

Single cell RNA-sequencing in the upper gastrointestinal tract



Richard Peter Owen

Exeter College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Clinical Medicine

Trinity 2018

Abstract

Single cell RNA-seq in the upper gastrointestinal tract

Richard Peter Owen, Exeter College

*A thesis submitted for the degree of Doctor of Philosophy in Clinical
Medicine, Trinity 2018*

Barrett's oesophagus is a precursor of oesophageal adenocarcinoma. In this common condition, squamous epithelium in the oesophagus is replaced by columnar epithelium in response to chronic acid reflux. Barrett's oesophagus is highly heterogeneous and its relationships to normal tissues are unclear. In this thesis, I investigate the cellular complexity of Barrett's oesophagus and the upper gastrointestinal tract using RNA-seq of single cells from multiple biopsies from six patients with Barrett's oesophagus and two patients without oesophageal pathology. I show that cell populations in Barrett's oesophagus, characterised by *LEFTY1* expression, exhibit a profound transcriptional overlap with oesophageal submucosal gland cells, but not with gastric or duodenal cells. Additionally, *SPINK4* appears to mark cells that precede morphologically identifiable goblet cells in Barrett's oesophagus, potentially aiding the identification of metaplasia. Then, using a machine learning-based approach to improve the accuracy of single nucleotide variant identification in single cell RNA-seq data ('snv.predict'), I show the mutational profiles and genetic relationships of cells in Barrett's oesophagus, demonstrating clonal phenotype heterogeneity and a common mutational link between Barrett's oesophagus and oesophageal submucosal glands. Finally, with the help of modifications designed to improve the practicality of plate based single cell RNA-seq in clinical science, I identify an adaptive-to-innate immune switch in tumour cells from a single patient with oesophageal adenocarcinoma treated with immunotherapy (anti PD-L1), as part of a clinical trial.

Acknowledgements

First I would like to thank all patients who participated in this work. I consented nearly 100 people overall and I hope they will be happy with their contribution to science. Next, I would like to thank my supervisors, Professors Xin Lu and Mark Middleton, for helping and supporting me to carry out an extremely ambitious project. I would also like to thank the many people I have met and who have helped me through this whole process including numerous post docs, technical and support staff, and other students alike. In particular, Michael White and David Severson have been excellent colleagues and I could not have completed my DPhil project without their advice, support and friendship. I am also very grateful for the support I received from Andrew Roth, who helped to supervise my computational analyses and sanity checked my methodological approaches. At the bench-side, both Ludo Buti and Stefanie Werner were a great help in showing me how to perform experiments, and how to analyse the results. With sequencing, I received an enormous amount of support from the sequencing team at the Wellcome Trust Centre for Human Genetics, especially Angie Green, Paolo Piazza and David Buck. Paul Sopp (Weatherall Institute of Molecular Medicine) was also very kind in spending time sharing some of his FACS expertise, which gave me the confidence to embark on a single cell project with clinical samples. My ever-growing family have been a great source of love and support throughout this DPhil and they have kept me grounded. Finally, I am very thankful for the funding I have received from the Ludwig Institute for Cancer Research, the NIHR Biomedical Research Centre, the Oxford Health Services Research Committee and Oxford University Clinical Academic Graduate School, without which I could not have completed this project.

Publications

- Owen RP, White MJ, Severson DT, Braden B, Bailey A, Goldin R, Wang LM, Ruiz-Puig C, Maynard ND, Green A, Piazza P, Buck D, Middleton MR, Ponting CP, Schuster-Böckler B, Lu X. Single cell RNA-seq reveals profound transcriptional similarity between Barrett's oesophagus and oesophageal submucosal glands. *Nat Commun.* 2018 Oct 15;9(1):4261.
- Severson DT, Owen RP, White MJ, Lu X, Schuster-Böckler B. Bearscc determines robustness of single-cell clusters using simulated technical replicates. *Nat Commun.* 2018 Mar 22;9(1):1187.
- Zhou FY, Ruiz-Puig C, Owen RP, White MJ, Rittscher J, Lu X. Motion Sensing Superpixels (MOSES): A systematic framework to quantify and discover cellular motion phenotypes. *Elife.* 2019 Feb 26;8.

Contents

1	Introduction	1
1.1	The upper gastrointestinal tract	1
1.1.1	Oesophagus	1
1.1.2	Stomach	2
1.1.3	Duodenum	3
1.1.4	Upper gastrointestinal cells	3
1.2	Upper gastrointestinal epithelia and development	4
1.2.1	Upper gastrointestinal development	4
1.2.2	Oesophageal epithelium	5
1.2.3	Gastric epithelium	7
1.2.4	Duodenal epithelium	12
1.3	Barrett's oesophagus, oesophagitis and gastro-oesophageal reflux disease	16
1.3.1	Gastro-oesophageal reflux disease and oesophagitis	16
1.3.2	Barrett's oesophagus	16
1.3.3	Origins of Barrett's oesophagus	18
1.3.3.1	Gastric migration	18
1.3.3.2	Squamous transdifferentiation	18
1.3.3.3	Squamous transcommitment	19
1.3.3.4	Junctional progenitor cell recruitment	19
1.3.3.5	Oesophageal submucosal gland origin	20
1.3.4	Gastrointestinal stem cells and Barrett's oesophagus	22
1.3.4.1	Intestinal stem cells	22
1.3.4.2	Gastric stem cells	23
1.3.4.3	Stem cells to identify an origin for Barrett's oesophagus	24
1.3.5	Molecular signalling in Barrett's	25
1.3.5.1	Notch	25
1.3.5.2	Bone morphogenetic protein	25

1.3.5.3	Wingless	26
1.3.5.4	Krüppel like factors	26
1.3.5.5	Hedgehog	26
1.3.5.6	Mechanisms and signalling pathways in oesophageal adenocarcinoma	27
1.3.6	Modelling gastro-oesophageal reflux and Barrett's oesophagus	28
1.3.6.1	Animal models	28
1.3.6.2	<i>In vitro</i> and <i>ex vivo</i> models	30
1.4	Genetic features of Barrett's oesophagus	32
1.4.1	Genomic landscape features of Barrett's oesophagus and oe- sophageal adenocarcinoma	32
1.4.2	Gene expression analyses in Barrett's oesophagus	34
1.5	Single cell sequencing	36
1.5.1	Single cell RNA-seq	36
1.5.2	Characterisation of the human upper gastrointestinal tract with single cell RNA-seq	39
1.5.3	Other single cell sequencing methods	41
1.5.3.1	Single cell genomics	41
1.5.3.2	Single cell epigenomics	42
1.5.3.3	Combined methods sequencing in single cells	43
1.5.4	Identifying single nucleotide variants in single cell RNA-seq . .	46
1.6	Immunotherapy and treatment of oesophageal cancer	47
1.6.1	Immunotherapy	49
1.6.1.1	Cancer vaccines	49
1.6.1.2	Adoptive T-cell transfer	49
1.6.1.3	Immune checkpoint inhibition	50
1.6.2	LUD2015-005 - a phase I/II study of anti PD-L1 in combination with chemo(radio)therapy for oesophageal cancer	52
2	Materials and Methods	55
2.1	Human tissue sampling	55
2.2	Cell isolation	57
2.3	TMB-HRP reporter assay	60
2.4	Single cell RNA-seq	60
2.5	Bulk RNA-seq	60
2.5.1	RNA-seq pre-processing and quality control filtering	61

