P. Quinn, K. L. Sacchi, P. A. Seubert, G. M. Shopp, E. D. Thorsett, J. S. Tung, J. Wu, S. Yang, C. T. Yin, D. B. Schenk, P. C. May, L. D. Altstiel, M. H. Bender, L. N. Boggs, T. C. Britton, J. C. Clemens, D. L. Czilli, D. K. Dieckman-McGinty, J. J. Droste, K. S. Fuson, B. D. Gitter, P. A. Hyslop, E. M. Johnstone, W. Y. Li, S. P. Little, T. E. Mabry, F. D. Miller and J. E. Audia (2001). "Functional gamma-secretase inhibitors reduce beta-amyloid peptide levels in brain." J Neurochem **76**(1): 173-181.

Esser, C., C. Gottlinger, J. Kremer, C. Hundeiker and A. Radbruch (1995). "Isolation of full-size mRNA from ethanol-fixed cells after cellular immunofluorescence staining and fluorescence-activated cell sorting (FACS)." Cytometry **21**(4): 382-386.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-194.

Fearnhead, N. S., J. L. Wilding and W. F. Bodmer (2002). "Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis." Br Med Bull **64**: 27-43.

Feldman, R. J., V. I. Sementchenko, M. Gayed, M. M. Fraig and D. K. Watson (2003). "Pdef expression in human breast cancer is correlated with invasive potential and altered gene expression." Cancer Res **63**(15): 4626-4631.

Fevr, T., S. Robine, D. Louvard and J. Huelsken (2007). "Wnt/beta-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells." Mol Cell Biol **27**(21): 7551-7559.

Garrett, W. S., J. I. Gordon and L. H. Glimcher (2010). "Homeostasis and Inflammation in the Intestine." Cell **140**(6): 859-870.

Geling, A., H. Steiner, M. Willem, L. Bally-Cuif and C. Haass (2002). "A gamma-secretase inhibitor blocks Notch signaling in vivo and causes a severe neurogenic phenotype in zebrafish." EMBO Rep **3**(7): 688-694.

Gersemann, M., S. Becker, I. Kubler, M. Koslowski, G. X. Wang, K. R. Herrlinger, J. Griger, P. Fritz, K. Fellermann, M. Schwab, J. Wehkamp and E. F. Stange (2009). "Differences in goblet cell differentiation between Crohn's disease and ulcerative colitis." Differentiation **77**(1): 84-94.

Godl, K., M. E. Johansson, M. E. Lidell, M. Morgelin, H. Karlsson, F. J. Olson, J. R. Gum, Jr., Y. S. Kim and G. C. Hansson (2002). "The N terminus of the MUC2 mucin forms trimers that are held together within a trypsin-resistant core fragment." J Biol Chem **277**(49): 47248-47256.

Goldsworthy, S. M., P. S. Stockton, C. S. Trempus, J. F. Foley and R. R. Maronpot (1999). "Effects of fixation on RNA extraction and amplification from laser capture microdissected tissue." Mol Carcinog **25**(2): 86-91.

Gray, G. E., R. S. Mann, E. Mitsiadis, D. Henrique, M. L. Carcangiu, A. Banks, J. Leiman, D. Ward, D. Ish-Horowitz and S. Artavanis-Tsakonas (1999). "Human ligands of the Notch receptor." Am J Pathol **154**(3): 785-794.

Gregorieff, A., D. E. Stange, P. Kujala, H. Begthel, M. van den Born, J. Korving, P. J. Peters and H. Clevers (2009). "The ets-domain transcription factor Spdef promotes maturation of goblet and paneth cells in the intestinal epithelium." Gastroenterology **137**(4): 1333-1345 e1331-1333.

Gött, Peter, et al. "Human trefoil peptides: genomic structure in 21q22. 3 and coordinated expression." European journal of human genetics 4 (1996): 308-315.

Gum, J. R., Jr., J. W. Hicks, N. W. Toribara, B. Siddiki and Y. S. Kim (1994). "Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of the amino terminus and overall sequence similarity to prepro-von Willebrand factor." J Biol Chem **269**(4): 2440-2446.

Hair, Dark Hair D. Light. "Multiple comparisons." (1984).

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Hanski, C., M. Hofmeier, A. Schmitt-Graff, E. Riede, M. L. Hanski, F. Borchard, E. Sieber, F. Niedobitek, H. D. Foss, H. Stein and E. O. Riecken (1997). "Overexpression or ectopic expression of MUC2 is the common property of mucinous carcinomas of the colon, pancreas, breast, and ovary." J Pathol **182**(4): 385-391.

Hanski, C., E. Riede, A. Gratchev, H. D. Foss, C. Bohm, E. Klussmann, M. Hummel, B. Mann, H. J. Buhr, H. Stein, Y. S. Kim, J. Gum and E. O. Riecken (1997). "MUC2 gene suppression in human colorectal carcinomas and their metastases: in vitro evidence of the modulatory role of DNA methylation." Lab Invest **77**(6): 685-695.

Hansson, G. C. and M. E. Johansson (2010). "The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria." Gut Microbes **1**(1): 51-54.

Harada, N., S. Iijima, K. Kobayashi, T. Yoshida, W. R. Brown, T. Hibi, A. Oshima and M. Morikawa (1997). "Human IgGFc binding protein (FcgammaBP) in colonic epithelial cells exhibits mucin-like structure." J Biol Chem **272**(24): 15232-15241.

Harbers, Matthias. "The current status of cDNA cloning." Genomics **91.3** (2008): 232-242.

Hawkins, Nicholas, et al. "CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability." Gastroenterology 122.5 (2002): 1376-1387

Head, Steven R., et al. "Library construction for next-generation sequencing: overviews and challenges." Biotechniques **56.2** (2014): 61.

Heringlake, S., M. Hofdmann, A. Fiebeler, M. P. Manns, W. Schmiegel and A. Tannapfel (2010). "Identification and expression analysis of the aldo-ketoreductase1-B10 gene in primary malignant liver tumours." J Hepatol **52**(2): 220-227.

Herrmann, A., J. R. Davies, G. Lindell, S. Martensson, N. H. Packer, D. M. Swallow and I. Carlstedt (1999). "Studies on the "insoluble" glycoprotein complex from human colon. Identification of reduction-insensitive MUC2 oligomers and C-terminal cleavage." J Biol Chem **274**(22): 15828-15836.

Hoffmann, W., W. Jagla and A. Wiede (2001). "Molecular medicine of TFF-peptides: from gut to brain." Histology and Histopathology **16**(1): 319-334.

Hollingsworth, M. A. and B. J. Swanson (2004). "Mucins in cancer: protection and control of the cell surface." Nat Rev Cancer **4**(1): 45-60.

Hrvatin, S., F. Deng, C. W. O'Donnell, D. K. Gifford and D. A. Melton (2014). "MARIS: method for analyzing RNA following intracellular sorting." PLoS One **9**(3): e89459.

Humphries, A. and N. A. Wright (2008). "Colonic crypt organization and tumorigenesis." Nat Rev Cancer **8**(6): 415-424.

Iacopetta, Barry. "Are there two sides to colorectal cancer?." International journal of cancer 101.5 (2002): 403-408.

Ivanov, S. V., I. Kuzmin, M. H. Wei, S. Pack, L. Geil, B. E. Johnson, E. J. Stanbridge and M. I. Lerman (1998). "Down-regulation of transmembrane carbonic anhydrases in renal cell carcinoma cell lines by wild-type von Hippel-Lindau transgenes." Proc Natl Acad Sci U S A **95**(21): 12596-12601.

Jackerott, M., Y. C. Lee, K. Mollgard, H. Kofod, J. Jensen, S. Rohleder, N. Neubauer, L. W. Gaarn, J. Lykke, R. Dodge, L. T. Dalgaard, B. Sostrup, D. B. Jensen, L. Thim, E. Nexo, P. Thams, H. C. Bisgaard and J. H. Nielsen (2006). "Trefoil factors are expressed in human and rat endocrine pancreas: differential regulation by growth hormone." Endocrinology **147**(12): 5752-5759.

Jagla, W., A. Wiede, K. Dietzmann, K. Rutkowski and W. Hoffmann (2000). "Co-localization of TFF3 peptide and oxytocin in the human hypothalamus." FASEB J **14**(9): 1126-1131.

Jagla, W., A. Wiede, M. Hinz, K. Dietzmann, D. Gulicher, K. L. Gerlach and W. Hoffmann (1999). "Secretion of TFF-peptides by human salivary glands." Cell Tissue Res **298**(1): 161-166.

Jagla, W., A. Wiede and W. Hoffmann (1999). "Localization of TFF3 peptide to porcine conjunctival goblet cells." Cell Tissue Res **296**(3): 525-530.

Jaitin, Diego Adhemar, et al. "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types." Science 343.6172 (2014): 776-779.

Jamur, M. C. and C. Oliver (2010). "Permeabilization of cell membranes." Methods Mol Biol **588**: 63-66.

Jasperson, K. W., T. M. Tuohy, D. W. Neklason and R. W. Burt (2010). "Hereditary and familial colon cancer." Gastroenterology **138**(6): 2044-2058.

Jass, Jeremy R., et al. "Emerging concepts in colorectal neoplasia." *Gastroenterology* 123.3 (2002): 862-876.

Jellema, P., D. A. van der Windt, D. J. Bruinvels, C. D. Mallen, S. J. van Weyenberg, C. J. Mulder and H. C. de Vet (2010). "Value of symptoms and additional diagnostic tests for colorectal cancer in primary care: systematic review and meta-analysis." BMJ **340**: c1269.

Jensen, J., E. E. Pedersen, P. Galante, J. Hald, R. S. Heller, M. Ishibashi, R. Kageyama, F. Guillemot, P. Serup and O. D. Madsen (2000). "Control of endodermal endocrine development by Hes-1." Nat Genet **24**(1): 36-44.

Johansson, M. E., J. M. Larsson and G. C. Hansson (2011). "The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions." Proc Natl Acad Sci U S A **108 Suppl 1**: 4659-4665.

Johansson, M. E. V., K. A. Thomsson and G. C. Hansson (2009). "Proteomic Analyses of the Two Mucus Layers of the Colon Barrier Reveal That Their Main Component,

the Muc2 Mucin, Is Strongly Bound to the Fcgbp Protein." Journal of Proteome Research **8**(7): 3549-3557.

Jones DE, Bevins CL. Defensin-6 mRNA in human Paneth cells: implications for antimicrobial peptides in host defense of the human bowel. FEBS Lett. 1993;315:187–192. doi: 10.1016/0014-5793(93)81160-2

Karam, S. M. (1999). "Lineage commitment and maturation of epithelial cells in the gut." Front Biosci **4**: D286-298.

Katz, J. P., N. Perreault, B. G. Goldstein, C. S. Lee, P. A. Labosky, V. W. Yang and K. H. Kaestner (2002). "The zinc-finger transcription factor Klf4 is required for terminal differentiation of goblet cells in the colon." Development **129**(11): 2619-2628.

Kazanjian, A., T. Noah, D. Brown, J. Burkart and N. F. Shroyer (2010). "Atonal homolog 1 is required for growth and differentiation effects of notch/gamma-secretase inhibitors on normal and cancerous intestinal epithelial cells." Gastroenterology **139**(3): 918-928, 928 e911-916.

Kazanjian, A. and N. F. Shroyer (2011). "NOTCH Signaling and ATOH1 in Colorectal Cancers." Curr Colorectal Cancer Rep **7**(2): 121-127.

Khochbin, S., D. Grunwald, M. Pabion and J. J. Lawrence (1990). "Recovery of RNA from flow-sorted fixed cells." Cytometry **11**(8): 869-874.

Kim, D. H., J. W. Kim, J. H. Cho, S. H. Baek, S. Kakar, G. E. Kim, M. H. Sleisenger and Y. S. Kim (2005). "Expression of mucin core proteins, trefoil factors, APC and p21 in subsets of colorectal polyps and cancers suggests a distinct pathway of pathogenesis of mucinous carcinoma of the colorectum." Int J Oncol **27**(4): 957-964.

Kim, T. H. and R. A. Shivdasani (2011). "Genetic evidence that intestinal Notch functions vary regionally and operate through a common mechanism of Math1 repression." J Biol Chem **286**(13): 11427-11433.

Kim, Y. S. and G. Deng (2008). "Aberrant expression of carbohydrate antigens in cancer: the role of genetic and epigenetic regulation." Gastroenterology **135**(1): 305-309.

Kim, Y. S. and S. B. Ho (2010). "Intestinal goblet cells and mucins in health and disease: recent insights and progress." Curr Gastroenterol Rep **12**(5): 319-330.

Kindon, H., C. Pothoulakis, L. Thim, G. Lynchdevaney and D. K. Podolsky (1995). "Trefoil Peptide Protection of Intestinal Epithelial Barrier Function - Cooperative Interaction with Mucin Glycoprotein." Gastroenterology **109**(2): 516-523.

Kiyohara H, Egami H, Shibata Y, Murata K, Ohshima S, Ogawa M. Light microscopic immunohistochemical analysis of the distribution of group II phospholipase A2 in human digestive organs. J Histochem Cytochem.

Kinoshita, K., D. R. Taupin, H. Itoh and D. K. Podolsky (2000). "Distinct pathways of cell migration and antiapoptotic response to epithelial injury: structure-function analysis of human intestinal trefoil factor." Mol Cell Biol **20**(13): 4680-4690.

Kivela, A., S. Parkkila, J. Saarnio, T. J. Karttunen, J. Kivela, A. K. Parkkila, A. Waheed, W. S. Sly, J. H. Grubb, G. Shah, O. Tureci and H. Rajaniemi (2000). "Expression of a novel transmembrane carbonic anhydrase isozyme XII in normal human gut and colorectal tumors." Am J Pathol **156**(2): 577-584.

Kivela, A. J., J. Saarnio, T. J. Karttunen, J. Kivela, A. K. Parkkila, S. Pastorekova, J. Pastorek, A. Waheed, W. S. Sly, T. S. Parkkila and H. Rajaniemi (2001). "Differential expression of cytoplasmic carbonic anhydrases, CA I and II, and membrane-associated isozymes, CA IX and XII, in normal mucosa of large intestine and in colorectal tumors." Dig Dis Sci **46**(10): 2179-2186.

Kjellev, S. (2009). "The trefoil factor family - small peptides with multiple functionalities." Cellular and Molecular Life Sciences **66**(8): 1350-1369.

Kobayashi, K., Y. Hamada, M. J. Blaser and W. R. Brown (1991). "The molecular configuration and ultrastructural locations of an IgG Fc binding site in human colonic epithelium." J Immunol **146**(1): 68-74.

Kobayashi, K., H. Ogata, M. Morikawa, S. Iijima, N. Harada, T. Yoshida, W. R. Brown, N. Inoue, Y. Hamada, H. Ishii, M. Watanabe and T. Hibi (2002). "Distribution and partial characterisation of IgG Fc binding protein in various mucin producing cells and body fluids." Gut **51**(2): 169-176.

Kouznetsova, I., T. Kalinski, U. Peitz, K. E. Monkemuller, H. Kalbacher, M. Vieth, F. Meyer, A. Roessner, P. Malfertheiner, H. Lippert and W. Hoffmann (2007). "Localization of TFF3 peptide in human esophageal submucosal glands and gastric cardia: differentiation of two types of gastric pit cells along the rostro-caudal axis." Cell Tissue Res **328**(2): 365-374.

Kouznetsova, I., U. Peitz, M. Vieth, F. Meyer, E. M. Vestergaard, P. Malfertheiner, A. Roessner, H. Lippert and W. Hoffmann (2004). "A gradient of TFF3 (trefoil factor family 3) peptide synthesis within the normal human gastric mucosa." Cell Tissue Res **316**(2): 155-165.

Kruskal, Joseph B., and Myron Wish. Multidimensional scaling. Vol. 11. Sage, 1978.