2.5.2	Analysis of processed RNA-seq data	61
2.5.3	Whole genome sequencing	62
2.5.4	Single nucleotide variant calling	63
2.6	Tissue staining	63
2.7	Cell preservatives for single cell RNA-seq	64
2.8	Cell culture and staining	65
2.9	Immunoblotting	65
2.10	Cell co-culture	66
2.11	Organoid culture	66
2.12	Data presentation	67
2.13	Buffers	67
3	Barrett’s oesophagus and upper gastrointestinal cell populations	69
3.1	Single cell RNA-seq identifies subpopulations in normal upper gastroin- testinal epithelia	70
3.2	Barrett’s oesophagus is enriched for LEFTY1-expressing cells	79
3.3	Oesophageal submucosal glands share an RNA composition profile with Barrett’s oesophagus	81
3.4	ITLN1 and SPINK4 mark early goblet cells	91
3.5	<i>OLFM4</i> marks a stem-like transcript profile in Barrett’s oesophagus and oesophageal submucosal gland cells	94
3.6	Summary	102
4	Calling mutations in single cell RNA-seq data	103
4.1	Single nucleotide variants in single cell RNA-seq	104
4.1.1	Identifying ground truths in single cell RNA-seq variant calling	104
4.2	Machine learning augments SNV calling in single cell RNA-seq	108
4.3	Single nucleotide variants in Barrett’s oesophagus	114
4.3.1	Mutational landscape of phenotyped single cells in Barrett’s oesophagus	115
4.3.2	Phenotype heterogeneity exists within the clonal structure of Barrett’s oesopahgus	116
4.3.3	Unique mutations in goblet cells	117
4.3.4	Unique mutations in Barrett’s oesophagus	118
4.3.5	Mutational signatures in Barrett’s oesophagus	121
4.3.6	Genomic links between Barrett’s oesophagus and oesophageal submucosal glands	125

4.4	Generalising SNV prediction models	126
4.5	Single cell mutations in breast carcinoma	128
4.6	Summary	131
5	Enhancements for the clinical application of single cell RNA-seq	132
5.1	Tissue preservatives and single cell RNA sequencing	133
5.1.1	RNA stabilisation	133
5.1.1.1	Cryopreservation	133
5.1.1.2	RNA <i>later</i>	133
5.1.1.3	Dithio-bis(succinimidyl propionate)	135
5.1.2	Data quality with RNA preservatives	136
5.2	Accelerating throughput with robotic sample preparation	139
5.3	Miniaturising sequencing reactions	143
5.3.1	Reagent reduction	143
5.3.2	Sequencing depth requirements	146
5.3.3	Cell selection	149
5.4	Single cell RNA-seq in a phase I/II trial of immunotherapy for oe- sophageal cancer	151
5.4.1	Single cell RNA-seq identifies a response signature in oesophageal cancer treated with immunotherapy	154
5.5	Summary	160
6	Tissue homeostasis at the upper gastrointestinal squamocolumnar junction	161
6.1	Squamocolumnar junction in humans and mice	162
6.2	Effects of gastro-oesophageal reflux on a squamous and columnar de- rived cell line	164
6.2.1	p63	165
6.2.2	Keratin 14	166
6.2.3	Tumor Protein P53 Binding Protein 2 and Caudal Type Home- obox 2	167
6.2.4	Effects of ASPP2 on organoid development in murine gastric tissue	171
6.2.5	Co-culture of squamous and columnar oesophageal cells	173
6.3	Summary	176

7	Conclusions	177
7.1	Single cell RNA-seq in upper gastrointestinal tissues and Barrett's oesophagus	177
7.2	Identifying single nucleotide variations in single cell RNA-seq data . .	182
7.3	Enhancing the application of single cell RNA-seq in clinical trials . .	184
A	Differentially expressed genes in cell clusters from each tissue	186
A.1	Duodenum cluster D1 - enterocytes	186
A.2	Duodenum cluster D2 - enteroendocrine cells	187
A.3	Duodenum cluster D3 - crypt cells	188
A.4	Duodenum cluster D4 - non epithelial cells	189
A.5	Gastric cluster G1 - enteroendocrine cells	189
A.6	Gastric cluster G2 - surface mucus cells	190
A.7	Gastric cluster G3 - mucus neck cells	191
A.8	Oesophagus cluster O1 - squamous cells	191
A.9	Oesophagus cluster O2 - squamous cells	192
A.10	Oesophagus cluster O3 - columnar cells	193
A.11	Oesophagus cluster O4 - columnar cells	194
A.12	Barrett's oesophagus cluster B1 - goblet cells	195
A.13	Barrett's oesophagus cluster B2 - low proliferating Barrett's cells . .	195
A.14	Barrett's oesophagus cluster B3 - high proliferating Barrett's cells . .	196
A.15	Barrett's oesophagus cluster B4 - enteroendocrine cells	197
B	Differentially expressed genes in cell clusters from all tissues combined	199
B.1	Non epithelial	199
B.2	Barrett's type	200
B.3	Enterocytes	200
B.4	Mucus neck	202
B.5	Goblet	203
B.6	Squamous	204
B.7	Enteroendocrine	206
C	Classifiers used in snv.predict	208

D Functions for the prediction of nucleotide variation in single cell	
RNA-seq	210
D.1 snv.bench	210
D.2 snv.performance	212
D.3 snv.cross.predict	214
D.4 snv.predict	215
References	220

Chapter 1

Introduction

The UK has the highest global incidence of oesophageal adenocarcinoma (OAC), though OAC in itself is not a common cancer, affecting around 5500 people a year in the UK (Offman et al., 2018). Barrett’s oesophagus (BO), a precursor lesion of OAC, is far more common, affecting 1 in 50 people, with approximately 20,000 new cases per year in the UK alone (Zagari et al., 2016). There are many areas of controversy in BO, such as how and where it originates, its relationship with OAC, and how it should be managed in a pre malignant setting. Although many other diseases occur within the upper gastrointestinal tract, this thesis is mostly limited to common events that occur in the lower part of the oesophagus.

1.1 The upper gastrointestinal tract

The upper gastrointestinal tract extends from the mouth to the duodenum (the first part of the small intestine). Between this are the pharynx, oesophagus and stomach. This clinically useful definition hides the wider relations of the upper gastrointestinal tract with the embryological foregut from which it develops, which includes parts of the respiratory tract, the biliary system, liver, pancreas and spleen. In this thesis, for clinical and practical reasons, the organs investigated are the most endoscopically accessible, limiting study to the duodenum, stomach and oesophagus.