Ku, J. L., Y. K. Shin, D. W. Kim, K. H. Kim, J. S. Choi, S. H. Hong, Y. K. Jeon, S. H. Kim, H. S. Kim, J. H. Park, I. J. Kim and J. G. Park (2010). "Establishment and characterization of 13 human colorectal carcinoma cell lines: mutations of genes and expressions of drug-sensitivity genes and cancer stem cell markers." Carcinogenesis **31**(6): 1003-1009.

Lacroix, B., M. Kedinger, P. Simonassmann, M. Rousset, A. Zweibaum and K. Haffen (1984). "Developmental Pattern of Brush-Border Enzymes in the Human-Fetal Colon

- Correlation with Some Morphogenetic Events." Early Human Development **9**(2): 95-103.

Laffin, B. and J. M. Petrash (2012). "Expression of the Aldo-Ketoreductases AKR1B1 and AKR1B10 in Human Cancers." Front Pharmacol **3**: 104.

Lala, Sanjay, et al. "Crohn's disease and the NOD2 gene: a role for paneth cells." Gastroenterology 125.1 (2003): 47-57.

Lang, T., M. Alexandersson, G. C. Hansson and T. Samuelsson (2004). "Bioinformatic identification of polymerizing and transmembrane mucins in the puffer fish Fugu rubripes." Glycobiology **14**(6): 521-527.

Lang, T., G. C. Hansson and T. Samuelsson (2007). "Gel-forming mucins appeared early in metazoan evolution." Proc Natl Acad Sci U S A **104**(41): 16209-16214.

Langer, G., W. Jagla, W. Behrens-Baumann, S. Walter and W. Hoffmann (1999). "Secretory peptides TFF1 and TFF3 synthesized in human conjunctival goblet cells." Invest Ophthalmol Vis Sci **40**(10): 2220-2224.

Larsson, J. M. H., H. Karlsson, J. G. Crespo, M. E. V. Johansson, L. Eklund, H. Sjovall and G. C. Hansson (2011). "Altered O-glycosylation Profile of MUC2 Mucin Occurs in Active Ulcerative Colitis and Is Associated with Increased Inflammation." Inflammatory Bowel Diseases **17**(11): 2299-2307.

Lemercinier, X., F. W. Muskett, B. Cheeseman, P. B. McIntosh, L. Thim and M. D. Carr (2001). "High-resolution solution structure of human intestinal trefoil factor and functional insights from detailed structural comparisons with the other members of the trefoil family of mammalian cell motility factors." Biochemistry **40**(32): 9552-9559.

Leow, C. C., M. S. Romero, S. Ross, P. Polakis and W. Q. Gao (2004). "Hath1, down-regulated in colon adenocarcinomas, inhibits proliferation and tumorigenesis of colon cancer cells." Cancer Research **64**(17): 6050-6057.

Lievin-Le Moal, V. and A. L. Servin (2006). "The front line of enteric host defense against unwelcome intrusion of harmful microorganisms: mucins, antimicrobial peptides, and microbiota." Clin Microbiol Rev **19**(2): 315-337.

Lindblom, Annika. "Different mechanisms in the tumorigenesis of proximal and distal colon cancers." Current opinion in oncology13.1 (2001): 63-69.

Liu, K., J. Fan and J. Wu (2017). "Forkhead Box Protein J1 (FOXJ1) is Overexpressed in Colorectal Cancer and Promotes Nuclear Translocation of beta-Catenin in SW620 Cells." Med Sci Monit **23**: 856-866.

Lo, Y. H., E. Chung, Z. Li, Y. W. Wan, M. M. Mahe, M. S. Chen, T. K. Noah, K. N. Bell, H. K. Yalamanchili, T. J. Klisch, Z. Liu, J. S. Park and N. F. Shroyer (2017). "Transcriptional Regulation by ATOH1 and its Target SPDEF in the Intestine." Cell Mol Gastroenterol Hepatol **3**(1): 51-71.

Ma, J., D. X. Luo, C. Huang, Y. Shen, Y. Bu, S. Markwell, J. Gao, J. Liu, X. Zu, Z. Cao, Z. Gao, F. Lu, D. F. Liao and D. Cao (2012). "AKR1B10 overexpression in breast cancer: association with tumor size, lymph node metastasis and patient survival and its potential as a novel serum marker." Int J Cancer **131**(6): E862-871.

Ma, Y., Y. Yang, F. Wang, P. Zhang, C. Shi, Y. Zou and H. Qin (2013). "Obesity and risk of colorectal cancer: a systematic review of prospective studies." PLoS One **8**(1): e53916.

MacDonald, T. T., I. Monteleone, M. C. Fantini and G. Monteleone (2011). "Regulation of Homeostasis and Inflammation in the Intestine." Gastroenterology **140**(6): 1768-1775.

Madsen, J., O. Nielsen, I. Tornoe, L. Thim and U. Holmskov (2007). "Tissue localization of human trefoil factors 1, 2, and 3." J Histochem Cytochem **55**(5): 505-513.

Marcos, N. T., E. P. Bennett, J. Gomes, A. Magalhaes, C. Gomes, L. David, I. Dar, C. Jeanneau, S. DeFrees, D. Krustrup, L. K. Vogel, E. H. Kure, J. Burchell, J. Taylor-Papadimitriou, H. Clausen, U. Mandel and C. A. Reis (2011). "ST6GalNAc-I controls expression of sialyl-Tn antigen in gastrointestinal tissues." Front Biosci (Elite Ed) **3**: 1443-1455.

McCauley, H. A. and G. Guasch (2015). "Three cheers for the goblet cell: maintaining homeostasis in mucosal epithelia." Trends Mol Med **21**(8): 492-503.

McDonald, John H. Handbook of biological statistics. Vol. 2. Baltimore, MD: Sparky House Publishing, 2009.

Medeiros, F., C. T. Rigl, G. G. Anderson, S. H. Becker and K. C. Halling (2007). "Tissue handling for genome-wide expression analysis: a review of the issues, evidence, and opportunities." Arch Pathol Lab Med **131**(12): 1805-1816.

Medema, J. P. and L. Vermeulen (2011). "Microenvironmental regulation of stem cells in intestinal homeostasis and cancer." Nature **474**(7351): 318-326.

Metsis, M., A. Cintra, V. Solfrini, P. Ernfors, F. Bortolotti, D. G. Morrasutti, C. G. Ostenson, S. Efendic, B. Agerberth, V. Mutt and et al. (1992). "Molecular cloning of PEC-60 and expression of its mRNA and peptide in the gastrointestinal tract and immune system." J Biol Chem **267**(28): 19829-19832.

Milano, J., J. McKay, C. Dagenais, L. Foster-Brown, F. Pognan, R. Gadient, R. T. Jacobs, A. Zacco, B. Greenberg and P. J. Ciaccio (2004). "Modulation of notch processing by gamma-secretase inhibitors causes intestinal goblet cell metaplasia and induction of genes known to specify gut secretory lineage differentiation." Toxicol Sci **82**(1): 341-358.

Mizoshita, T., T. Tsukamoto, K. I. Inada, N. Hirano, M. Tajika, T. Nakamura, H. Ban and M. Tatematsu (2007). "Loss of MUC2 expression correlates with progression along

the adenoma-carcinoma sequence pathway as well as de novo carcinogenesis in the colon." Histol Histopathol **22**(3): 251-260.

Morey, Jeanine S., James C. Ryan, and Frances M. Van Dolah. "Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR." Biological procedures online 8.1 (2006): 175-193.

Mori, M., R. J. Staniunas, G. F. Barnard, J. M. Jessup, G. D. Steele, Jr. and L. B. Chen (1993). "The significance of carbonic anhydrase expression in human colorectal cancer." Gastroenterology **105**(3): 820-826.

Mortazavi, Ali, et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature methods 5.7 (2008): 621-628.

Muskett, F. W., F. E. May, B. R. Westley and J. Feeney (2003). "Solution structure of the disulfide-linked dimer of human intestinal trefoil factor (TFF3): the intermolecular orientation and interactions are markedly different from those of other dimeric trefoil proteins." Biochemistry **42**(51): 15139-15147.

Nawa, Toru, et al. "Differences between right-and left-sided colon cancer in patient characteristics, cancer morphology and histology." Journal of gastroenterology and hepatology 23.3 (2008): 418-423.

NCI. http://www.cancer.gov/cancertopics/types/colon-and-rectal.

Noah, T. K., A. Kazanjian, J. Whitsett and N. F. Shroyer (2010). "SAM pointed domain ETS factor (SPDEF) regulates terminal differentiation and maturation of intestinal goblet cells." Exp Cell Res **316**(3): 452-465.

Noah, T. K. and N. F. Shroyer (2013). "Notch in the intestine: regulation of homeostasis and pathogenesis." Annu Rev Physiol **75**: 263-288.

Noble, William S. "How does multiple testing correction work?." Nature biotechnology 27.12 (2009): 1135-1137.

Oertel, M., A. Graness, L. Thim, F. Buhling, H. Kalbacher and W. Hoffmann (2001). "Trefoil factor family-peptides promote migration of human bronchial epithelial cells: synergistic effect with epidermal growth factor." Am J Respir Cell Mol Biol **25**(4): 418-424.

Oettgen, P., E. Finger, Z. J. Sun, Y. Akbarali, U. Thamrongsak, J. Boltax, F. Grall, A. Dube, A. Weiss, L. Brown, G. Quinn, K. Kas, G. Endress, C. Kunsch and T. A. Libermann (2000). "PDEF, a novel prostate epithelium-specific Ets transcription factor, interacts with the androgen receptor and activates prostate-specific antigen gene expression." Journal of Biological Chemistry **275**(2): 1216-1225.

Ogura, Y., et al. "Expression of NOD2 in Paneth cells: a possible link to Crohn's ileitis." Gut 52.11 (2003): 1591-1597.

Ohashi, T., M. Idogawa, Y. Sasaki, H. Suzuki and T. Tokino (2013). "AKR1B10, a transcriptional target of p53, is downregulated in colorectal cancers associated with poor prognosis." Mol Cancer Res **11**(12): 1554-1563.

Okudaira, K., S. Kakar, L. Cun, E. Choi, R. Wu Decamillis, S. Miura, M. H. Sleisenger, Y. S. Kim and G. Deng (2010). "MUC2 gene promoter methylation in mucinous and non-mucinous colorectal cancer tissues." Int J Oncol **36**(4): 765-775.

Ouellette, André Joseph. "Paneth cell α-defensins in enteric innate immunity." Cellular and Molecular Life Sciences 68.13 (2011): 2215-2229.

Pan, Y., Z. Ouyang, W. H. Wong and J. C. Baker (2011). "A new FACS approach isolates hESC derived endoderm using transcription factors." PLoS One **6**(3): e17536.

Park, E. T., H. K. Oh, J. R. Gum, Jr., S. C. Crawley, S. Kakar, J. Engel, C. C. Leow, W. Q. Gao and Y. S. Kim (2006). "HATH1 expression in mucinous cancers of the colorectum and related lesions." Clin Cancer Res **12**(18): 5403-5410.

Park, K. S., T. R. Korfhagen, M. D. Bruno, J. A. Kitzmiller, H. Wan, S. E. Wert, G. K. Khurana Hershey, G. Chen and J. A. Whitsett (2007). "SPDEF regulates goblet cell hyperplasia in the airway epithelium." J Clin Invest **117**(4): 978-988.

Park, S. W., G. Zhen, C. Verhaeghe, Y. Nakagami, L. T. Nguyenvu, A. J. Barczak, N. Killeen and D. J. Erle (2009). "The protein disulfide isomerase AGR2 is essential for production of intestinal mucus." Proc Natl Acad Sci U S A **106**(17): 6950-6955.

Parkkila, S., H. Rajaniemi, A. K. Parkkila, J. Kivela, A. Waheed, S. Pastorekova, J. Pastorek and W. S. Sly (2000). "Carbonic anhydrase inhibitor suppresses invasion of renal cancer cells in vitro." Proc Natl Acad Sci U S A **97**(5): 2220-2224.

Pastorek, J., S. Pastorekova, I. Callebaut, J. P. Mornon, V. Zelnik, R. Opavsky, M. Zat'ovicova, S. Liao, D. Portetelle, E. J. Stanbridge and et al. (1994). "Cloning and characterization of MN, a human tumor-associated protein with a domain homologous to carbonic anhydrase and a putative helix-loop-helix DNA binding segment." Oncogene **9**(10): 2877-2888.

Paulsen, F., D. Varoga, A. Paulsen and M. Tsokos (2005). "Trefoil factor family (TFF) peptides of normal human Vater's ampulla." Cell Tissue Res **321**(1): 67-74.

Paulsen, F. P., M. Hinz, U. Schaudig, A. B. Thale and W. Hoffmann (2002). "TFF peptides in the human efferent tear ducts." Invest Ophthalmol Vis Sci **43**(11): 3359-3364.

Pelaseyed, T., J. H. Bergstrom, J. K. Gustafsson, A. Ermund, G. M. Birchenough, A. Schutte, S. van der Post, F. Svensson, A. M. Rodriguez-Pineiro, E. E. Nystrom, C. Wising, M. E. Johansson and G. C. Hansson (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." Immunol Rev **260**(1): 8-20.

Pellegrinet, L., V. Rodilla, Z. Liu, S. Chen, U. Koch, L. Espinosa, K. H. Kaestner, R. Kopan, J. Lewis and F. Radtke (2011). "Dll1- and dll4-mediated notch signaling are required for homeostasis of intestinal stem cells." Gastroenterology **140**(4): 1230-1240 e1231-1237.

Penning, T. M. (2005). "AKR1B10: a new diagnostic marker of non-small cell lung carcinoma in smokers." Clin Cancer Res **11**(5): 1687-1690.

Paterson JC, Watson SH. Paneth cell metaplasia in ulcerative colitis. Am J Pathol. 1961;38:243–249

Podolsky, D. K., D. A. Fournier and K. E. Lynch (1986). "Human colonic goblet cells. Demonstration of distinct subpopulations defined by mucin-specific monoclonal antibodies." J Clin Invest **77**(4): 1263-1271.

Podolsky, D. K., G. Gerken, A. Eyking and E. Cario (2009). "Colitis-Associated Variant of TLR2 Causes Impaired Mucosal Repair Because of TFF3 Deficiency." Gastroenterology **137**(1): 209-220.

Podolsky, D. K. and K. J. Isselbacher (1984). "Glycoprotein composition of colonic mucosa. Specific alterations in ulcerative colitis." Gastroenterology **87**(5): 991-998.

Porter, E. M., C. L. Bevins, D. Ghosh and T. Ganz (2002). "The multifaceted Paneth cell." Cell Mol Life Sci **59**(1): 156-170.

Potten, C. S., C. Booth and D. M. Pritchard (1997). "The intestinal epithelial stem cell: the mucosal governor." Int J Exp Pathol **78**(4): 219-243.

Poulsen, S. S., H. Kissow, K. Hare, B. Hartmann and L. Thim (2005). "Luminal and parenteral TFF2 and TFF3 dimer and monomer in two models of experimental colitis in the rat." Regul Pept **126**(3): 163-171.

Qiao, L. and B. C. Wong (2009). "Role of Notch signaling in colorectal cancer." Carcinogenesis **30**(12): 1979-1986.

Ramsköld, Daniel, et al. "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells." Nature biotechnology 30.8 (2012): 777-782.

Reya, T., S. J. Morrison, M. F. Clarke and I. L. Weissman (2001). "Stem cells, cancer, and cancer stem cells." Nature **414**(6859): 105-111.

Riccio, O., M. E. van Gijn, A. C. Bezdek, L. Pellegrinet, J. H. van Es, U. Zimber-Strobl, L. J. Strobl, T. Honjo, H. Clevers and F. Radtke (2008). "Loss of intestinal crypt progenitor cells owing to inactivation of both Notch1 and Notch2 is accompanied by derepression of CDK inhibitors p27Kip1 and p57Kip2." EMBO Rep **9**(4): 377-383.

Richman, P. I. and W. F. Bodmer (1987). "Monoclonal antibodies to human colorectal epithelium: markers for differentiation and tumour characterization." Int J Cancer **39**(3): 317-328.

Richman, P. I. and W. F. Bodmer (1988). "Control of Differentiation in Human Colorectal-Carcinoma Cell-Lines - Epithelial Mesenchymal Interactions." Journal of Pathology **156**(3): 197-211.

Rinnert, M., M. Hinz, P. Buhtz, F. Reiher, W. Lessel and W. Hoffmann (2010). "Synthesis and localization of trefoil factor family (TFF) peptides in the human urinary tract and TFF2 excretion into the urine." Cell Tissue Res **339**(3): 639-647.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Rockett, John C., and Gary M. Hellmann. "Confirming microarray data—is it really necessary?." Genomics 83.4 (2004): 541-549.

Rosler, S., T. Haase, H. Claassen, U. Schulze, M. Schicht, D. Riemann, J. Brandt, D. Wohlrab, B. Muller-Hilke, M. B. Goldring, S. Sel, D. Varoga, F. Garreis and F. P. Paulsen (2010). "Trefoil factor 3 is induced during degenerative and inflammatory joint disease, activates matrix metalloproteinases, and enhances apoptosis of articular cartilage chondrocytes." Arthritis Rheum **62**(3): 815-825.

Ross M, Pawlina W (2011). Histology: A Text and Atlas (6th ed.). Lippincott Williams & Wilkins. pp. 592–593.

Russell, J. N., J. E. Clements and L. Gama (2013). "Quantitation of gene expression in formaldehyde-fixed and fluorescence-activated sorted cells." PLoS One **8**(9): e73849.

Salzman NH, Underwood MA, Bevins CL. Paneth cells, defensins, and the commensal microbiota: a hypothesis on intimate interplay at the intestinal mucosa. Semin Immunol. 2007;19:70–83. doi: 10.1016/j.smim.2007.04.002.

Schwartz, G. J., A. M. Kittelberger, R. H. Watkins and M. A. O'Reilly (2003). "Carbonic anhydrase XII mRNA encodes a hydratase that is differentially expressed along the rabbit nephron." Am J Physiol Renal Physiol **284**(2): F399-410.

Shi, J. (2007). "Defensins and Paneth cells in inflammatory bowel disease." Inflamm Bowel Dis **13**(10): 1284-1292.

Shinoda, M., M. Shin-Ya, Y. Naito, T. Kishida, R. Ito, N. Suzuki, H. Yasuda, J. Sakagami, J. Imanishi, K. Kataoka, O. Mazda and T. Yoshikawa (2010). "Early-stage blocking of Notch signaling inhibits the depletion of goblet cells in dextran sodium sulfate-induced colitis in mice." Journal of Gastroenterology **45**(6): 608-617.

Shroyer, N. F., M. A. Helmrath, V. Y. Wang, B. Antalffy, S. J. Henning and H. Y. Zoghbi (2007). "Intestine-specific ablation of mouse atonal homolog 1 (Math1) reveals a role in cellular homeostasis." Gastroenterology **132**(7): 2478-2488.

Signore, Michele, and K. Alex Reeder. "Antibody validation by Western blotting." Molecular Profiling: Methods and Protocols(2012): 139-155.

Simmonds, Naomi, et al. "Paneth cell metaplasia in newly diagnosed inflammatory bowel disease in children." BMC gastroenterology 14.1 (2014): 93.

Sims, David, et al. "Sequencing depth and coverage: key considerations in genomic analyses." Nature Reviews Genetics15.2 (2014): 121-132.

Sly, W. S. and P. Y. Hu (1995). "Human carbonic anhydrases and carbonic anhydrase deficiencies." Annu Rev Biochem **64**: 375-401.

Smith, G., F. A. Carey, J. Beattie, M. J. Wilkie, T. J. Lightfoot, J. Coxhead, R. C. Garner, R. J. Steele and C. R. Wolf (2002). "Mutations in APC, Kirsten-ras, and p53--alternative genetic pathways to colorectal cancer." Proc Natl Acad Sci U S A **99**(14): 9433-9438.

Sood, A. K., R. Saxena, J. Groth, M. M. Desouki, C. Cheewakriangkrai, K. J. Rodabaugh, C. S. Kasyapa and J. Geradts (2007). "Expression characteristics of prostate-derived Ets factor support a role in breast and prostate cancer progression." Hum Pathol **38**(11): 1628-1638.

Specian, R. D. and M. G. Oliver (1991). "Functional biology of intestinal goblet cells." Am J Physiol **260**(2 Pt 1): C183-193.

Stanger, B. Z., R. Datar, L. C. Murtaugh and D. A. Melton (2005). "Direct regulation of intestinal fate by Notch." Proc Natl Acad Sci U S A **102**(35): 12443-12448.

Surawicz, C. M., R. C. Haggitt, M. Husseman and L. V. McFarland (1994). "Mucosal biopsy diagnosis of colitis: acute self-limited colitis and idiopathic inflammatory bowel disease." Gastroenterology **107**(3): 755-763.

Sugai, Tamotsu, et al. "Analysis of molecular alterations in left-and right-sided colorectal carcinomas reveals distinct pathways of carcinogenesis: proposal for new molecular profile of colorectal carcinomas." *The Journal of Molecular Diagnostics* 8.2 (2006): 193-201.

Tammali, R., A. B. Reddy, A. Saxena, P. G. Rychahou, B. M. Evers, S. Qiu, S. Awasthi, K. V. Ramana and S. K. Srivastava (2011). "Inhibition of aldose reductase prevents colon cancer metastasis." Carcinogenesis **32**(8): 1259-1267.

Tanaka, H., G. Deng, K. Matsuzaki, S. Kakar, G. E. Kim, S. Miura, M. H. Sleisenger and Y. S. Kim (2006). "BRAF mutation, CpG island methylator phenotype and microsatellite instability occur more frequently and concordantly in mucinous than non-mucinous colorectal cancer." Int J Cancer **118**(11): 2765-2771.

Tanaka, Masanori, et al. "Spatial distribution and histogenesis of colorectal Paneth cell metaplasia in idiopathic inflammatory bowel disease." Journal of gastroenterology and hepatology 16.12 (2001): 1353-1359.

Taupin, D. and D. K. Podolsky (2003). "Trefoil factors: initiators of mucosal healing." Nat Rev Mol Cell Biol **4**(9): 721-732.

Taupin, D. R., K. Kinoshita and D. K. Podolsky (2000). "Intestinal trefoil factor confers colonic epithelial resistance to apoptosis." Proc Natl Acad Sci U S A **97**(2): 799-804.

Terzic, J., S. Grivennikov, E. Karin and M. Karin (2010). "Inflammation and colon cancer." Gastroenterology **138**(6): 2101-2114 e2105.

Thomsen, E. R., J. K. Mich, Z. Yao, R. D. Hodge, A. M. Doyle, S. Jang, S. I. Shehata, A. M. Nelson, N. V. Shapovalova, B. P. Levi and S. Ramanathan (2016). "Fixed single-cell transcriptomic characterization of human radial glial diversity." Nat Methods **13**(1): 87-93.

Tichelaar, J. W., S. E. Wert, R. H. Costa, S. Kimura and J. A. Whitsett (1999). "HNF-3/forkhead homologue-4 (HFH-4) is expressed in ciliated epithelial cells in the developing mouse lung." J Histochem Cytochem **47**(6): 823-832.

Tureci, O., U. Sahin, E. Vollmar, S. Siemer, E. Gottert, G. Seitz, A. K. Parkkila, G. N. Shah, J. H. Grubb, M. Pfreundschuh and W. S. Sly (1998). "Human carbonic anhydrase XII: cDNA cloning, expression, and chromosomal localization of a carbonic anhydrase gene that is overexpressed in some renal cell cancers." Proc Natl Acad Sci U S A **95**(13): 7608-7613.

Turner, D. P., V. J. Findlay, A. D. Kirven, O. Moussa and D. K. Watson (2008). "Global gene expression analysis identifies PDEF transcriptional networks regulating cell migration during cancer progression." Mol Biol Cell **19**(9): 3745-3757.

Turner, J., J. Roger, J. Fitau, D. Combe, J. Giddings, G. V. Heeke and C. E. Jones (2011). "Goblet cells are derived from a FOXJ1-expressing progenitor in a human airway epithelium." Am J Respir Cell Mol Biol **44**(3): 276-284.

Tytgat, K. M., H. A. Buller, F. J. Opdam, Y. S. Kim, A. W. Einerhand and J. Dekker (1994). "Biosynthesis of human colonic mucin: Muc2 is the prominent secretory mucin." Gastroenterology **107**(5): 1352-1363.

van der Flier, L. G. and H. Clevers (2009). "Stem cells, self-renewal, and differentiation in the intestinal epithelium." Annu Rev Physiol **71**: 241-260.

Van der Sluis, M., B. A. E. De Koning, A. C. J. M. De Bruijn, A. Velcich, J. P. P. Meijerink, J. B. Van Goudoever, H. A. Buller, J. Dekker, I. Van Seuningen, I. B. Renes and A. W. C. Einerhand (2006). "Muc2-deficient mice spontaneously develop colitis, indicating that Muc2 is critical for colonic protection." Gastroenterology **131**(1): 117-129.

van Es, J. H., N. de Geest, M. van de Born, H. Clevers and B. A. Hassan (2010). "Intestinal stem cells lacking the Math1 tumour suppressor are refractory to Notch inhibitors." Nat Commun **1**: 18.

van Es, J. H., M. E. van Gijn, O. Riccio, M. van den Born, M. Vooijs, H. Begthel, M. Cozijnsen, S. Robine, D. J. Winton, F. Radtke and H. Clevers (2005). "Notch/gamma-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells." Nature **435**(7044): 959-963.

van Tetering, G., P. van Diest, I. Verlaan, E. van der Wall, R. Kopan and M. Vooijs (2009). "Metalloprotease ADAM10 is required for Notch1 site 2 cleavage." J Biol Chem **284**(45): 31018-31027.

VanDussen, K. L., A. J. Carulli, T. M. Keeley, S. R. Patel, B. J. Puthoff, S. T. Magness, I. T. Tran, I. Maillard, C. Siebel, A. Kolterud, A. S. Grosse, D. L. Gumucio, S. A. Ernst, Y. H. Tsai, P. J. Dempsey and L. C. Samuelson (2012). "Notch signaling modulates proliferation and differentiation of intestinal crypt base columnar stem cells." Development **139**(3): 488-497.

VanDussen, K. L. and L. C. Samuelson (2010). "Mouse atonal homolog 1 directs intestinal progenitors to secretory cell rather than absorptive cell fate." Dev Biol **346**(2): 215-223.

Velcich, A., W. Yang, J. Heyer, A. Fragale, C. Nicholas, S. Viani, R. Kucherlapati, M. Lipkin, K. Yang and L. Augenlicht (2002). "Colorectal cancer in mice genetically deficient in the mucin Muc2." Science **295**(5560): 1726-1729.

Verdugo, P. (1990). "Goblet cells secretion and mucogenesis." Annu Rev Physiol **52**: 157-176.

Vogelstein, B., E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger, M. Leppert, Y. Nakamura, R. White, A. M. Smits and J. L. Bos (1988). "Genetic alterations during colorectal-tumor development." N Engl J Med **319**(9): 525-532.

Wapenaar, M. C., A. J. Monsuur, J. Poell, R. van 't Slot, J. W. Meijer, G. A. Meijer, C. J. Mulder, M. L. Mearin and C. Wijmenga (2007). "The SPINK gene family and celiac disease susceptibility." Immunogenetics **59**(5): 349-357.

Wenzel, U. A., M. K. Magnusson, A. Rydstrom, C. Jonstrand, J. Hengst, M. E. Johansson, A. Velcich, L. Ohman, H. Strid, H. Sjovall, G. C. Hansson and M. J. Wick (2014). "Spontaneous colitis in Muc2-deficient mice reflects clinical and cellular features of active ulcerative colitis." PLoS One **9**(6): e100217.

Wiede, A., M. Hinz, E. Canzler, K. Franke, C. Quednow and W. Hoffmann (2001). "Synthesis and localization of the mucin-associated TFF-peptides in the human uterus." Cell Tissue Res **303**(1): 109-115.

Winawer, S. J., A. G. Zauber, M. N. Ho, M. J. O'Brien, L. S. Gottlieb, S. S. Sternberg, J. D. Waye, M. Schapiro, J. H. Bond, J. F. Panish and et al. (1993). "Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup." N Engl J Med **329**(27): 1977-1981.

Wong, W. M., R. Poulsom and N. A. Wright (1999). "Trefoil peptides." Gut **44**(6): 890-895.

Wu, H. J. and E. Wu (2012). "The role of gut microbiota in immune homeostasis and autoimmunity." Gut Microbes **3**(1): 4-14.

Wu, Y., C. Cain-Hom, L. Choy, T. J. Hagenbeek, G. P. de Leon, Y. Chen, D. Finkle, R. Venook, X. Wu, J. Ridgway, D. Schahin-Reed, G. J. Dow, A. Shelton, S. Stawicki, R. J. Watts, J. Zhang, R. Choy, P. Howard, L. Kadyk, M. Yan, J. Zha, C. A. Callahan, S. G. Hymowitz and C. W. Siebel (2010). "Therapeutic antibody targeting of individual Notch receptors." Nature **464**(7291): 1052-1057.

Xue, H., B. J. Li, J. Zhang, M. L. Wu, Q. Huang, Q. Wu, H. Q. Sheng, D. D. Wu, J. W. Hu and M. D. Lai (2010). "Identification of Serum Biomarkers for Colorectal Cancer Metastasis Using a Differential Secretome Approach." Journal of Proteome Research **9**(1): 545-555.

Yamada, N., Y. Tamai, H. Miyamoto and M. Nozaki (2000). "Cloning and expression of the mouse Pse gene encoding a novel Ets family member." Gene **241**(2): 267-274.

Yang, Q., N. A. Bermingham, M. J. Finegold and H. Y. Zoghbi (2001). "Requirement of Math1 for secretory cell lineage commitment in the mouse intestine." Science **294**(5549): 2155-2158.

Yao, H. B., Y. Xu, L. G. Chen, T. P. Guan, Y. Y. Ma, X. J. He, Y. J. Xia, H. Q. Tao and Q. S. Shao (2014). "AKR1B10, a good prognostic indicator in gastric cancer." Eur J Surg Oncol **40**(3): 318-324.

Yeung, T. M., S. C. Gandhi and W. F. Bodmer (2011). "Hypoxia and lineage specification of cell line-derived colorectal cancer stem cells." Proceedings of the National Academy of Sciences of the United States of America **108**(11): 4382-4387.

Yeung, T. M., S. C. Gandhi, J. L. Wilding, R. Muschel and W. F. Bodmer (2010). "Cancer stem cells from colorectal cancer-derived cell lines." Proc Natl Acad Sci U S A **107**(8): 3722-3727.

Young B, Woodford P, O'Dowd G (2013). Wheater's Functional Histology: A Text and Colour Atlas (6th ed.). Elsevier. p. 94. ISBN 978-0702047473.

Zhang, Y., L. Zhang, H. Sun, Q. Lv, C. Qiu, X. Che, Z. Liu and J. Jiang (2017). "Forkhead transcription factor 1 inhibits endometrial cancer cell proliferation via sterol regulatory element-binding protein 1." Oncol Lett **13**(2): 731-737.

Zhao, F., R. Edwards, D. Dizon, K. Afrasiabi, J. R. Mastroianni, M. Geyfman, A. J. Ouellette, B. Andersen and S. M. Lipkin (2010). "Disruption of Paneth and goblet cell homeostasis and increased endoplasmic reticulum stress in Agr2-/-mice." Developmental Biology **338**(2): 268-277.