1.1.1 Oesophagus

The oesophagus serves as an organ to deliver swallowed matter to the stomach. While this is aided by gravity, the oesophagus also has coordinated peristaltic contractions

and mechanisms to prevent the reflux of gastric content. Developmentally, the human oesophagus and trachea develop from a single columnar lined structure, which divides in the fourth week. The adult oesophagus is approximately 25cm long and is lined with non-keratinised stratified squamous epithelium, with underlying lamina propria and muscularis. Replacement of columnar epithelium for squamous epithelium occurs at 14-16 weeks development. Squamous cells develop from a 3-4 cell thick basal zone to a glycogen rich cell population in the suprabasal zone, followed by further maturation as cells flatten towards the luminal surface and finally shed (DeNardi and Riddell, 1991). Other important features of normal oesophagus are melanocytes, endocrine cells, Langerhans cells Merkel cells and oesophageal submucosal glands. The oesophagus terminates at the gastro-oesophageal junction (GOJ) which approximates the position of a sharp transition from squamous to columnar epithelium, which lines the stomach and intestine. The oesophagus has a rich blood supply in health, however after oesophagectomy, the blood supply to the replacement conduit, most commonly formed from the stomach, can be poor and contributes to the risk of anastomotic leak, an independent predictor of mortality after oesophagectomy (Raymond et al., 2016). Lymphatic drainage of the upper and middle parts of the oesophagus is into the thoracic duct and mediastinal nodes, with distal drainage into gastric and coeliac nodes, a frequent route of cancer spread contributing to the relatively low rates of curative therapy achieved in oesophageal cancer (Crosby et al., 2004, 2017).

1.1.2 Stomach

The stomach enables digestion and has endocrine roles. The stomach is divided from proximal to distal into cardiac, fundic (and corpus with respect to gastric gland type) and pyloric regions. Cardiac and pyloric glands are mainly mucus producing. Pyloric glands also contain gastrin producing enteroendocrine cells (G cells) which stimulate acid production. Fundic glands, the most common type, contain surface and neck mucus cells, chief cells in the gland base which produce a pro-protease (pepsinogen), and scattered hydrochloric acid producing parietal cells in the gland neck. Gastric enterochromaffin-like cells tend to be located in the vicinity of parietal cells and produce histamine for stimulation of acid secretion (Table 1.1).

1.1.3 Duodenum

The duodenum is segregated into four parts and connects the stomach to the jejunum. In its second part, the duodenum has an opening to allow secretions, such as bile, from the biliary system to enter the intestines. The duodenal mucosa is viliform columnar and thus distinct from the glandular gastric mucosa. Duodenal cells can be functionally separated into absorptive (enterocytes), and secretory types. Secretory cells are further divided into hormone producing (Paneth cells and enteroendocrine cells) and mucus secreting (goblet cells), among others (Table 1.1). Unique to the duodenum are Brunner’s glands, submucosal structures which produces an alkaline mucus secretion to protect the duodenum from the acid chyme from the stomach (Krause, 2000).

1.1.4 Upper gastrointestinal cells

The main gastroduodenal cell types are summarised in Table 1.1, with further characterisation of function, secretions and gene or protein marker expression used to identify cell type (Uhl et al., 2015; Haber et al., 2017; Barbera et al., 2015).

Cell name	Function	Organ	Secretion	Genes expressed
Suprabasal squamous cells	Proliferative	Oe	n/a	<i>KRT4, KRT14</i>
Basal squamous cells	Proliferative	Oe	n/a	<i>MCSP, ITGB1, TP63, KRT5, KRT14</i>
Submucosal gland	Secretory	Oe	Alkaline fluid	<i>KRT7</i>
Surface mucus	Secretory	Ga	Mucus	<i>GKN1, GKN2</i>
Neck mucus	Secretory	Ga	Mucus	<i>GKN2</i>
Parietal	Secretory	Ga	Acid; intrinsic factor	<i>ATP4A</i>
Chief	Secretory	Ga	Pepsinogen	<i>PGA3, PGA4</i>
Enterochromaffin-like	Secretory	Ga	Histamine	<i>CHGA, CHGB</i>
Enteroendocrine (G, S, I, D)	Secretory	Du/Ga	Secretin; somatostatin; cholecystokinin; gastrin	<i>CHGA, CHGB, REG4</i>
Enterocyte	Absorptive	Du	n/a	<i>ALPI, RBP2</i>
Goblet	Secretory	Du	Mucus	<i>MUC2</i>
Paneth	Secretory	Du	Anti-microbial peptides	<i>DEF5A, LYZ</i>
Tuft	Secretory	Du/Ga	Interleukin 25	<i>DCLK1</i>
Enterochromaffin	Secretory	Du	Serotonin	<i>CHGA</i>
M	Secretory	Du	Motilin	<i>MLN</i>
Cup	Unknown	Du	Unknown	

Table 1.1: Summary of the main oesophageal, gastric and duodenal cell types. Table showing the unique genes expressed by each cell type, or common genes expressed if no unique gene is known. Letters in parentheses indicate enteroendocrine cell subtypes with secretions listed in order of cell type. Oe=Oesophagus; Ga=gastric; Du=duodenum; KRT5/7/13/14=cytokeratin 5/7/13/14; MCSP=melanoma chondroitin sulfate proteoglycan; ITGB1= β 1-integrin; TP63=p63; GKN1/2=gastrophilin 1/2; ATP4A=ATPase H⁺/K⁺ transporting subunit alpha; PGA3/4=pepsinogen 3/4; CHGA/B=chromogranin A/B; REG4=regenerating family member 4; ALPI=alkaline phosphatase, intestinal; RBP2=retinol binding protein 2; MUC2=mucin 2; DEF5A=defensin alpha 5; LYZ=lysozyme; DCLK1=doublecortin like kinase 1; MLN=motilin.

1.2 Upper gastrointestinal epithelia and development

1.2.1 Upper gastrointestinal development

The gastrointestinal tract develops from the endoderm, one of the primary germ layers organised during gastrulation and responsible for nutrient uptake and transport in the developing embryo. Endodermal development depends on nodal signalling, stimulated by wingless (Wnt) signalling and inhibited by left-right determination factor 2 (LEFTY2) in a negative feedback loop (Takaoka et al., 2017). Correct folding of the endoderm, dependent on GATA binding protein 4 (GATA4) activity (Narita et al., 1997), generates the primitive gut tube. Regionalisation of the early embryo is key to proper organ development, and the gut tube is modelled from proximal to distal to give a fore, mid and hind gut, each of which are fated to specific organogenesis. The upper gastrointestinal tract derives from the foregut (along with the liver, pancreas and respiratory tract, as previously mentioned), most of the small intestine and proximal colon derives from the midgut, and the distal colon, rectum and upper anal canal derive from the hindgut. This longitudinal patterning is controlled mainly by the homeobox (HOX) family transcription factors (TFs) (Zorn and Wells, 2009), demonstrated in transgenic mouse models where disrupting HOX expression leads to underdevelopment of several key gut structures (Zkny and Duboule, 1999). Simplistically, the regulation of HOX involves Wnt, Fibroblast Growth Factor (FGF) and Bone Morphogenetic Proteins (BMP) ligands in the distal gut, and transforming growth factor- β -ligands (TGF β) and Wnt/BMP inhibitors in the proximal gut. Retinoic acid has also been shown to support the formation of organs throughout the gut tube axis (Davenport et al., 2016) (Figure 1.1). Interestingly, the clear differences seen between the epithelium types, discussed in more detail below, do not match the anatomical boundaries defined by primitive gut development, suggesting that epithelial type is driven by later molecular mechanisms which can be manipulated in adulthood, as seen in pathological metaplasia.

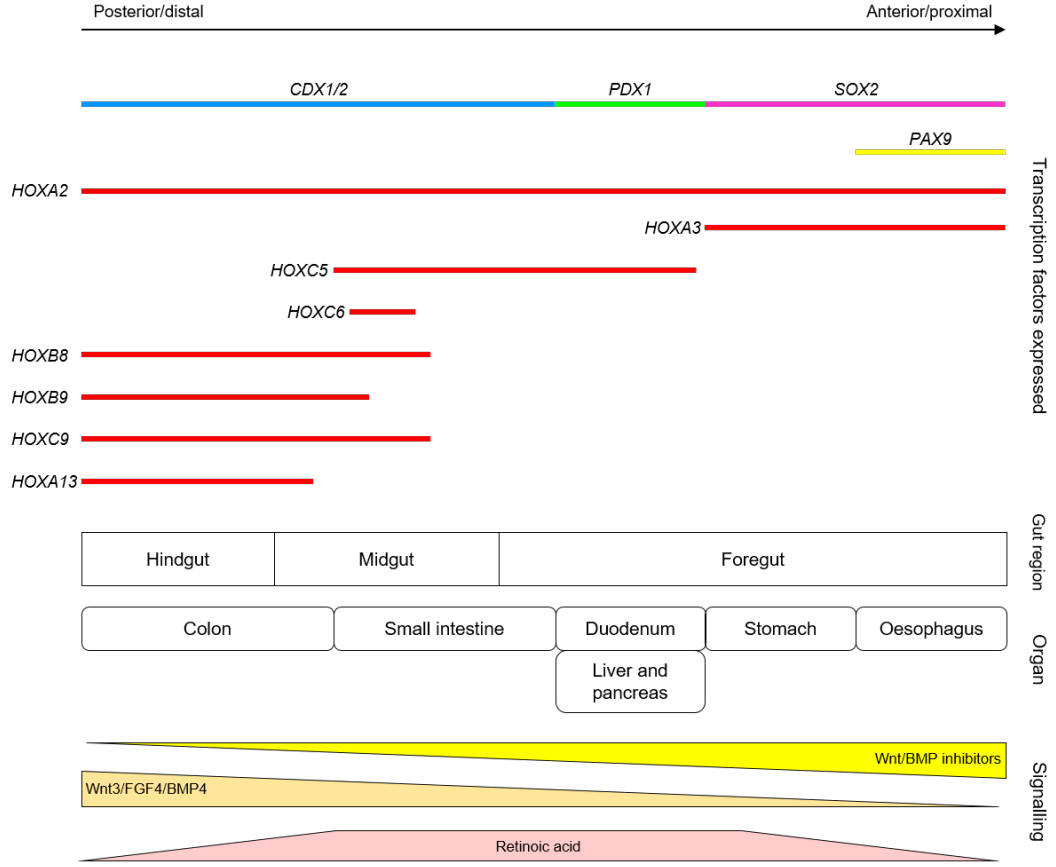


Figure 1.1: Schematic showing development of the gut. Summary of the main transcription factors and signalling pathways expressed in the longitudinal development of the gut, from posterior (hindgut, left side) to proximal (foregut, right side), with approximations of the main gastrointestinal organ fates from each developing gut region. Signalling activity is indicated by polygon or triangle heights.

1.2.2 Oesophageal epithelium

The maintenance and control of oesophageal epithelium is important to understanding how these mechanisms can fail or change in oesophageal diseases such as BO.

In the oesophagus, the epithelium consists of a two layers of keratinocytes, a proliferative layer and a suprabasal layer. The suprabasal layer undergoes terminal differentiation and passive migration before shedding into the oesophageal lumen. This proliferation-termination gradient is maintained by control of epithelial renewal through balanced cell shedding and basal layer generation. Markers of these layers are well defined with basal keratinocytes expressing keratins 5 and 14 (KRT5 and KRT14, respectively), integrins ($\alpha 6$ and $\beta 1$) and the transcription factors (TFs) p63 and SRY (Sex Determining Region Y)-Box 2 (SOX2) (Que et al., 2007; Wang

et al., 2011), and suprabasal layers expressing keratins 4 and 13 (KRT4 and KRT13, respectively), filaggrin and involucrin (Barbera et al., 2015). As described above, the oesophageal epithelium transitions from columnar-to-squamous epithelium during development. This has been demonstrated by staining mouse embryos at different Carnegie stages for the columnar marker keratin 8 (KRT8) and the squamous marker keratin 14 (KRT14) (Rishniw et al., 2011), but this study also observed that during squamous development some cells expressed both KRT8 and KRT14, suggesting a phenotypic conversion, rather than epithelial replacement or death, could be taking place (Figure 1.2).

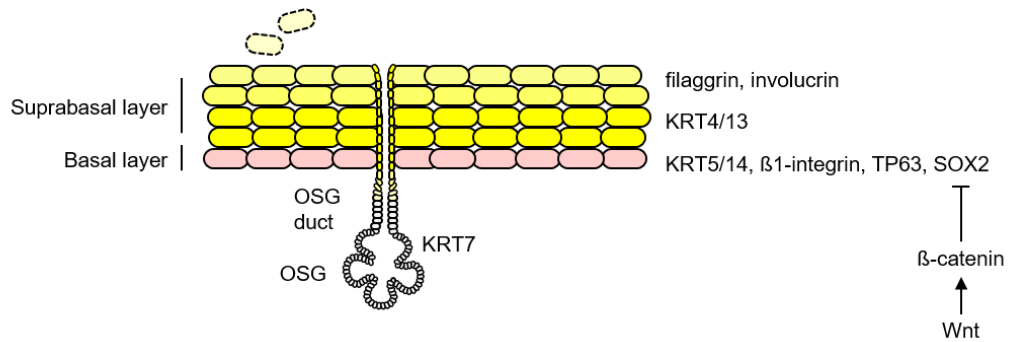


Figure 1.2: Oesophageal mucosa. Cartoon summarising the arrangement of oesophageal mucosa and key protein markers of oesophageal submucosal glands, basal epithelium and suprabasal epithelium.

In the maintenance of the adult oesophageal epithelium, *in vivo* observations and 3D culture of human oesophageal samples has shown that a niche of quiescent and slowly proliferating cells existed at the tip of squamous papillae (Barbera et al., 2015). However, although these cells expressed the putative stem cell markers β 1-integrin, MCSP and CD34, clonogenicity was not confined and other cell lineages, isolated by fluorescence activated cell sorting (FACS), were able to reconstitute stratified epithelium in organotypic culture. This suggests the progenitor capability of the oesophagus is cellularly widespread, though this observation may have been affected by the loss of tissue organisation inherent in *ex vivo* culture. In support of this widespread oesophageal progenitor theory, genetic lineage tracing, 5-ethynyl-2-deoxyuridine and transgenic cell proliferation in a mouse oesophagus injury model showed that a range of proliferative cells contributed to repair (Doup et al., 2012).