Zheng, W., P. Rosenstiel, K. Huse, C. Sina, R. Valentonyte, N. Mah, L. Zeitlmann, J. Grosse, N. Ruf, P. Nurnberg, C. M. Costello, C. Onnie, C. Mathew, M. Platzer, S. Schreiber and J. Hampe (2006). "Evaluation of AGR2 and AGR3 as candidate genes for inflammatory bowel disease." Genes Immun **7**(1): 11-18.

Zheng, X., K. Tsuchiya, R. Okamoto, M. Iwasaki, Y. Kano, N. Sakamoto, T. Nakamura and M. Watanabe (2011). "Suppression of hath1 gene expression directly regulated by

hes1 via notch signaling is associated with goblet cell depletion in ulcerative colitis." Inflamm Bowel Dis **17**(11): 2251-2260.

APPENDIX

Figure Appendix.1 Representative PR5D5 staining in 64 CRC cell lines
5000 cells of each cell line were seeded into 96-well plates and grown for three days before intracellular staining using PR5D5 (1:200).

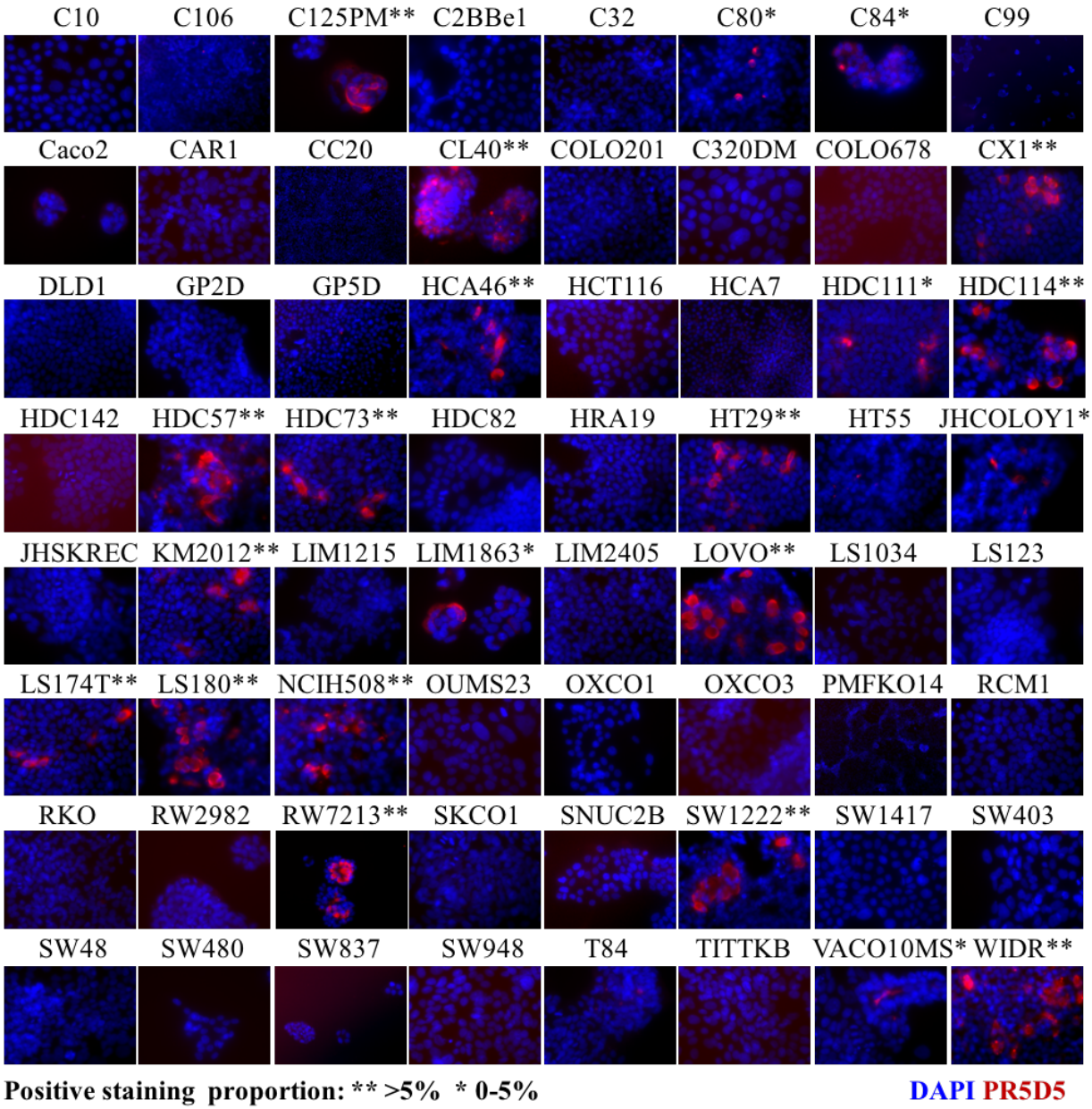


Figure Appendix.2 Representative MUC2D staining in 64 CRC cell lines
5000 cells of each cell line were seeded into 96-well plates and grown for three days before intracellular staining using MUC2D (1:200).

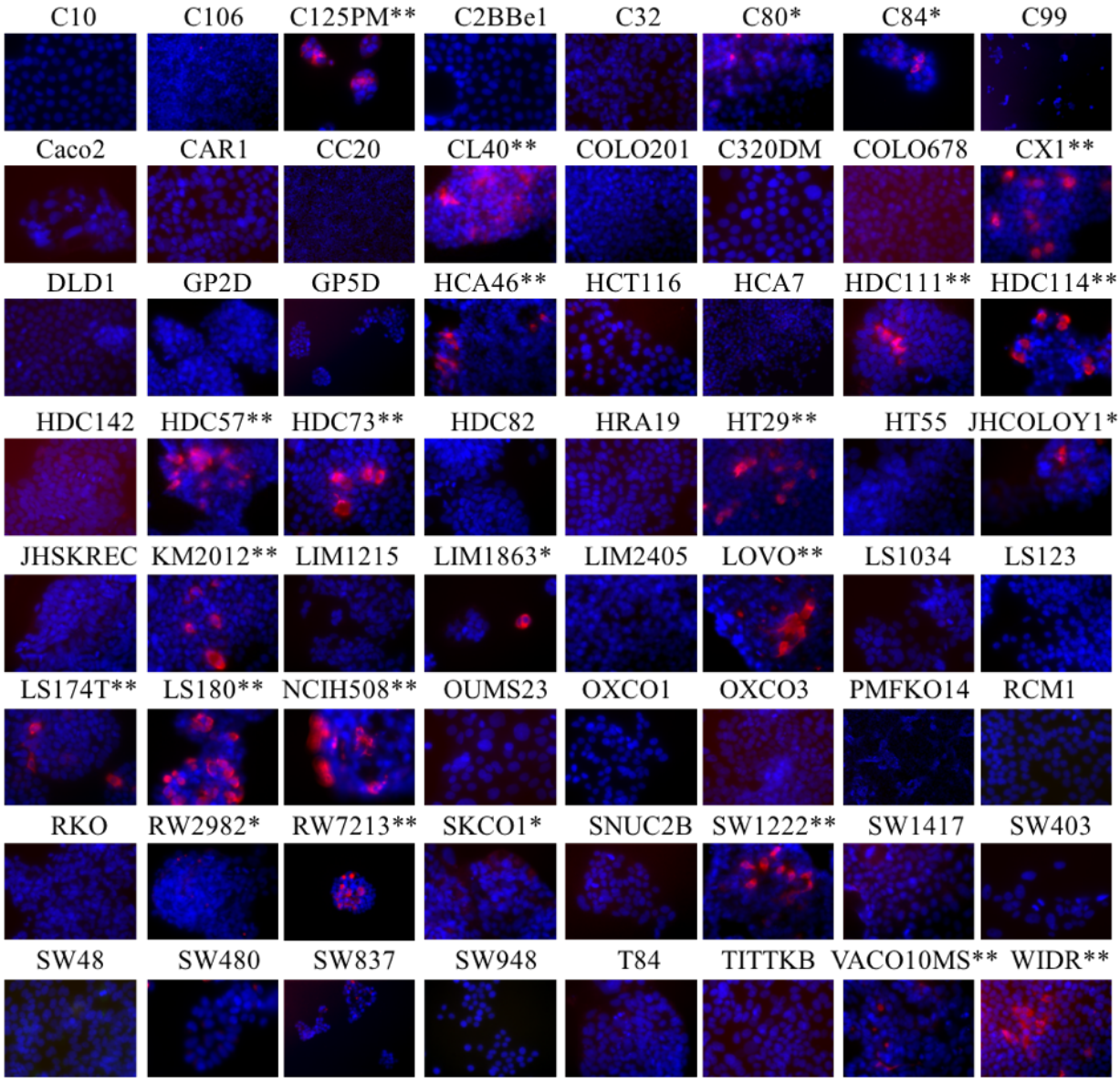


Table Appendix.1 List of 500 most significantly up-regulated genes in goblet cell-positive cell lines, sorted by fold change. Fold change given as goblet cell-positive versus negative cell lines. P value adjusted for multiple testing using false discovery rate FDR step up.

Gene Symbol	Set-up p-value	Fold change
AGR3	0.0083	22.7101
LGALS4	0.0276	21.4264
AKR1B10	0.0032	16.8033
AGR2	0.0285	15.0087
TSPAN8	0.0599	12.0510
GPX2	0.0557	11.1586
UGT1A1	0.0323	11.0559
TOX3	0.0351	10.6526
CA12	0.0083	10.5218
SPINK1	0.0276	10.2973
CEACAM5	0.0736	10.2805
ALDH1A1	0.0560	9.9991
HMGCS2	0.0577	9.6027
REG4	0.0438	9.4208
TOX3	0.0414	9.3540
REG4	0.0273	9.3330
HEPH	0.0754	9.0945
CEACAM6	0.0599	8.9000
CA12	0.0073	8.3755
NOX1	0.0629	8.3553
CA12	0.0032	8.3397
UGT1A1	0.0353	8.2421
CEACAM6	0.0577	8.2160
PPP1R1B	0.0448	7.8870
UGT1A1	0.0399	7.6074
TRIM31	0.0379	7.5888
PIGR	0.0423	7.5621
CFTR	0.0567	7.4941
UGT1A1	0.0423	7.2898
UGT1A1	0.0333	7.1103
TOX3	0.0468	7.0974
CA12	0.0032	7.0336
CCL14-CCL15 /// CCL15	0.0512	6.4191
GCNT3	0.0423	6.2379
CDC42EP5	0.0242	6.1225
ST6GALNAC1	0.0532	6.0752
LYZ	0.2813	6.0627
S100P	0.0639	5.9525
FABP1	0.2748	5.8113
BHLHE41	0.0276	5.7750
C9orf152	0.0887	5.7623
CLDN2	0.0543	5.7592
KRT20	0.1835	5.6910
CLRN3	0.1923	5.6776
CA12	0.0083	5.6063
AGR2	0.0238	5.5977
KIAA1199	0.1296	5.4600
C11orf93	0.0276	5.4261
CFTR	0.0923	5.4241
UGT1A8 /// UGT1A9	0.0276	5.3177
LYZ	0.3069	5.2783
DDC	0.1566	5.2669
AKR1C3	0.2155	5.2339
LCN2	0.1211	5.2200
FCGBP	0.0182	5.1814
MAOA	0.0363	5.1432
SLC44A4	0.0438	5.0520
PAPSS2	0.1257	4.9974

MECOM	0.0314	4.9583
MECOM	0.0276	4.9248
GATA6	0.0647	4.7510
IL33	0.0918	4.6555
NOX1	0.0858	4.6344
ATP10B	0.1330	4.5977
---	0.0356	4.5532
CYP2B6 /// CYP2B7P1	0.1109	4.5232
CYP3A5	0.0323	4.4349
ETS2	0.0262	4.4301
TNFRSF11A	0.0457	4.3879
PAPSS2	0.1151	4.3320
ACSM3	0.0291	4.3302
MAOA	0.0672	4.2630
TRIM31	0.0411	4.1553
---	0.1034	4.1199
CYP3A5	0.0273	4.1084
UGT1A6	0.1668	4.0791
SYTL5	0.0973	3.9469
CYP3A5	0.0341	3.9188
DEPDC6	0.0438	3.8906
EHF	0.1148	3.8541
HNMT	0.1461	3.8520
TM4SF4	0.1031	3.8399
MUC2	0.0032	3.8060
OLFM4	0.4938	3.8023
POF1B	0.0736	3.7948
ATP8B1	0.0379	3.7695
CRIP1	0.1082	3.7545
POF1B	0.1637	3.7478
AKR1C2	0.1109	3.7338
AKR1C2	0.1312	3.6901
---	0.0515	3.6619
GPA33	0.2950	3.6558
C15orf48	0.3292	3.6445
CDH17	0.4615	3.6204
SLC40A1	0.2839	3.6132
PRSS3	0.0588	3.6122
NOSTRIN	0.0406	3.5766
GPX2	0.0560	3.5747
PTPRO	0.3246	3.5692
MAOA	0.0602	3.5570
IQGAP2	0.1737	3.5500
ANXA13	0.1037	3.5498
FBP1	0.0714	3.5303
TRIM31	0.0388	3.5177
ERBB3	0.0413	3.5039
CALB1	0.3763	3.4936
LOC100505633	0.1439	3.4892
C10orf99	0.3383	3.4792
ACSM3	0.0578	3.4694
CALB1	0.4697	3.4629
FAM3D	0.0423	3.4593
IFI27	0.3789	3.4495
CALML4	0.0349	3.4344
CD44	0.1256	3.4320
FAR2	0.0520	3.4303
FA2H	0.0083	3.4214
XIST	0.3745	3.4188
MYB	0.1630	3.4159
C10orf99	0.2902	3.3783
SLC12A2	0.0238	3.3463
FAM84A	0.4640	3.3351
FOXA2	0.0825	3.3325
AKR1C1	0.1296	3.3273
ACSL5	0.1290	3.3253
LGR5	0.3128	3.3207

ATP2C2	0.0341	3.3204
AZGP1	0.2657	3.3200
PIP5K1B	0.1916	3.3165
ERBB3	0.0468	3.3051
RBM47	0.0276	3.3038
ARL14	0.0543	3.3019
SAMD5	0.2899	3.2990
HNF4G	0.0332	3.2923
SLC44A4	0.0822	3.2727
SLC44A3	0.0622	3.2634
BHLHE41	0.0468	3.2537
VIL1	0.3275	3.2440
CALML4	0.0323	3.2291
CD24	0.1814	3.2252
PTPRB	0.0522	3.2207
CA2	0.2073	3.1981
TMEM45B	0.1361	3.1946
GALNT3	0.0878	3.1924
CDH1	0.1774	3.1812
XIST	0.4872	3.1806
LOC100507192	0.1446	3.1792
CYP2B6	0.1907	3.1788
NFIB	0.0622	3.1731
PPARG	0.0340	3.1653
REPS2	0.0351	3.1573
LRRC31	0.1051	3.1363
CLDN3	0.1037	3.1254
MUC13	0.3718	3.1237
MUC13	0.2447	3.1136
GMDS	0.0117	3.1070
AKR1C1	0.1348	3.1005
FXYD3	0.1724	3.0928
LOC100505946	0.0522	3.0921
DENND2D	0.0083	3.0892
BCL2L14	0.0622	3.0844
CEACAM1	0.1966	3.0785
ADRA2A	0.0560	3.0785
ETS2	0.0291	3.0772
C2orf89	0.1425	3.0701
CD44	0.0656	3.0653
RASEF	0.0423	3.0639
ADH1C	0.1463	3.0633
PRSS3	0.0825	3.0617
RAB25	0.1089	3.0513
SGPP2	0.0598	3.0492
SGPP2	0.0560	3.0386
NR3C2	0.0192	3.0350
GIPC2	0.2652	3.0336
CD24	0.1694	3.0332
AUTS2	0.3285	3.0298
CD24	0.1761	3.0245
KLF5	0.0399	3.0231
KLK10	0.3681	3.0195
PRR15L	0.1689	3.0176
ANKRD22	0.1885	2.9863
KLK11	0.0556	2.9827
ANXA10	0.1630	2.9825
PCK1	0.4765	2.9632
SAMD13	0.0276	2.9629
ELF3	0.0457	2.9567
RBM47	0.0294	2.9404
MOBKL2B	0.0560	2.9372
---	0.1502	2.9289
C4orf19	0.0734	2.9199
RASEF	0.1196	2.9150
IL1R2	0.2626	2.9128
SEL1L3	0.0652	2.9064