The oesophageal mucosa serves as a barrier to injury, whether mechanical, thermal or chemical. The lack of keratinisation and thick mucus coating, as seen in rodent oesophagus and columnar epithelium respectively, means that oesophageal mucosal

defence is mainly comprised of cell junction complexes, such as adherens junctions, desmosomes and tight junctions. These cell junctions complexes are formed by adhesion molecules such as the cadherins, catenins and claudins. Dysfunction of such adhesion molecules has been attributed to the development and progression of BO (Bailey et al., 1998). Further, relatively uncharacterised, structures contributing to oesophageal mucosal defence are the oesophageal submucosal glands (OSGs), which localise in the most proximal and distal parts of the oesophagus. OSGs secrete some mucus, epidermal growth factors and bicarbonate, which provides an acid buffering effect to help neutralise gastro-oesophageal reflux (Orlando, 2010). How OSGs develop is unknown, however they have been proposed to arise postnatally (in mouse studies), from minor salivary gland ingrowth or directly from ingrowths of the squamous epithelium (Long and Orlando, 1999). Since many OSGs are found in the distal oesophagus in humans, distant from the main salivary glands, another tempting proposal is that OSGs derive from residual columnar tissue present in the oesophagus before squamous epithelial replacement occurs during embryogenesis, which might explain their location and partial columnar lining (Rishniw et al., 2011).

Only a handful of molecular pathways have been reported in the development and maintenance of the oesophagus, but together they represent a complex web of signalling pathways. During development, Hedgehog (Hh) signalling allows the foregut to properly separate into trachea and oesophagus (Litingtung et al., 1998). Murine studies have shown that p63 and SOX2 function cooperatively to enable squamous epithelial development, whereas Wnt signalling functions to downregulate SOX2, promoting columnar cell differentiation in the trachea via its effector, β -catenin (Harris-Johnson et al., 2009). BMPs also play an important role, as excess BMP signalling results in a failure to complete the transition from ciliated columnar to mature stratified squamous epithelium (Rodriguez et al., 2010). Finally, oesophageal neural development has been shown to depend on SULF1/SULF2 mediated regulation of Glial cell line-derived neurotrophic factor/RET (GDNF/RET) signalling (Ai et al., 2007).

1.2.3 Gastric epithelium

Understanding how gastric epithelium develops, is regionalised and regulated remains critical to understanding oesophageal pathology, since the stomach is adjacent to the oesophagus and the columnar epithelium encountered at this boundary was once the

most obvious culprit of columnar-lined oesophagus, when BO was originally described (Barrett, 1957).

Common to all stomach regions in humans is the glandular nature of the epithelium. The gastric epithelium is regionalised into four main parts: cardia, fundus, body and pylorus. The development of the glandular epithelium of the stomach seems to lack any specific TF activation, which is surprising considering how sharp the epithelial demarcations are between the oesophagus and stomach, and duodenum and stomach. Gene expression profiling has shown tight restriction of TFs key to intestinal development, such as Caudal Type Homeobox 1 and 2 (CDX1 and CDX2) (Grainger et al., 2010), to the intestinal endoderm, pointing to a common precursor cell pool and later molecular signalling determining epithelial fate. Thus, in contrast to the selective expression of CDX1/2 in intestinal epithelium, SOX2 is expressed in the gastric epithelium and oesophagus, and reduced SOX2 expression results in defective differentiation in both organs, while ectopic SOX2 expression in intestine results in some upregulation of gastric genes (Raghoebir et al., 2012). Furthermore, intestinal metaplasia can also be induced in gastric epithelium by over expression of CDX2, as seen in transgenic mouse models. The stomach seems to regionalise normally in CDX2^{-/-} mice without reduction in gastric specific gene expression, showing that absence of CDX2 is not important in gastric differentiation (Silberg et al., 2002). These observations suggest more is at play in patterning the foregut epithelium and mesodermal factors have also been investigated. A key TF expressed selectively in mesoderm is BARX Homeobox 1 (BARX1), with BARX1^{-/-} mice showing loss of a clear gastric-duodenal boundary (Kim et al., 2005), presumably through loss of stimulation of the Wnt antagonists Secreted Frizzled Related Protein 1 and 2 (SFRP1/2), in turn preventing the suppression of CDX. Rescue of this phenotype was possible with additive Wnt antagonists, demonstrating a crosstalk which promotes intestinal development. Overexpression of BARX1 does not however induce a gastric mucosa phenotype (Jayewickreme and Shivdasani, 2015). An intriguing observation made in rat xenografts showed that gastric endoderm can develop normally on intestinal mesoderm, but equivalent experiments in chick embryos show a requirement for region specific gastric mesoderm (Koike and Yasugi, 1999).

BMPs also have a role in gastric patterning, with BMP inhibition by noggin (NOG) showing a reduction of gastric gland development, and columnar gland formation in the dorsal oesophagus at E18.5 in NOG^{-/-} mice (Rodriguez et al., 2010). Also, activation of SOX9 in mesoderm by BMP signalling induces the development of pyloric

features in the overlying epithelium (Theodosiou and Tabin, 2005), and the pyloric sphincter itself requires the TFs SOX9, Gata3 and NKX2.5 to properly form (Udager et al., 2014). Cranio-caudal gastric development requires the homeobox domain TF HOXA5 (Aubin et al., 2002), while gastric growth is modulated by Hh and Notch. For instance, Notch dysregulation impacts gastric growth by depleting underlying mesoderm, though this phenotype is rescued with the addition of recombinant Sonic hedgehog (SHH), indicating important crosstalk between these signalling pathways (Kim et al., 2011).

Taken together, these observations demonstrate the importance of region specific TF expression and restriction in gastric development (Figure 1.3).

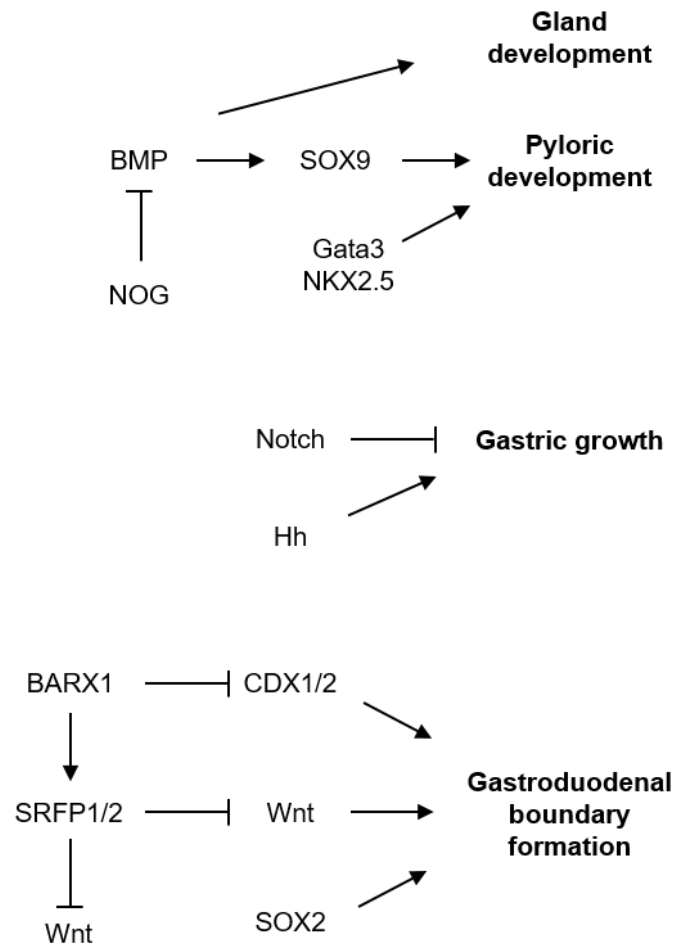


Figure 1.3: Transcription factors in gastric development. Schematic summarising the interactions and effects of some key transcription factors in gastric development and epithelial patterning.