NUPR1	0.1936	2.8859
RARRES1	0.3065	2.8715
APOH	0.0273	2.8710
NCRNA00261	0.0819	2.8684
VAV3	0.4036	2.8569
ALOX5	0.2044	2.8550
WNK4	0.0656	2.8526
SLC12A2	0.0782	2.8519
SLITRK6	0.1188	2.8502
---	0.0340	2.8491
TDGF1 /// TDGF3	0.4240	2.8475
LOC100506781	0.1206	2.8367
AIM1	0.1228	2.8261
BHLHE41	0.0228	2.8196
---	0.0443	2.8151
ALPK3	0.0276	2.8116
TTC22	0.0672	2.8087
LOC400573	0.1595	2.8013
CXCL3	0.2320	2.7929
IGFBP4	0.1755	2.7909
HNMT	0.1034	2.7826
NFIB	0.1535	2.7806
FOXA3	0.0885	2.7780
CD24	0.2660	2.7698
CD24	0.1927	2.7659
PDZK1IP1	0.1576	2.7574
VIL1	0.4520	2.7534
RSPH1	0.0238	2.7467
CKB	0.1109	2.7450
BCL11B	0.1813	2.7387
SLITRK6	0.1161	2.7350
ACSL5	0.2296	2.7340
SLITRK6	0.1161	2.7280
FAM84A	0.4015	2.7277
METTL7A	0.3843	2.7264
CLDN3	0.1054	2.7110
FAM13A	0.1781	2.7099
SGK2	0.1654	2.7098
XIST	0.5495	2.7065
USH1C	0.3192	2.7041
MOBKL2B	0.0591	2.7008
CEACAM7	0.1966	2.6983
HKDC1	0.3280	2.6949
ALDH3A1	0.1690	2.6927
LOC100506966	0.0423	2.6860
LGR4	0.1061	2.6832
CCND2	0.6256	2.6697
TSPAN1	0.0750	2.6673
GPR160	0.1905	2.6544
EPS8L3	0.3928	2.6500
SERPINB1	0.0323	2.6485
CTSE	0.1963	2.6454
TMC5	0.1057	2.6406
XK	0.1914	2.6392
FOXA2	0.2187	2.6387
LXN	0.2141	2.6220
GIPC2	0.2379	2.6202
NOX1	0.1336	2.6137
FAM110C	0.1227	2.6117
VAV3	0.3765	2.6110
ARSE	0.2029	2.6032
PTPRH	0.0207	2.6021
PARM1	0.3427	2.6007
GABRP	0.0523	2.5979
ALPK3	0.0238	2.5947
CAPN8	0.2322	2.5892
LOC283352	0.2485	2.5883

ELF3	0.0820	2.5859
CREB3L1	0.0656	2.5850
---	0.1703	2.5830
MLLT3	0.1412	2.5808
GDA	0.3717	2.5683
SERPINB1	0.0520	2.5581
RASSF6	0.2348	2.5578
KLF5	0.0670	2.5577
TCEA3	0.3024	2.5572
STARD10	0.0768	2.5534
PRR5L	0.0853	2.5463
CLDN1	0.4036	2.5394
ANXA4	0.0294	2.5379
TINAG	0.1774	2.5354
KRTCAP3	0.0991	2.5332
ESRP1	0.2508	2.5303
CEACAM1	0.1991	2.5273
FAM105A	0.2121	2.5271
CYP4X1	0.2751	2.5253
TMC5	0.0792	2.5209
GMDS	0.0182	2.5193
STARD10	0.0390	2.5185
NRARP	0.0714	2.5165
CYP1B1	0.3880	2.5161
LRRC19	0.2445	2.5135
S100A14	0.2725	2.5100
ABCC3	0.0785	2.5079
GDF15	0.2653	2.5027
ATP1B1	0.0864	2.5003
NFIB	0.1568	2.4992
NFIB	0.1500	2.4968
NOX1	0.2524	2.4962
TFF3	0.4045	2.4814
C4orf19	0.1561	2.4806
ELF3	0.0670	2.4762
SLC3A1	0.2723	2.4760
EIF1AY	0.5900	2.4753
C1orf21	0.2310	2.4722
C9orf150	0.1115	2.4721
GFPT1	0.0059	2.4685
XIST	0.5099	2.4665
FUT3	0.2322	2.4636
TNFRSF11A	0.0469	2.4608
LY75	0.4440	2.4585
VSIG2	0.0337	2.4555
SIAE	0.0918	2.4552
MAL2	0.2039	2.4524
EPS8	0.0337	2.4477
FAR2	0.2410	2.4432
GRTP1	0.1412	2.4368
XIST	0.5705	2.4331
CD24	0.3119	2.4321
PRR15	0.2228	2.4321
C2orf82	0.1054	2.4298
ONECUT2	0.0708	2.4295
---	0.1632	2.4235
RASSF6	0.2145	2.4147
CYP39A1	0.1362	2.4135
CFTR	0.0622	2.4107
CD44	0.2505	2.4083
ETV6	0.0083	2.4069
CYP39A1	0.0719	2.3968
KIAA1244	0.0715	2.3968
DNAJC12	0.3122	2.3908
CD44	0.2567	2.3902
MYH14	0.0323	2.3865
MEST	0.1633	2.3845

NME7	0.1447	2.3841
GALNT4	0.0781	2.3828
TMEM45B	0.2108	2.3744
FOXP1	0.2493	2.3734
PRKACB	0.4548	2.3730
CLDN7	0.1135	2.3730
PYCARD	0.1991	2.3671
SYTL2	0.2842	2.3669
TESC	0.3083	2.3669
ETHE1	0.0351	2.3645
CALML4	0.0117	2.3618
SYTL2	0.3113	2.3600
BTNL9	0.1045	2.3584
MUC3B	0.0944	2.3542
SYK	0.1644	2.3542
NFIA	0.0999	2.3507
FERMT1	0.0807	2.3506
FOS	0.2157	2.3499
NR1I2	0.4328	2.3497
TP53I11	0.0182	2.3470
PYGB	0.0399	2.3468
EPCAM	0.1837	2.3426
NFIA	0.0652	2.3405
CLMN	0.1667	2.3386
GFPT1	0.0083	2.3318
UGT8	0.2095	2.3287
---	0.1222	2.3255
PTPRK	0.0992	2.3253
GUCY2C	0.5038	2.3238
GGT6	0.2419	2.3234
FAM13A	0.1898	2.3228
RNF128	0.5720	2.3210
XIST	0.4593	2.3183
FABP6	0.2990	2.3123
MAML3	0.1095	2.3079
EHF	0.3259	2.3071
HSD17B2	0.4317	2.3066
CELF2	0.2509	2.3024
MYO5B	0.0396	2.2986
PLEKHA6	0.0864	2.2965
CD44	0.2475	2.2957
TC2N	0.0622	2.2914
PKDCC	0.2316	2.2901
MUC5AC	0.0332	2.2899
FERMT1	0.0754	2.2892
ABCC3	0.0853	2.2876
SEL1L3	0.1064	2.2867
OAS1	0.2679	2.2862
IL1RN	0.2399	2.2856
ITGB6	0.3424	2.2855
AIFM3	0.0925	2.2830
GRIN2D	0.1463	2.2805
AKAP7	0.1668	2.2797
GSTT1	0.3800	2.2794
PDK4	0.1898	2.2789
MYH14	0.0639	2.2764
---	0.0679	2.2739
RARRES1	0.3972	2.2737
---	0.0924	2.2715
MYO1D	0.1190	2.2705
ATP1B1	0.0522	2.2691
---	0.0182	2.2674
KCNQ1	0.1223	2.2658
DPP4	0.3887	2.2621
MYH14	0.0750	2.2599
TMC4	0.0652	2.2571
IL1R2	0.3494	2.2560

CYP1B1	0.3840	2.2527
FA2H	0.0276	2.2500
SERPINB1	0.0622	2.2454
PDZK1IP1	0.2296	2.2429
EPPK1	0.3000	2.2423
---	0.0470	2.2404
---	0.0923	2.2401
NFIB	0.1666	2.2399
DDC	0.2322	2.2398
PDE10A	0.1577	2.2380
NFIA	0.0975	2.2368
CACNA1D	0.1572	2.2315
PIGR	0.1367	2.2291
TMPRSS2	0.3845	2.2272
CYP2C18	0.0315	2.2270
---	0.1445	2.2268
MLLT3	0.2805	2.2218
EPPK1	0.2839	2.2212
SLC22A18	0.0522	2.2210
RAB11FIP4	0.0337	2.2177
CD200	0.3329	2.2147
CD44	0.2807	2.2129
PTGER4	0.3569	2.2118
FAM129A	0.5118	2.2108
---	0.4566	2.2094
ABCB1 /// ABCB4	0.5123	2.2081
Mar-03	0.0719	2.2046
CCND2	0.5430	2.2040
LOC151009	0.0577	2.2013
TSPAN12	0.2000	2.1998
ARSJ	0.1580	2.1997
LOC730102	0.3139	2.1996
TFF1	0.4264	2.1957
ZG16B	0.2142	2.1914
ANKRD22	0.2596	2.1901
RPS4Y1	0.5881	2.1899
MANSC1	0.1991	2.1896
IHH	0.1211	2.1893
DPEP1	0.4805	2.1883
CA9	0.1986	2.1855
DHRS11	0.0207	2.1820
SULT1C2	0.3765	2.1806
TPPP	0.0278	2.1800
MGC11082	0.0853	2.1781
TTC39A	0.0946	2.1758
DPP4	0.4240	2.1671
SELENBP1	0.5490	2.1649
CYP2S1	0.2347	2.1636
KBTBD11	0.2377	2.1601
SLC43A1	0.0754	2.1557
PLCB4	0.5204	2.1475
FOXP1	0.3169	2.1445
SIDT1	0.2919	2.1438
---	0.0656	2.1437
SPINK4	0.4587	2.1405
ZBED3	0.0358	2.1399
TNFSF13	0.0842	2.1394
FTH1	0.0853	2.1393
USH1C	0.4440	2.1391
EPHB3	0.2923	2.1385
HHLA2	0.1296	2.1319
VEGFA	0.0917	2.1251
FUT4	0.1666	2.1206
---	0.0276	2.1198
---	0.0323	2.1176
FGD4	0.0225	2.1160
CXCL2	0.4566	2.1128

OAS1	0.4273	2.1123
COBLL1	0.1017	2.1105
SLC5A1	0.2760	2.1081
TC2N	0.1336	2.1074
PRSS2	0.2937	2.1068
CBLC	0.0622	2.1057
SPRR1A	0.2681	2.1042
MYH14	0.0762	2.1031
LNX1	0.2155	2.1001
---	0.0995	2.0989
TSPAN12	0.3681	2.0983
EHF	0.3651	2.0961
SAMD5	0.5123	2.0958
NBL1	0.0807	2.0947
PLXNA2	0.0880	2.0937
FUT3	0.2919	2.0933
ELL3 /// SERINC4	0.0520	2.0927
NRIP1	0.1251	2.0919
C2CD4A	0.4407	2.0907
PIK3R1	0.0532	2.0906
SIPA1L2	0.3026	2.0902
NRIP1	0.1038	2.0883
AHCYL2	0.1064	2.0862
SATB1	0.5036	2.0861
GBP3	0.4309	2.0858
RAP1GAP	0.0887	2.0857
LOC339290	0.0827	2.0854
EHF	0.3136	2.0830
ABHD12B	0.6715	2.0825
SEMA4G	0.1905	2.0813
SLC16A5	0.1666	2.0810
ISX	0.4240	2.0802
C7orf46	0.1925	2.0773
RHBDL2	0.2377	2.0766
ARHGAP32	0.0380	2.0742
TST	0.1064	2.0695
SLC7A11	0.2295	2.0686
AMACR	0.4094	2.0673
FAM134B	0.3304	2.0663
---	0.3641	2.0661

Table Appendix.2 List of 350 most significantly up-regulated genes in goblet cell-negative cell lines, sorted by fold change. Fold change given as goblet cell-positive versus negative cell lines. P value adjusted for multiple testing using false discovery rate FDR step up.

Gene Symbol	Set-up p-value	Fold change
SLC2A3	0.0559834	-8.23727
AKAP12	0.0994298	-7.18119
SLC2A14 /// SLC2A3	0.0609784	-6.58308
WWTR1	0.0467359	-6.42299
SLC2A3	0.0754027	-6.11255
SLC2A14 /// SLC2A3	0.0634405	-5.6679
FAM171B	0.0655816	-5.4227
CAV1	0.222196	-5.13698
FSCN1	0.0622081	-5.11668
SLC2A3	0.0789853	-5.02937
FN1	0.0995123	-4.96047
FERMT2	0.16542	-4.75451
FN1	0.102874	-4.58239
CALD1	0.250461	-4.56879
OBSL1	0.0931812	-4.55184
PHLDB2	0.238162	-4.54109
TNFRSF19	0.0675811	-4.5312
ANXA6	0.0212059	-4.51974
GJC1	0.166701	-4.41543
CYR61	0.092948	-4.38298
TIMP2	0.162686	-4.27106
FN1	0.0972997	-4.1338
CHST15	0.14669	-4.06861
AKAP12	0.223109	-4.04158
FRMD6	0.0824608	-4.03986
CAV1	0.267998	-4.03961
TUBA1A	0.166468	-4.01164
TUBB6	0.254974	-3.96823
FN1	0.0841101	-3.94785
DLC1	0.0468292	-3.92725
TIMP2	0.149999	-3.92433
CYR61	0.0912938	-3.87882
TFAP2C	0.0550027	-3.7577
MYO5A	0.0828335	-3.7313
GJA1	0.269622	-3.72358
FRMD6	0.103727	-3.68441
TMEM45A	0.044755	-3.67527
ZBED2	0.183667	-3.67327
SRPX	0.0872737	-3.66296
HLTF	0.291905	-3.60842
L1CAM	0.0901706	-3.60218
ARL4C	0.199081	-3.58866
DEGS1	0.00830119	-3.57363
ARL4C	0.172215	-3.56877
RUNX2	0.139991	-3.53193
CSRP2	0.124455	-3.52724
PLAU	0.107639	-3.46161
APOBEC3F /// APOBEC3G	0.0524285	-3.42063
MUM1L1	0.0549986	-3.37707
MYBL1	0.0522256	-3.35338
LBH	0.115478	-3.29062
APOBEC3G	0.114142	-3.28284
ATP10D	0.0352549	-3.27575
FSCN1	0.0610889	-3.27207
LOC730755	0.181519	-3.26088
AP1S2	0.0290847	-3.23944
NAV2	0.0714238	-3.23572
FOXD1	0.230159	-3.22194