The body (or corpus) of the stomach is responsible for most functions and contains

five differentiated cell types within gastric gland units. At the top and in the neck of the gastric glands are the surface and neck mucus cell respectively. At the bottom of the glands are chief cells, responsible for secretion of digestive enzymes. Acid and intrinsic factor secreting parietal cells are interspersed with the mucus neck cells. More rare are endocrine cells, which secrete hormones, and Tuft cells, which have no clear function. In the distal stomach in the pyloric region, gastric glands change conformation and contain less specialised cell types (Figure 1.4).

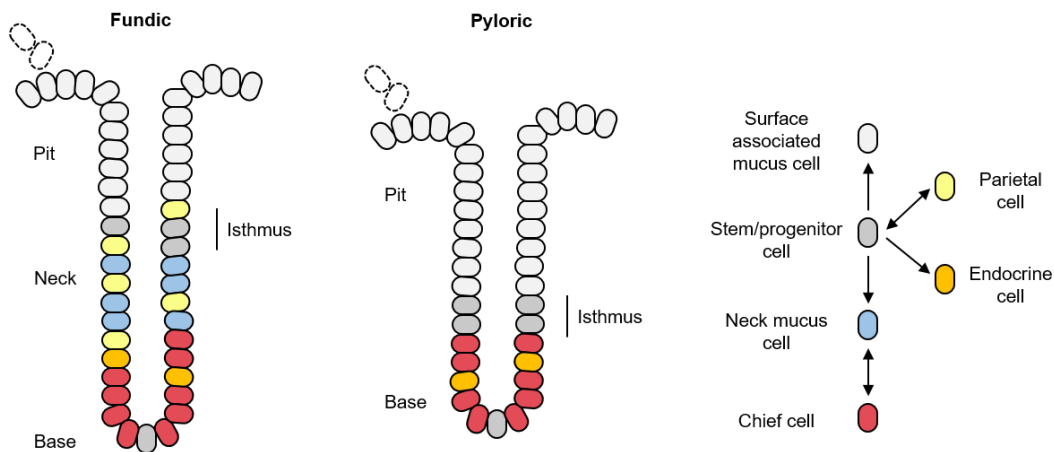


Figure 1.4: Gastric gland arrangement and cell types. Cartoons summarising the arrangement of gastric gland units in the fundus (left image) and pyloric regions (middle image). The flow chart on the right indicates the differentiation pathways of gastric cells, and the colour code for each cell type in the gland cartoons.

The isthmus of gastric glands provides the stem cell niche from which other cell types arise, as previously shown in radioactive labelling studies (Lee and Leblond, 1985), though other stem cell niches have also been described, as will be discussed. Glands start from polyclonal units, but quickly enlarge into monoclonal gland units which can subsequently undergo gland fission, leading to rapid expansion in gland density (Nomura et al., 1998). Whether gastric glands arise from a single progenitor cell is controversial, with conflicting evidence presented between species (Bjerknes and Cheng, 2002; McDonald et al., 2008). Each differentiated cell type exhibits specific molecular features, and studies in mice have identified dependent TFs for some discrete cell types though, as previously mentioned, a dominant gastric TF has not yet been identified. Important cell specific TFs that have been discovered include FOXQ1-dependent MUC5AC expression defining surface and neck mucus cells, Estrogen-related receptor gamma (ESRRG) controlling parietal cell-defining ATP4B, required for acid secretion, and MIST1 enabling chief cell differentiation (Verzi et al.,

2008; Alaynick et al., 2010; Ramsey et al., 2007). The endocrine cell population in gastric glands is subdivided into five main types, as summarised above (Table 1.1), and numerous transgenic mouse experiments have defined a regulatory network of TFs. For instance, ASCL1 is required for all endocrine differentiation, PAX6 and NGN3 are required for G and D cell differentiation, then further downstream NKX-6.3, PDX1 and ARX are required for G cells, and PAX4 is required for D cells (Kokubu et al., 2008; Jenny et al., 2002; Larsson et al., 1998, 1996; Choi et al., 2008; Du et al., 2012).

To maintain the gastric epithelium, self renewal is necessary and this relies on progenitor cells. Confusingly, different stem-markers have been observed in different cells within the same gastric glands, suggesting either multiple sources or some stem-hierarchy in gastric glands. Leucine Rich Repeat Containing G Protein-Coupled Receptor 5 (LGR5), a marker of intestinal stem cells (Barker et al., 2007), localises to cells which exhibit stem cell behaviour in the base of gastric glands in the gastric cardia and pylorus (Barker et al., 2010), but not in the body, whereas SOX2 expresses more readily in mostly spatially separate cells which are not restricted to the gland base (Arnold et al., 2011). A further VIL1+/LGR5- cell population has also been observed to activate with cytokine stimulation (Qiao et al., 2007) (Figure 1.5).

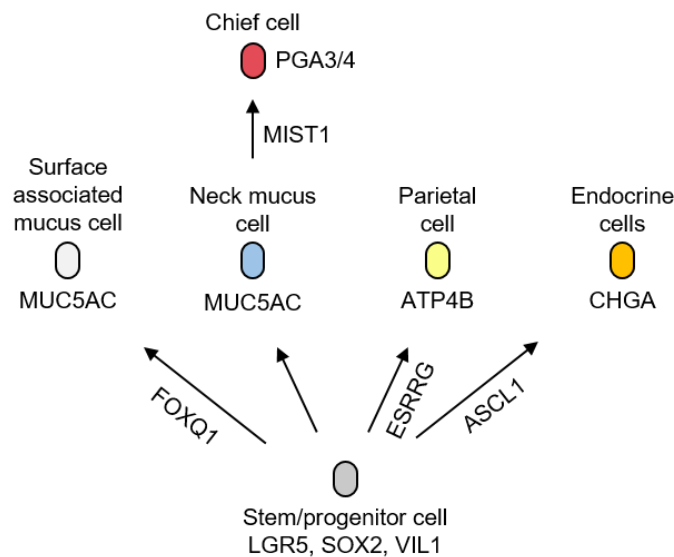


Figure 1.5: Gastric cell fates. Schematic summarising some key cell markers and proteins required for the differentiation of the main gastric cell types.

The stomach can also renew from de-differentiation of mature cells, as seen in parietal cells with ectopic Notch expression and TNFRSF19+ chief cells de-differentiating in

response to injury (Kim and Shivdasani, 2011; Stange et al., 2013). The mechanisms behind this gastric cell plasticity is unknown, though broadly permissive chromatin has been shown to underlie intestinal cell plasticity (Kim et al., 2014).

1.2.4 Duodenal epithelium

The presence of intestinal cells in BO means it is important to use intestine as a control when studying BO. Usefully, accessing the duodenum marks the completion of an upper gastrointestinal endoscopic examination, and thus duodenal tissues can be readily obtained without exposing patients with BO to prolonged discomfort or additional tests in routine practice. Duodenal mucosa comprises of numerous villi, between which are the crypts of Lieberkühn, containing specialised cell types.

As already discussed, SHH and SOX2 are required in the proximal gastrointestinal tract for proper development of the oesophagus, and to determine the GOJ, while BARX1 suppresses CDX via stimulation of Wnt antagonists. Since BARX1 is absent distal to the stomach, unsuppressed CDX expression facilitates the differentiation of intestinal tissues. The structure of mature duodenal columnar epithelium is arranged in a series of crypts, which requires Wnt, BMP and CDX2 expression for proper development (Grainger et al., 2010).