MSX2	0.0462303	-3.21504
DCBLD2	0.199081	-3.2012
AXL	0.274998	-3.1896
IQCJ-SCHIP1 /// SCHIP1	0.054829	-3.17782
MYO5A	0.0678636	-3.17519
---	0.0858431	-3.17465
MBOAT2	0.0261552	-3.17163
WNT5A	0.208335	-3.17004
DLC1	0.0609553	-3.16947
AKAP12	0.31701	-3.16295
C13orf15	0.257814	-3.14407
MAP1B	0.21033	-3.13106
FERMT2	0.172362	-3.10512
OBSL1	0.0613203	-3.08847
MYL9	0.0522916	-3.08736
MAP1B	0.157183	-3.08243
CD109	0.407023	-3.0656
MITF	0.0683276	-3.05235
MSX2	0.163154	-3.04134
TRIB2	0.23481	-3.03935
GPX8	0.247952	-3.01516
DEGS1	0.0250091	-3.0073
SCD5	0.294627	-2.98531
---	0.230517	-2.96994
MICB	0.169427	-2.96053
WWTR1	0.111095	-2.95751
DOCK4	0.142417	-2.95732
HMGA2	0.162707	-2.93392
RDX	0.358587	-2.9214
CFL2	0.110863	-2.9174
MLLT11	0.0923292	-2.91473
EMP3	0.150923	-2.89538
CALD1	0.313316	-2.89378
PLAU	0.179002	-2.88728
ALDH1A3	0.4284	-2.8779
OLR1	0.181296	-2.84931
TIMP2	0.0853312	-2.83232
AMOTL1	0.0613203	-2.82574
DCLK1	0.105435	-2.80752
PLAT	0.225058	-2.77865
PDP1	0.0968493	-2.77324
KRT23	0.500776	-2.77237
HEG1	0.0734324	-2.7715
SMURF2	0.0275917	-2.76782
AP1S2	0.0275917	-2.76252
SLC16A6	0.19977	-2.75467
ETS1	0.311318	-2.75438
WNT5A	0.202503	-2.75239
NXN	0.437856	-2.75154
TACSTD2	0.647337	-2.74799
RDX	0.367189	-2.74187
LOC399959	0.307472	-2.74098
FZD10	0.201578	-2.73545
VGLL1	0.37033	-2.73483
KATNAL1	0.0277906	-2.7262
ZEB1	0.169705	-2.72311
PTK7	0.0399278	-2.72231
TMEM47	0.507298	-2.71861
CPVL	0.265661	-2.71613
DCBLD2	0.37437	-2.70741
FAM92A1	0.19557	-2.70111
CFL2	0.135972	-2.69762
RAB31	0.1345	-2.69233
DCBLD2	0.336046	-2.68474
AKT3	0.0905808	-2.68281
LGALS1	0.441242	-2.67832
PRKAA2	0.140767	-2.67293

HOXC10	0.0392048	-2.6515
TMEM158	0.111368	-2.64293
CACHD1	0.133569	-2.63789
PHLDB2	0.394657	-2.63019
CAV2	0.515933	-2.62236
ETS1	0.134469	-2.60964
---	0.0917428	-2.59947
LEF1	0.478941	-2.5967
RAB31	0.199081	-2.59616
RAB31	0.249159	-2.59064
ZNF512B	0.0807372	-2.59047
ABCC4	0.0772496	-2.58607
ANTXR1	0.303792	-2.58078
OGFRL1	0.234177	-2.58068
B4GALT6	0.0341456	-2.57799
SCARA3	0.0290847	-2.57466
RDX	0.600459	-2.57058
GULP1	0.399011	-2.56029
PMP22	0.230028	-2.55113
OBFC2A	0.0524285	-2.53685
CXorf57	0.0792374	-2.53673
TUSC3	0.28573	-2.53239
B3GALNT1	0.167686	-2.53004
TRPC1	0.11914	-2.51943
SYT1	0.323969	-2.51666
DPYSL3	0.130293	-2.51133
PECI	0.263078	-2.50995
STMN3	0.254294	-2.50527
PDP1	0.141038	-2.49988
TMEFF1	0.0847182	-2.48676
NANOS1	0.156045	-2.48419
MXRA7	0.1596	-2.47991
MSX1	0.244546	-2.47238
CLU	0.208335	-2.47216
SLC16A6	0.125875	-2.46125
CDC42EP3	0.170333	-2.46038
INHBB	0.169522	-2.44785
AKT3	0.14857	-2.44421
RBMS1	0.453027	-2.44241
HOXA10 /// HOXA9	0.595084	-2.44223
IL17RD	0.179678	-2.43946
DSE	0.31948	-2.42982
C12orf59	0.187483	-2.42932
SNX10	0.267109	-2.42322
NGFRAP1	0.59664	-2.41921
ADA	0.163154	-2.41557
MSRB3	0.258008	-2.41543
TRIB2	0.168881	-2.41436
IKBIP	0.160528	-2.40851
ELK3	0.168986	-2.40469
APOBEC3C	0.0350732	-2.40264
ZC3H12C	0.332135	-2.40178
SRGAP2P1	0.169427	-2.40088
PEG10	0.632815	-2.39324
CYBRD1	0.333231	-2.39322
---	0.0655816	-2.39294
WWTR1	0.106233	-2.38833
SMURF2	0.0958816	-2.38381
DNAJC6	0.252397	-2.3752
OLFML3	0.253	-2.37392
TIAM1	0.388499	-2.3687
CLDN11	0.491383	-2.36671
MCOLN3	0.331473	-2.36651
CAB39L	0.156836	-2.36355
IDS	0.0978179	-2.36325
LPHN2	0.0876525	-2.35956
CHST11	0.265847	-2.35749

DKK3	0.305712	-2.35412
CCDC88A	0.288446	-2.3536
PPFIBP1	0.102901	-2.34557
DPYSL2	0.0736152	-2.33496
CLU	0.22553	-2.33361
OGFRL1	0.505191	-2.33357
GULP1	0.313142	-2.33201
SGCE	0.548993	-2.32879
DPYSL3	0.285551	-2.32285
NAV3	0.280663	-2.31677
CD99	0.0598998	-2.31569
TMEFF1	0.0807372	-2.31538
OBFC2A	0.0590751	-2.31452
PRKAA2	0.227679	-2.31032
ABCC4	0.103727	-2.30946
TFAP2C	0.130378	-2.30885
TNFRSF19	0.172027	-2.30851
NRG1	0.380067	-2.30579
CST6	0.429574	-2.30475
COL6A1	0.248174	-2.3046
RBMS1	0.311331	-2.30362
PTPN14	0.0719421	-2.30269
HDGFRP3	0.598828	-2.2978
DFNA5	0.156836	-2.29762
PRKAR2B	0.354274	-2.29642
GPNMB	0.23305	-2.28974
---	0.0672257	-2.28551
GPX8	0.234304	-2.27688
SYT1	0.391061	-2.27485
NAV3	0.270192	-2.27342
CHST11	0.220399	-2.27072
DUSP27	0.657217	-2.26909
SLC1A3	0.214682	-2.2677
CPPED1	0.0781069	-2.26748
CDC42EP3	0.284212	-2.26486
ELOVL5	0.619496	-2.26023
GJC1	0.313804	-2.25588
FAM92A1	0.153419	-2.24365
FHOD3	0.16393	-2.24248
LOC100506377	0.392748	-2.22994
---	0.274763	-2.2281
MITF	0.0785404	-2.22641
PTPRN2	0.223192	-2.22517
B3GALNT1	0.265847	-2.2251
OGFRL1	0.326622	-2.21693
CLIP4	0.423974	-2.21665
LATS2	0.194096	-2.21335
TRIM6	0.257522	-2.21026
STXBP1	0.232208	-2.20714
ARL4C	0.21301	-2.20472
GULP1	0.495297	-2.18744
SERPINE1	0.14943	-2.18634
LPAR1	0.303118	-2.18559
KLHL5	0.558902	-2.1842
ZNF697	0.0495658	-2.18314
ADA	0.151869	-2.1831
CFL2	0.246211	-2.18272
RBMS1	0.374544	-2.17914
TNC	0.49449	-2.17717
EMB	0.548993	-2.17461
TRPC1	0.0994298	-2.1733
CD99	0.0655816	-2.17186
GJC1	0.283872	-2.17148
SORBS1	0.398294	-2.16337
RBMS1	0.466819	-2.16029
MOSPD1	0.0644311	-2.15558
GSTM3	0.636427	-2.15037

MCOLN3	0.254783	-2.14817
TIMP3	0.367057	-2.14813
CHN1	0.268342	-2.1457
HOXA1	0.230159	-2.13601
UBASH3B	0.238162	-2.13452
OXR1	0.0350732	-2.13394
TNNC1	0.446533	-2.1339
B3GALNT1	0.191503	-2.13348
RAB38	0.303792	-2.13083
SLC7A8	0.368585	-2.12932
B4GALT6	0.0532341	-2.12615
GPR161	0.141096	-2.12192
PTPRN2	0.473205	-2.12113
AP1S2	0.0971832	-2.12059
CCNE2	0.121617	-2.11996
SEC14L1	0.0520348	-2.11478
GPC1	0.14387	-2.10947
LHFP	0.223006	-2.10874
IGF2 /// INS-IGF2	0.655907	-2.10118
DCLK1	0.317191	-2.10084
RBP1	0.59664	-2.09569
PEG10	0.64494	-2.09424
PSTPIP2	0.331774	-2.09406
EPHB1	0.387002	-2.09238
TIMP3	0.283872	-2.09166
PTPN14	0.163362	-2.09113
PRKCDBP	0.31948	-2.08976
SMURF2	0.133569	-2.08517
CALD1	0.336352	-2.0839
PRR9	0.400227	-2.08308
GPR137B	0.0610889	-2.0821
PTGES	0.364585	-2.08185
ZCCHC24	0.139972	-2.07966
CD274	0.377008	-2.07896
TWSG1	0.14525	-2.07793
QKI	0.706487	-2.07713
FAM101B	0.557456	-2.07655
CTGF	0.481778	-2.07595
EDIL3	0.465227	-2.0746
CSRP2	0.168986	-2.07117
REEP1	0.457413	-2.0696
GATA3	0.325273	-2.06856
TMX4	0.223006	-2.06805
MDFIC	0.575688	-2.06605
SACS	0.441242	-2.0657
BAMBI	0.520954	-2.06305
FLNA	0.12985	-2.06198
SHROOM2	0.235921	-2.05876
SH3BP5	0.15138	-2.05766
CPPED1	0.105435	-2.05477
VIM	0.59784	-2.05077
DKK3	0.230028	-2.04998
SEPP1	0.412757	-2.04961
ALPP	0.322043	-2.04875
LOC541471 /// NCRNA00152	0.103577	-2.04641
OSMR	0.508769	-2.04457
HSD17B1	0.116414	-2.04321
WWC2	0.19123	-2.0411
ATP6V1C1	0.0212059	-2.0405
SOCS2	0.442601	-2.04034
GLS	0.0493784	-2.03927
BMP7	0.571349	-2.03875
CSGALNACT2	0.0556618	-2.03764
NINL	0.138157	-2.03591
IL18	0.447241	-2.03438
NLRP2	0.603654	-2.03434

UCA1	0.675177	-2.03397
MPP1	0.246552	-2.03367
MOSPD1	0.182677	-2.03341
BUB1	0.0242264	-2.02983
SOBP	0.263281	-2.02677
RBMS1	0.499232	-2.02543
AKAP12	0.558658	-2.0235
KIF3C	0.0350732	-2.02181
AKAP2 /// PALM2-AKAP2	0.536388	-2.02108
SNRPN /// SNURF	0.53371	-2.02074
CYTH3	0.0566606	-2.02037
MBOAT2	0.138706	-2.01932
FYN	0.38939	-2.01863
SOCS2	0.504235	-2.01838
OSMR	0.43522	-2.016
FBXO6	0.301566	-2.01569
ANKRD1	0.443978	-2.01408
TOR1AIP1	0.0181951	-2.01281
NBEA	0.146347	-2.00784
CST7	0.390234	-2.00622
BLVRA	0.215687	-2.00353
TET1	0.137162	-1.99931
GXYLT2	0.377008	-1.99913
KIAA1949	0.13936	-1.99894

Table Appendix.3 List of 300 most significantly up-regulated genes in goblet cells via RNA-seq, sorted by fold change. Fold change given as goblet cells versus non-goblet cells. P value adjusted for multiple testing using false discovery rate FDR step up.

Gene Symbol	log2(fold change)	-log2(FDR)
FCGBP	7.43347	78.07663
ATP2A3	7.00794	44.20212
MUC2	6.76090	46.15105
ANO7	6.53752	104.48742
HEPACAM2	5.96125	22.60720
ZG16	5.79090	16.57144
SPDEF	5.76295	36.29376
KLK3	5.64949	27.44391
TRPA1	5.50181	17.09180
GPR153	5.42477	42.73162
LINC00261	5.29837	48.20138
SPINK4	5.14561	77.12086
DEFA6	5.06675	3.67414
TFF3	5.03563	180.31016
NEURL	4.78038	35.95059
TMEM61	4.70302	24.21313
TFF2	4.57565	13.04021
KLK15	4.55754	13.44399
ENG	4.28848	16.85724
CBFA2T3	4.22296	29.19466
NXPE1	4.13923	12.78881
SERPINA1	4.04802	37.37245
RAB3B	3.86579	15.18224
TFF1	3.82439	66.50705
RASD1	3.68558	17.00524
RETNLB	3.61838	24.27481
FAM83E	3.59218	13.47705
KLK1	3.54014	99.65765
GSN	3.53034	67.82517
L1TD1	3.45919	4.91459
REG4	3.33262	48.02874
KLK12	3.33089	22.07417
RAB26	3.31316	49.28312
CAPN9	3.30609	14.55242
ANXA13	3.29364	7.48067
MST1	3.26939	13.92950
SELM	3.25887	26.09136
DLL1	3.17178	18.46573
RAP1GAP	3.15213	41.68618
CACNA2D2	3.12112	16.47997
CA8	3.05194	6.59083
WFDC2	3.01235	22.26327
RASD2	3.00208	2.27743
DLL4	2.96339	17.71984
FOSB	2.94150	20.27930
MLPH	2.84874	44.60914
ALDH1A1	2.80022	8.97606
NTN4	2.78711	8.06521
TP53INP2	2.68457	5.78959
KLF4	2.67622	46.39351
CREB3L1	2.65584	44.80511
TMEM184A	2.63750	13.13879
SMPD3	2.61063	15.49798
LRRC26	2.49480	17.09455
IGFBP3	2.46790	24.81849
FABP2	2.45662	8.84739
MCF2L	2.40715	24.93708
FOXA3	2.35746	8.52242

MICAL1	2.34151	17.89403
CDC42EP5	2.30899	9.57153
ERN2	2.29316	41.14269
SYT7	2.24992	6.53748
VWA2	2.24573	5.46893
SYTL2	2.24267	33.61940
MALAT1	2.22269	61.96237
CSRNP3	2.21990	7.72973
MDK	2.21363	8.55503
GNE	2.19579	15.63047
C11orf92	2.18789	14.68838
C2orf54	2.17764	4.30171
ATF3	2.17262	6.41699
SLC22A23	2.17097	15.65457
HSD11B2	2.14705	61.52446
MIR210HG	2.14446	9.38629
NDRG1	2.13693	9.74023
RNASE1	2.10171	54.92477
ABP1	2.09648	13.42612
NEAT1	2.06179	48.09460
SIDT1	2.03300	11.14040
TMEM238	1.99509	6.82371
DNAJB2	1.98594	6.53897
DUSP1	1.95496	25.59122
HMGCS2	1.95182	9.95404
FOXA2	1.92641	14.03253
SH3PXD2A	1.92589	20.68443
ABCA5	1.87900	6.21293
GALNT12	1.87408	5.39907
BAIAP2L2	1.82878	8.78354
AQP3	1.82831	13.09672
DYRK4	1.82626	40.20773
RILP	1.80168	4.32234
CBFA2T2	1.79268	13.18529
SPNS2	1.78798	6.56204
YPEL3	1.77541	5.91109
SRGAP3	1.75523	6.42165
C11orf93	1.75177	9.39068
KCNQ1OT1	1.75120	5.22363
CEACAM5	1.74586	26.07462
SCNN1A	1.74288	48.19444
MOGS	1.74164	5.18173
SEMA3B	1.73520	9.75633
PVRL1	1.72411	14.32827
MID1IP1	1.71997	9.56469
ST6GALNAC1	1.71912	33.24634
SEMA3F	1.70101	6.71046
HPCAL1	1.69971	16.98728
NR4A1	1.68597	20.44278
ZNF814	1.67969	4.02019
KLK11	1.67576	13.83214
RGL3	1.65748	5.73278
ANKRD9	1.64534	5.95812
CAPN8	1.64119	6.37454
MIA3	1.63907	26.09136
HES6	1.63177	20.68443
C3orf62	1.62908	2.28328
STARD10	1.61291	28.96548
SYTL1	1.59372	11.81301
LOC100506990	1.58026	3.09205
LOC100289137	1.56111	4.53504
HID1	1.54926	13.95360
EPS8L1	1.54150	3.46221
CYTIP	1.54139	3.20658
CAPN5	1.53829	11.93828
SMPD1	1.52577	4.24604
MCTP2	1.52422	8.15536

FABP1	1.52260	1.69721
IL4R	1.52027	8.11103
CCNJL	1.51250	4.88159
JUND	1.51120	5.37799
KCNAB1	1.51036	2.37024
AGR3	1.50706	13.11644
CBX4	1.50389	8.35127
CEACAM6	1.50363	16.70679
WNK4	1.49872	5.37982
KCNK6	1.49843	7.62576
PPM1K	1.49776	8.00546
ACAP3	1.49757	6.02399
AGR2	1.49408	35.14709
PPP1R21	1.49267	4.94329
FAM214A	1.49156	5.42908
FAM198B	1.48898	7.92813
FAM174B	1.48478	8.30051
TOX	1.48060	5.86909
ACPP	1.48035	3.79959
ANG	1.48010	6.35526
ESRP1	1.46930	6.61333
SLC16A3	1.46781	4.71784
APP	1.46594	11.42735
BTN3A1	1.46320	1.80915
ADARB1	1.45984	2.30143
TGFBI	1.42479	20.50083
ISG20	1.42154	5.60185
TRPM4	1.41422	13.84525
LOC155060	1.41342	3.00988
HSF4	1.41254	8.01913
SULT1C2	1.40346	7.43640
ZNF444	1.40056	5.00648
WIPI1	1.39867	2.23333
GUSBP11	1.39563	3.45948
CHCHD10	1.39229	19.69237
PLXNA3	1.39048	5.77622
CSAD	1.38451	3.98739
MUC5B	1.37812	18.06289
POFUT2	1.37577	2.89418
ARHGEF38	1.37537	3.52572
MIR22HG	1.37500	2.09319
ACOX2	1.36844	4.69945
SBNO2	1.36642	4.05025
GHDC	1.36616	3.18057
MTRNR2L1	1.35999	7.93643
SEC24D	1.35908	8.24949
MXRA7	1.35477	7.79191
SLC17A9	1.35427	7.25030
ABCC3	1.35051	9.39090
CREB3L4	1.34568	6.20515
CCDC64B	1.34475	2.90518
CYTH1	1.34425	6.13053
C3orf52	1.33985	6.34054
ERMAP	1.33669	2.30267
LFNG	1.33636	4.18932
ARVCF	1.33617	3.94632
EIF2AK3	1.33102	3.69351
ARFGAP1	1.32524	16.37239
PTCH1	1.29923	11.60747
EFCAB4A	1.29690	18.62831
PTOV1-AS1	1.29382	2.24093
ANO9	1.28379	9.92795
RPS6KA4	1.27884	4.46518
ZNF160	1.27845	2.43279
ZBTB7A	1.27043	3.93415
SLC44A5	1.26752	2.81081
MAGED2	1.26468	6.69317

TECPR1	1.25863	5.49410
TSPAN1	1.25644	10.25091
RGMB	1.25562	31.96874
PDXDC2P	1.25496	7.69914
KIAA0319L	1.25390	6.99858
DOPEY2	1.25262	3.65863
FASN	1.24366	15.62464
DGKD	1.23776	12.57707
C2CD2L	1.23394	2.71946
ARHGEF10	1.22939	5.73677
SIPA1L2	1.22915	4.67853
MXD1	1.22852	6.53400
TNFRSF25	1.22795	2.48650
GATA2	1.22790	3.97783
MTMR11	1.22400	4.10724
C8orf4	1.21137	2.29485
RHBDF1	1.21132	6.09816
GRAMD4	1.20631	3.52572
LOC113230	1.20574	4.66549
METRN	1.20450	2.65895
TLX1	1.20344	2.70894
S100A6	1.20337	27.64014
DNAJC10	1.19592	14.47402
PRPF19	1.19568	7.40817
MANF	1.18590	10.81965
LGALS4	1.17924	24.99499
NABP1	1.17390	1.80532
KIAA1217	1.17364	6.31492
MED15	1.17150	4.66935
BTG2	1.16393	10.85213
DENND3	1.16291	2.74280
RNF207	1.16092	3.21976
FAM46A	1.16012	5.39202
TBC1D8B	1.15708	4.81490
S100A4	1.15645	5.82960
GSDMB	1.15450	1.83922
ITM2C	1.15423	9.34350
ABAT	1.15321	1.97370
TMED9	1.15092	5.28361
CAPRIN2	1.14687	5.01458
ALDH3B1	1.14684	1.32214
MICALL2	1.14580	5.97349
PITX1	1.14437	9.43385
CLCN6	1.14311	2.85443
TEP1	1.14025	4.29723
C12orf23	1.13860	5.88002
SHD	1.13517	2.44854
GOLGA8A	1.13392	15.54429
EDEM3	1.13199	8.50066
MARVELD1	1.13147	1.64373
MVP	1.13097	15.02296
PLEKHG3	1.12908	3.75013
NRBP2	1.12396	3.61696
ZNF839	1.12252	2.94328
IFI27L2	1.12025	3.31113
IER5	1.11935	2.52772
BAIAP3	1.11699	1.58636
ORMDL3	1.11610	4.26625
ELF3	1.11374	15.26233
SPIRE2	1.11280	2.72966
PPP1R9B	1.10644	2.48783
LRFN4	1.10462	3.07245
TMC4	1.10379	5.74521
PPDPF	1.09252	2.73594
NPDC1	1.09053	20.67058
LINC00324	1.08891	1.70198
GLTSCR2	1.08879	7.79000

TPM1	1.08153	16.20981
MLL4	1.08084	4.54908
DGAT1	1.07935	12.33606
FAM73B	1.07627	2.23795
SLC39A8	1.07492	5.17186
OSBPL7	1.07156	2.93865
RAPGEF5	1.07048	2.18466
LIMD1	1.06720	1.67882
LOC100507410	1.06364	1.31543
ENTHD2	1.05722	2.48888
C7orf43	1.05598	1.43175
SLC12A9	1.05567	2.29937
SMARCAD1	1.05560	8.92143
CKAP4	1.05135	10.71326
ARHGEF2	1.04932	2.10323
TMSB4X	1.04789	3.76130
CHPF	1.04166	8.89182
MON2	1.04068	3.69275
IRS2	1.03441	12.78881
OGFR	1.03339	2.19607
LPCAT1	1.03066	5.69680
CCNL2	1.03004	15.30474
CCPG1	1.02920	3.56299
CMIP	1.02830	2.47964
ZNF823	1.02643	1.43812
CCDC125	1.02479	2.34696
KIAA1324	1.02464	8.77725
CKMT1B	1.02448	3.49545
METTL21B	1.02432	1.66387
TMEM259	1.02372	5.77623
SPHK2	1.02139	4.11009
FAM114A1	1.02093	4.23095
IFT172	1.01797	1.88138
FGFR3	1.01622	1.90809
SEC31A	1.01220	6.29041
ZNF320	1.01188	2.82981
TRIB1	1.01161	6.85029
ZNF503	1.01004	2.05601
KIAA1244	1.01002	6.37200
LSM14B	1.00390	2.17757
LINC00659	1.00161	1.87399

Table Appendix.3 List of 50 most significantly up-regulated genes in non-goblet cells via RNA-seq, sorted by fold change. Fold change given as goblet cells versus non-goblet cells. P value adjusted for multiple testing using false discovery rate FDR step up.

Gene Symbol	log2(fold change)	-log2(FDR)
DKK1	-2.80585	2.33277
CXCL11	-2.49815	12.33606
EMP1	-2.30889	3.25917
APCDD1	-2.03189	2.99001
WWTR1	-1.84387	4.76306
PTGDR	-1.70106	6.62513
FAM81A	-1.66135	3.09056
CALB2	-1.62601	6.87013
IGFL2	-1.61646	7.59889
HNRNPA1L2	-1.52819	4.68550
FAM111B	-1.51981	2.72771
MOSPD2	-1.48582	4.59811
SLC43A3	-1.43945	3.55669
GREM1	-1.40329	1.44668
HIF1A	-1.38214	3.85922
UBASH3B	-1.36035	1.51271
PCGF6	-1.33210	1.51975
LOC100190940	-1.31168	1.33851
VCAN	-1.30674	14.23613
MCTP1	-1.29360	2.06783
KLF11	-1.28969	2.31574
E2F8	-1.28818	1.65610
FEN1	-1.27437	4.37489
HHLA2	-1.24653	2.30948
FAM175B	-1.24364	1.94599
PARP16	-1.24349	1.93086
NXT2	-1.24247	2.44854
ACO2	-1.23605	3.53975
ITGB8	-1.22521	1.82062
LRP4	-1.19008	2.34833
SMOC2	-1.17711	1.84335
RAD51	-1.15432	2.95550
PRIM1	-1.13950	3.69351
TRIM21	-1.13935	2.04941
SLC35G1	-1.12604	1.58117
GULP1	-1.12140	2.12879
MAN2A1	-1.11493	3.07156
C18orf8	-1.10814	1.55648
PKMYT1	-1.10709	2.48248
CHRNB1	-1.07912	1.64264
UGT1A1	-1.07626	1.76349
RAB38	-1.07479	1.44627
HRH1	-1.06667	1.74181
BPGM	-1.06209	2.08341
SCLT1	-1.05734	1.59716
PKIA	-1.05524	1.43275
ASB4	-1.05005	5.80547
SFMBT1	-1.04871	1.67376
PAIP2B	-1.04636	1.48726
M6PR	-1.04561	1.93132

Table Appendix.5 RPKM values of 50 most significantly up-regulated genes in goblet cells. Samples 1-7 are goblet cells. Samples 8-14 are non-goblet cells.

Sample No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
FCGBP	119.99	142.63	109.22	102.40	69.84	71.85	93.99	1.15	0.95	0.19	0.17	0.79	0.26	0.58
ATP2A3	4.99	4.43	5.71	6.81	5.49	3.28	6.44	0.13	0.04	0.00	0.00	0.00	0.00	0.00
MUC2	161.41	198.60	201.03	191.55	156.55	168.00	167.09	4.82	1.40	0.60	0.32	2.68	0.90	0.67
ANO7	36.15	33.47	25.21	49.59	30.24	34.39	38.42	0.66	0.44	0.42	0.07	0.11	0.04	0.70
HEPACAM2	34.74	62.31	32.44	53.35	47.33	30.59	48.78	1.93	1.94	0.16	0.00	0.23	0.00	0.44
SPDEF	91.74	141.92	116.87	103.14	84.47	76.40	97.91	3.08	3.80	0.67	0.00	1.59	0.53	3.34
ZG16	108.71	196.31	92.79	68.88	94.48	121.33	100.48	1.66	1.37	0.58	0.00	7.02	0.00	3.76
KLK3	80.01	51.11	134.84	72.69	46.93	32.71	52.69	3.96	1.12	0.32	1.31	0.00	0.92	1.31
TRPA1	6.68	3.74	6.82	7.55	5.40	8.43	12.17	0.17	0.00	0.00	0.00	0.24	0.03	0.71
GPR153	11.60	7.53	11.01	13.16	7.33	7.61	14.29	0.20	0.35	0.26	0.12	0.43	0.00	0.38
LINC00261	42.58	19.12	30.85	18.42	24.84	11.52	29.79	1.35	0.40	0.21	0.33	0.57	0.86	0.63
SPINK4	4853.01	5632.90	4416.66	3913.84	2893.25	2851.44	3233.13	172.61	164.30	46.93	47.17	121.98	116.08	115.41
DEFA6	83.70	99.06	91.19	639.31	808.42	79.91	251.95	50.36	9.85	0.00	0.00	0.00	0.00	0.00
TFF3	5661.59	5953.54	5698.80	4612.99	4623.03	4083.10	5142.07	221.91	213.31	99.09	130.39	170.35	117.08	138.00
NEURL	26.92	39.99	27.50	19.25	23.71	22.20	24.39	2.25	1.36	0.22	0.52	0.82	0.27	1.08
TMEM61	42.64	44.45	33.29	40.44	36.17	19.66	26.12	0.78	2.10	1.49	0.29	0.00	1.08	3.54
TFF2	280.88	287.54	211.19	261.79	288.25	119.79	216.28	3.91	20.57	1.32	2.96	36.59	4.08	1.37
KLK15	23.23	28.46	44.02	37.84	14.23	22.37	34.68	0.98	2.17	0.00	0.00	1.94	0.26	3.56
CBFA2T3	8.00	6.45	5.74	5.52	4.17	8.26	6.63	0.47	0.58	0.28	0.32	0.11	0.22	0.30
ENG	10.37	8.54	9.36	3.93	3.59	4.21	3.09	0.29	0.00	0.18	0.35	0.56	0.27	0.61
NXPE1	15.17	18.73	16.54	31.85	6.64	8.17	13.57	0.88	2.44	0.00	1.36	0.94	0.45	0.21
SERPINA1	31.98	41.91	21.35	33.42	29.93	25.92	29.30	3.53	2.12	1.20	0.44	1.24	1.29	2.89
RAB3B	4.54	4.33	4.97	3.61	3.98	1.31	3.20	0.54	0.18	0.16	0.05	0.12	0.08	0.60
TFF1	2300.96	1660.44	821.60	1752.25	1673.80	1220.23	1528.69	94.56	130.82	87.04	120.22	184.02	64.77	94.21
RASD1	19.09	20.36	36.26	17.79	9.57	23.91	24.84	1.27	2.74	0.87	0.28	1.31	3.30	2.11
RETNLB	235.02	130.05	137.06	175.49	96.28	127.77	119.04	22.92	18.88	4.89	4.49	5.31	9.76	15.57
FAM83E	8.24	8.88	11.06	7.35	8.31	9.54	8.45	1.07	0.18	0.06	0.40	0.63	1.94	0.77
GSN	90.98	99.04	60.09	90.22	75.90	72.15	86.71	8.72	4.63	3.99	7.43	4.81	10.07	9.97
KLK1	230.56	329.45	281.92	225.03	207.68	224.14	260.54	18.47	24.95	17.89	20.60	22.05	31.23	16.64
L1TD1	14.95	8.74	6.50	3.03	5.10	3.12	5.40	0.25	1.42	0.06	0.04	0.89	0.00	1.65
KLK12	168.21	110.95	240.96	96.61	87.66	83.28	117.35	24.97	16.67	3.91	5.99	11.81	12.94	13.18
REG4	526.75	666.07	490.01	632.27	505.01	444.04	509.29	85.28	59.65	46.53	33.97	78.70	20.53	49.73
ANXA13	38.54	31.58	18.12	14.39	16.39	9.69	31.04	4.70	4.56	1.98	0.00	3.34	0.43	1.10
CAPN9	12.85	9.14	11.75	8.84	6.64	11.90	16.43	1.04	2.26	0.41	0.62	0.39	1.24	1.87
RAB26	96.96	50.85	69.34	78.16	75.98	69.82	95.56	7.90	9.94	4.89	4.95	12.34	6.58	7.88
MST1	16.22	20.48	26.06	39.33	12.17	7.13	28.50	2.00	4.26	3.55	2.21	1.63	0.58	1.16
SELM	90.53	86.97	80.38	65.63	45.08	67.87	50.28	7.55	7.70	5.74	7.21	8.07	2.70	12.07
DLL1	21.37	20.59	22.07	17.13	12.15	13.51	26.66	3.82	3.59	1.37	1.60	0.61	0.82	2.75
RAP1GAP	56.52	65.75	79.33	70.73	62.58	32.69	69.53	9.15	8.28	5.26	9.83	2.82	7.92	5.61
CACNA2D2	4.72	5.89	3.35	7.17	3.15	2.29	6.76	0.68	0.46	0.54	0.63	0.47	0.54	0.43
CA8	6.50	11.61	5.82	12.44	7.18	10.60	16.14	3.53	1.25	1.43	0.00	0.78	0.22	0.94
RASD2	13.44	9.13	6.59	8.29	5.56	3.40	15.36	0.00	5.01	0.67	0.00	0.00	0.00	2.02
WFDC2	74.72	119.22	74.23	111.04	52.58	92.30	133.51	11.06	9.40	6.39	6.53	17.47	15.27	16.10
DLL4	42.00	31.63	45.76	18.02	18.12	16.89	31.59	4.18	2.33	3.34	3.76	8.08	1.19	3.40
FOSB	9.55	7.96	14.91	19.71	19.03	14.57	22.30	1.19	2.49	3.37	2.64	1.61	1.21	1.49
MLPH	41.56	62.26	41.97	47.05	43.60	34.77	43.28	3.92	9.21	4.72	9.29	4.59	6.36	5.68
ALDH1A1	37.84	26.02	32.47	36.62	17.78	21.47	50.42	9.82	5.61	1.08	3.60	1.61	1.07	8.92
NTN4	5.01	16.62	7.72	6.07	7.94	8.16	10.73	0.46	2.60	0.55	2.45	1.96	0.56	0.54
TP53INP2	6.80	12.24	8.99	4.90	5.53	6.57	11.25	0.54	0.69	0.48	4.09	0.74	0.16	2.07
KLF4	83.45	76.04	92.98	63.44	49.54	54.51	91.04	9.92	14.27	8.43	14.80	7.61	12.89	12.03

Table Appendix.6 RPKM values of 50 most significantly up-regulated genes in non-goblet cells. Samples 1-7 are goblet cells. Samples 8-14 are non-goblet cells.