Broadly, cellular phenotypes in the intestine are controlled by Notch, via its effector hairy and enhancer of split-1 (HES1), which segregates cell lineages from a stem cell niche, via transit amplifying cells, into secretory (Paneth, enteroendocrine and goblet cells) and absorptive (enterocytes) types by repressing the TF ATOH1 (McCauley and Guasch, 2015) (Figure 1.6). This is demonstrated when Notch is inhibited with dibenzazepine, resulting in a precocious secretory cell phenotype, and reduction in OLFM4 expressing crypt base cells (VanDussen et al., 2012). Closely intertwined with Notch signalling, Wnt also has a role in driving a secretory phenotype, as disruption of canonical Wnt signalling also results in a reduction in secretory cells (Pinto et al., 2003).

Goblet cells are responsible for maintaining the mucosal barrier through mucus production, and specifically express MUC2. Goblet cell differentiation is dependent on the TF SAM Pointed Domain Containing ETS Transcription Factor (SPDEF), a downstream target of Notch, Wnt and TGF β signalling pathways, in all tissues (Noah et al., 2010). The effects of SPDEF upstream signalling pathways can be highly tissue context-dependent, possibly due to the presence other tissue specific promoters

or developmental expression chronology. For example, Krüppel like factors 4 and 5 are downstream of Notch, and knock out of KLF4 results in a 90% reduction of goblet cells in the mouse colon at postnatal day 1, but goblet cell numbers in the small intestine are relatively unchanged (McCauley and Guasch, 2015; Katz et al., 2002). Interestingly, MUC5AC, which is specifically expressed in gastric tissues, can be induced by overexpression of Notch *in vitro* (Kang et al., 2011), which may be a further mechanism by which gastric tissues avoid intestinalisation. The TFs FOXA1 and FOXA2 have also been shown to be important for goblet cell differentiation and MUC2 expression, as knock out of these genes together reduces the intestinal goblet cell population, however FOXA2 deletion in respiratory epithelium resulted in goblet cell hyperplasia, again indicating a tissue dependent context to goblet cell control (Ye and Kaestner, 2009; Wan et al., 2004). Finally, TGF β has a multitude of roles in the intestine, making it difficult to define its role in secretory lineages, however TGF β has been shown to restrict goblet cell differentiation in normal conjunctival tissues through SMA and MAD related protein 3 (SMAD3)-dependent negative regulation of SPDEF (McCauley et al., 2014). This is intriguing as loss of SMAD3 or SMAD4 is associated with gastro-oesophageal (GOe) adenocarcinoma development, which would suggest an important role for TGF β in the secretory intestinal lineage, as shown with BMP signalling, a TGF family member (Wang et al., 2014a).

Due to the shared signalling pathways that define secretory cell types, the mechanisms controlling enteroendocrine and Paneth cell (PCs) differentiation is controlled by similar mechanisms as goblet cells, though some further downstream factors have been identified that fine tune terminal secretory cell fate.

PCs protect intestinal stem cells by stimulating stem cell differentiation, control bacterial populations in the gut lumen and assist in mucosal repair. PCs secrete defensins, thought to be associated with adapting immunity, and can be marked by lysozyme expression. Control of PC differentiation is not well defined, beyond that of all secretory cell types, it has been suggested that modulation of Growth Factor Independent 1 Transcriptional Repressor (GFI1)-mediated repression of a pro-endocrine TF Neurogenin 3 (NEUROG3) may influence specific secretory lineage allocation (Bjerknes and Cheng, 2010), though NEUROG3 is a requirement for enteroendocrine cell differentiation, so if there is a role in PCs, it will be nuanced (Lee et al., 2002). Inactivation of LGR4, a paralogue of the stem cell marker LGR5, resulted in an 80% reduction in terminal differentiation of Paneth cells, but there was also a substantial overall

reduction in cell proliferation making the specificity of this finding difficult to interpret (Mustata et al., 2011). Knock out of SOX9 in mice did result in complete loss of lysozyme expression, a marker of PCs, in the crypt base, but this transgenic model also caused a marked increase in cell proliferation, although other secretory phenotype cells were unchanged (Mori-Akiyama et al., 2007).

Enteroendocrine cell differentiation in the intestine is driven by the same regulatory networks as seen in the stomach (discussed above, summarised in Table 1.1) with the TFs Pax4, Pax6 and PDX1 all implicated (Larsson et al., 1996, 1998).

Enterocytes (and colonocytes in the colon) are the only cells in the absorptive lineage, and are also the most abundant intestinal cell type. Enterocytes possess a characteristic microvilli structure at the apical border (brush border) to absorb nutrients, which are exported at their basal border. Their primary function is to absorb nutrients apically and export them basally. Counter intuitively, Notch activity is not required for enterocyte differentiation, suggesting that the absorptive lineage is default (Artavanis-Tsakonas et al., 1999). Several genes have been shown to alter enterocyte differentiation, but these have also had wide ranging effects on other cell lineages suggesting that, if there is active regulation of enterocyte differentiation, it is controlled by a more complex mechanism. For example, E74 Like ETS Transcription Factor 3 (ELF3) functions in the differentiation of both enterocytes and goblet cells, but GADD45G Interacting Protein 1 (CRIF1) has been shown to be a co-activator of ELF3, and CRIF1 knock out mice produced abnormal differentiation of enterocytes (Kwon et al., 2009). This same study also suggests a role for $TGF\beta$ in enterocyte differentiation, as RNA interference of CRIF1 attenuated ELF3 transcription, resulting in a significant reduction in $TGF\beta$ type II receptor, a target of ELF3.

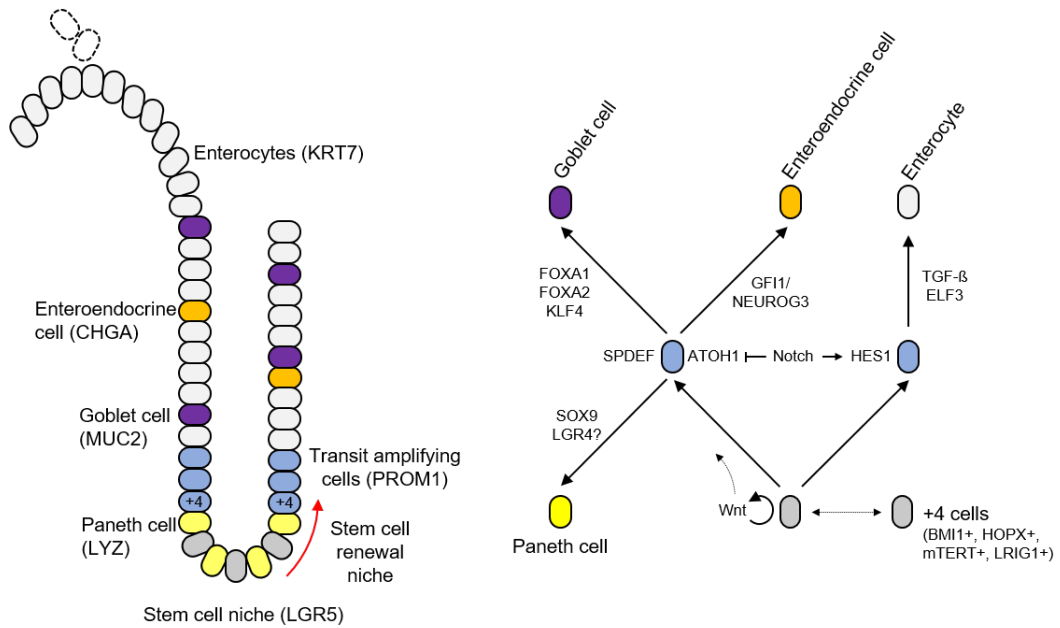


Figure 1.6: Duodenal crypt arrangement and cell types. Cartoon summarising the arrangement of the duodenal crypt structure, with cells labelled with known marker gene expression and a flow chart (right side) detailing the differentiation pathways of intestinal cells with some key factors known to influence cell lineage allocation. Tuft cells are omitted as their function is unknown.