Sample No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
DKK1	0.00	14.17	0.00	10.24	7.23	8.88	0.00	11.05	80.16	52.62	55.30	27.40	9.95	47.17
CXCL11	2.80	9.12	2.07	7.70	3.26	3.09	6.34	35.46	23.48	38.34	25.00	19.01	25.62	25.32
EMP1	0.21	2.20	0.51	2.55	4.16	4.66	0.72	3.66	5.50	8.12	28.09	4.25	6.65	18.19
APCDD1	2.89	0.00	1.40	4.52	1.43	1.47	1.78	3.95	8.79	8.52	18.57	5.94	4.57	4.54
WWTR1	0.25	2.34	2.06	0.78	2.01	2.43	0.57	4.97	5.62	4.04	5.12	6.24	5.94	6.00
PTGDR	6.23	11.26	18.36	12.98	14.26	7.26	1.98	40.26	40.20	26.97	26.99	32.58	40.85	28.41
FAM81A	0.22	2.33	2.70	3.84	4.06	4.05	0.66	9.45	9.71	6.12	7.51	11.26	7.62	5.03
CALB2	8.75	2.92	9.98	6.37	13.02	15.21	11.21	38.45	29.55	26.51	24.49	26.91	32.63	29.68
IGFL2	44.89	24.45	32.40	40.78	61.05	30.87	37.70	90.23	77.61	144.19	76.24	59.22	165.25	221.55
HNRNPA1L2	1.07	1.67	2.32	2.08	2.13	1.09	1.50	5.66	4.57	5.62	4.28	4.61	4.80	5.00
FAM111B	6.18	0.43	1.43	1.38	4.15	3.22	1.50	8.32	8.76	6.10	7.04	9.83	5.66	6.86
MOSPD2	2.25	1.32	3.81	2.92	2.27	2.87	2.40	4.89	11.03	4.70	6.67	7.51	7.03	8.56
SLC43A3	2.49	2.13	2.45	3.58	6.60	2.10	3.86	7.55	7.23	14.68	7.40	9.53	4.75	11.75
GREM1	1.51	0.00	1.99	1.25	2.06	4.93	0.61	6.07	4.07	3.64	8.04	3.92	4.10	2.91
HIF1A	1.29	5.94	1.53	3.08	3.76	4.01	1.96	8.11	10.25	6.77	6.25	10.85	7.14	6.93
UBASH3B	0.28	0.47	0.05	2.18	1.06	0.95	0.59	2.48	2.70	2.28	1.91	1.10	2.14	1.55
PCGF6	2.36	3.46	0.04	3.98	4.84	1.24	4.78	4.76	9.70	5.02	7.28	8.87	9.20	7.04
RPS6KA5	0.32	0.12	6.10	4.46	0.97	1.84	2.53	3.81	5.88	7.85	5.52	5.12	5.35	7.19
LOC100190940	4.25	0.00	0.94	2.76	5.31	5.48	4.11	11.26	6.85	8.72	6.96	4.03	9.13	9.60
VCAN	5.74	10.05	5.40	7.54	6.99	5.00	9.09	18.04	17.47	16.94	24.05	15.50	15.02	16.07
E2F8	3.88	0.04	6.15	1.86	3.02	5.18	5.43	12.03	7.32	7.17	8.58	8.94	6.49	12.02
MCTP1	0.90	0.36	0.50	0.91	2.57	2.38	1.35	3.06	3.56	3.32	2.12	3.34	3.21	3.32
KLF11	2.63	1.24	2.07	6.08	1.58	3.73	0.83	6.09	6.93	5.27	9.54	8.09	3.99	4.38
FEN1	10.96	25.45	4.50	10.74	20.63	14.32	10.39	28.89	39.42	29.55	27.89	49.75	29.04	30.27
NXT2	7.75	2.49	1.95	3.16	3.87	3.67	1.45	9.97	11.96	5.78	9.86	7.18	5.03	7.78
ITGB8	1.04	0.61	1.50	0.98	0.89	0.16	0.38	1.24	2.91	1.01	2.25	1.24	2.32	2.17
PARP16	4.00	1.42	0.60	2.14	6.18	2.81	2.30	7.85	7.98	4.88	5.20	7.61	7.09	5.36
ACO2	5.05	2.82	8.73	10.07	7.17	9.96	2.78	17.89	14.23	15.67	10.03	21.65	16.88	13.66
FAM175B	0.98	6.43	1.11	6.26	2.85	4.17	5.51	8.70	8.12	4.68	12.52	12.80	6.57	11.02
HHLA2	2.27	2.19	1.42	2.41	2.09	2.98	5.82	5.93	7.07	10.00	2.78	8.01	5.61	5.69
PGM2L1	0.57	0.46	3.72	0.70	0.66	1.07	0.21	1.86	1.81	2.57	1.42	2.24	2.71	4.24
LRP4	1.37	0.46	1.73	1.95	1.84	3.87	3.70	6.93	2.76	5.79	3.34	4.16	6.26	4.64
SMOC2	4.45	0.55	2.55	4.30	3.87	2.07	3.54	5.40	8.33	3.78	7.96	7.73	10.87	4.12
RAD51	3.87	5.96	5.71	2.14	9.64	5.07	8.07	17.96	14.65	13.74	13.42	9.22	9.81	11.22
PRIM1	32.92	11.98	13.92	11.13	17.26	11.50	25.50	48.23	27.17	54.20	25.62	43.19	42.09	32.50
TRIM21	3.43	4.36	2.50	12.01	8.03	3.80	4.73	11.45	19.40	8.35	12.92	13.54	10.28	9.25
GULP1	1.89	2.73	7.51	1.69	5.73	3.28	9.55	7.90	7.69	14.82	10.66	13.34	6.43	9.79
SLC35G1	3.00	5.22	1.46	8.41	4.71	1.65	2.22	8.02	7.97	3.88	5.62	11.29	8.61	12.67
MAN2A1	2.20	1.44	3.16	3.75	2.14	1.93	1.68	4.38	4.56	6.76	3.98	4.80	5.32	5.56
PKMYT1	2.44	5.61	6.57	7.39	3.97	2.66	3.75	11.02	8.26	11.97	8.21	7.84	10.40	12.25
C18orf8	6.13	0.77	3.40	12.51	11.34	4.53	3.19	15.16	12.51	11.90	13.58	13.56	12.41	10.89
UGT1A1	7.44	1.46	11.44	6.27	7.01	4.28	2.26	16.53	10.61	13.42	8.25	16.68	10.21	9.54
ASB9	8.81	0.25	4.69	6.59	5.16	9.04	0.49	12.26	7.06	10.94	12.65	13.31	9.97	7.82
RAB38	0.27	13.17	8.81	8.95	8.70	9.35	6.95	17.99	20.89	16.15	14.53	19.45	14.51	15.14
RAC2	1.06	8.89	3.67	1.34	2.02	1.75	8.46	6.15	8.70	6.28	7.01	7.60	10.51	10.96
RAD54L	0.25	5.08	1.62	4.78	2.99	1.20	1.93	6.28	6.65	3.86	5.21	5.56	6.25	3.70
CHRNB1	4.16	1.06	4.62	11.12	3.36	5.93	10.41	16.12	12.66	9.24	7.65	10.82	16.97	11.86
HRH1	0.81	1.89	0.88	1.23	1.08	3.37	1.42	3.72	2.47	2.35	3.80	4.00	2.15	3.86
CLDN2	1.70	0.90	7.18	2.44	7.64	9.27	19.32	14.35	25.43	12.73	14.77	11.29	15.53	7.05

Code Appendix.1: RNA-seq differential expression analysis in R package.

```
library(edgeR)

## Import the counts table derived from featureCounts
unfiltered_data_haoyu<-read.table("/t1-
data/user/hliu/friscr_liu/all_fastq_files/bamtobigwig/friscr_muc2_features_counted.txt", header = T,
stringsAsFactors=F, sep="\t")

## Rename the column names in both tables in a user-friendly way
colnames(unfiltered_data_haoyu)[7:20] <-
c("unfiltered_1","unfiltered_2","unfiltered_3","unfiltered_4","unfiltered_5","unfiltered_6","unfiltered_7","unfilt
ered_8","unfiltered_9","unfiltered_10","unfiltered_11","unfiltered_12","unfiltered_13","unfiltered_14")
colnames(unfiltered_data_haoyu)

## Rename the row names in both tables in a user-friendly way
rownames(unfiltered_data_haoyu)<-unfiltered_data_haoyu$Geneid
names(unfiltered_data_haoyu)

## Create a factor named factor_goblet_cell with "GC", "GC", "GC", "GC", "GC", "GC", "GC", "NGC", "NGC",
"NGC", "NGC", "NGC", "NGC", "NGC" (14 elements to clarify samples in the tables)
factor_goblet_cell <- factor(rep(c("GC", "NGC"), each=7))
factor_goblet_cell

## Create the DGEList object
unfiltered_counts_data <- DGEList(counts=unfiltered_data_haoyu[, 7:20],
group=factor_goblet_cell,genes=unfiltered_data_haoyu[,1:6])

## Filter for lowly expressed genes
unfiltered_cpm<-cpm(unfiltered_counts_data)
unfiltered_cpm_fil<-rowSums(unfiltered_cpm>1)>=7

## Use the vector as an index to filter the object unfiltered_counts_data. Save the filtered results in the same
objects.
unfiltered_counts_data<-unfiltered_counts_data[unfiltered_cpm_fil,]

## Normalisation
unfiltered_counts_data_norm <- calcNormFactors(unfiltered_counts_data)

## Output tables with normalised values in txt files
unfilter_norm <- cpm(unfiltered_counts_data_norm, normalized.lib.sizes=FALSE)
write.table(unfilter_norm, file="normalised_table_unfilter.txt", quote=T, sep="\t", col.names=T, row.names = T)

## Estimation of the Biological Coefficient of Variation (BCV) using the function estimateDisp
set_d_unfilter<-estimateDisp(unfiltered_counts_data_norm)
set_d_unfilter$common.dispersion

## Extract from DGEList
et_unfilter <- exactTest(set_d_unfilter)

## Two ways to test for differentially expresed genes
## Function decideTestsDGE returns a matrix of -1, 0, 1, denoting down-regulated, non-differentially expressed
and up-regulated genes respectively.

deg.table_unfilter <- decideTestsDGE(et_unfilter, adjust.method = "BH", p.value = 0.05)
summary(deg.table_unfilter)

## Function topTags returns the top differentially expressed genes (sorted by adjusted p-value by default).
top.table_unfilter <- topTags(et_unfilter, n = nrow(unfiltered_counts_data_norm))
top.table.sign_unfilter <- top.table_unfilter$table[,c("logFC","logCPM","PValue","FDR")]
top.table.sign_unfilter <- top.table.sign_unfilter[top.table.sign_unfilter$FDR<0.05,]

## Create tables with rpkm values
```

```
rpkm_obj_unfilter <- rpkm(unfiltered_counts_data_norm,
gene.length=unfiltered_counts_data_norm$genes$Length)
head(rpkm_obj_unfilter)

## Output the tables with RPKM values in the txt files
write.table(rpkm_obj_unfilter, file="rpkm_table_unfilter.txt", quote=F, sep="\t", col.names=T, row.names = T)
write.table(top.table.sign_unfilter, file = "DE_genes_unfilter.txt", quote = FALSE, sep = "\t",col.names=T,
row.names = T)

## Plot the values of logFC versus logCPM
plot(top.table_unfilter[[1]]$logCPM, top.table_unfilter[[1]]$logFC, main = "Smear plot (unfilter)",xlab =
"logCPM", ylab = "logFC", pch=".")
abline(h=c(-1,1))

## Make histogram of the logFC
hist(top.table.sign_unfilter$logCPM[top.table.sign_unfilter$logFC>0], main="histogram of logCPM    values for
the\n significantly upregulated genes (unfilter)", xlab="logCPM", col="grey")
hist(top.table.sign_unfilter$logCPM[top.table.sign_unfilter$logFC<0], main="histogram of logCPM    values for
the\n significantly downregulated genes (unfilter)", xlab="logCPM", col="grey")

## Make boxplots of the logFC
boxplot_MUC2_unfilter <- data.frame(rpkm=rpkm_obj_unfilter["MUC2",], group=rep(c("GC", "NGC"), each=7))
boxplot(boxplot_MUC2_unfilter$rpkm ~ boxplot_MUC2_unfilter$group, main="boxplot of MUC2
expression_unfilter", ylab="rpkm")

nrow(sixpos_vs_sevenneg_counts_data_norm)

## Multi-dimensional scaling to examine the clustering of samples
plotMDS(unfiltered_counts_data_norm, top=nrow(unfiltered_counts_data_norm), main="PCA_unfilter",
col=rep(c("blue","red"),    each=7), labels=c("Goblet Cell 1","Goblet Cell 2", "Goblet Cell 3", "Goblet Cell 4",
"Goblet Cell 5", "Goblet Cell 6", "Goblet Cell 7", "Non-Goblet Cell 1", "Non-Goblet Cell 2", "Non-Goblet Cell 3",
"Non-Goblet Cell 4", "Non-Goblet Cell 5", "Non-Goblet Cell 6", "Non-Goblet Cell 7"))

## Plot the genewise biological coefficient of variation (BCV) against gene abundance (in log2 counts per million)
plotBCV(set_d_unfilter)

## Plot smears only with the signicant values highlighted in red
plotSmear(et_unfilter, de.tags=row.names(top.table.sign_unfilter))
abline(h=c(-1,1), col="blue")

## Make heatmaps
library(gplots)
logCPM_unfilter <- cpm(unfiltered_counts_data_norm$counts, log = TRUE)
temp_heatmap_unfilter <- logCPM_unfilter[rownames(top.table.sign_unfilter), ]
heatmap.2(temp_heatmap_unfilter, scale="row", trace="none", labRow="",main="heatmap_unfilter")

## Analysis done by Haoyu
## 26/Sep/2016
## Special thankfulness is given to Neil, Nicolaos, Nicki, Emma and Chi
```