Brunner's glands (BGs) are located beneath the surface mucosa in the duodenum. The secretions of BGs (such as bicarbonate and epidermal growth factor (Konturek et al., 1989)) have been well characterised, however the development of BGs is quite mysterious - they are known to mostly develop postnatally, and published thesis work has recognised that they derive from Leucine Rich Repeats And Immunoglobulin Like Domains 1 (LRIG1)+ stem cells in mice, and that BGs in SPDEF deficient mice do not have a secretory phenotype (Powell et al., 2016). Nonetheless, similarities between OSGs and BGs are striking, and it is tempting to think that the development of BGs might be analogous to the theory that OSGs develop from remnant columnar tissue in the oesophagus during embryogenesis.

1.3 Barrett’s oesophagus, oesophagitis and gastro-oesophageal reflux disease

1.3.1 Gastro-oesophageal reflux disease and oesophagitis

Gastro-oesophageal reflux disease (GORD) is a condition which develops when the reflux of gastroduodenal content causes symptoms or complications (Vakil et al., 2006). GORD is very common and affects most people occasionally, and up to a fifth of the UK population as a chronic issue (Dent et al., 2005). The refluxate in GORD contains two main components, acid (from gastric secretions) and conjugated bile salts (from hepatic cholesterol metabolism); also, some proteolytic enzymes have been identified. The main bile salts observed in humans are cholic acid, chenodeoxycholic acid and deoxycholic acid, which can be subdivided into primary bile acids (cholic acid and chenodeoxycholic acid), and secondary bile acids (deoxycholic acid) which result from bacterial dehydroxylation of cholic acid (Hofmann and Hagey, 2008). Bile acids observed in lower esophageal aspirates in GORD tended to be primary and of far higher concentrations in patients with oesophagitis and BO than in normal controls, underlining the association between GORD, oesophagitis and BO (Kauer et al., 1997; Nehra et al., 1999), though *in vitro* studies have most frequently used deoxycholic acid (DCA), probably because little differential effect between bile acid types has been shown in most cell culture or explant studies. In the presence of GORD, oesophagitis occurs when epithelial protective mechanisms fail and the squamous epithelium becomes inflamed due to direct exposure to refluxate. Molecularly, GORD results in a T-helper-1 driven mucosal immune response, characterised by release of pro-inflammatory cytokines such as interleukins 1 and 8 (Isomoto et al., 2003; Kandulski and Malfertheiner, 2012).

1.3.2 Barrett’s oesophagus

At least 80% of cancers arise from epithelial cells. In many tumours a change in cell type, referred to as metaplasia, is a key step in cancer initiation. Barrett’s oesophagus (BO) is one such common example of metaplasia in the distal oesophagus, although its prevalence is likely to be a significantly underestimated in the UK, as BO requires endoscopic diagnosis which can be difficult to achieve in some patient groups, and may depend on regional referral practices (Shawihdi et al., 2014). BO is defined as an oesophagus in which any portion of the normal distal squamous epithelial lining has

been replaced by columnar epithelium, which is clearly visible endoscopically above the GOJ and confirmed histopathologically from oesophageal biopsies (Barrett, 1950; Fitzgerald et al., 2014). BO is associated with GORD (Vaezi and Richter, 1996), suggesting that BO occurs in response to a chronically inflamed environment (Kavanagh et al., 2014). Furthermore, controlling GORD with proton pump inhibitor medication has been shown to reduce Barrett’s carcinogenesis (Kastelein et al., 2013b). The complex molecular relationship between BO and GORD has been shown to involve DNA damage and instability, TGF β signalling loss (due to methylation, gene deletion and protein modification), epigenetic and nuclear factor kappa-light-chain-enhancer of activated B cells (NF- κ B) mediated silencing of CDX1/2 and activation of mitogen-activated protein kinase (MAPK) signalling pathways (Kazumori et al., 2006; Wong et al., 2005; Fitzgerald et al., 2002; Fitzgerald, 2006; Souza et al., 2002).

BO gives a 30x increased risk of oesophageal adenocarcinoma (OAC) (Solaymani-Dodaran et al., 2004), a highly fatal cancer with a 15% 5 year survival, and the most common oesophageal cancer type in the UK (Cunningham et al., 2010; CRUK, 2016). OAC arises in BO via dysplasia, a stage at which relatively minor clinical procedures can halt progression to OAC (Haidry et al., 2015) (Figure 1.7).

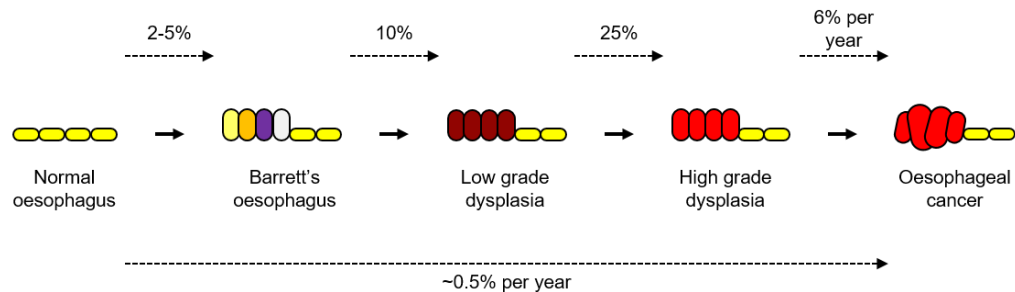


Figure 1.7: Progression of Barrett’s oesophagus to cancer. Cartoon summarising the progression of non-dysplastic Barrett’s oesophagus to oesophageal adenocarcinoma. Percentages given are for overall lifetime risk, unless stated otherwise.

Interestingly, BO has high cellular heterogeneity, with histologically identifiable gastric acid-producing cells, intestinal goblet cells, Paneth cells and, more rarely, enteroendocrine and pancreatic-type cells all being described (Chaves et al., 2005; Griffin and Sweeney, 1987; Krishnamurthy and Dayal, 1995). This apparent plasticity in BO has obscured its relationship with normal gastrointestinal tissues, as no normal gastrointestinal tissue is as heterogeneous as BO, and has led to disagreement regarding the diagnostic criteria for BO. The need for intestinal metaplasia (IM, commonly

defined as presence of goblet cells) to be histologically confirmed to diagnose BO is controversial, with practice differing in the UK and the US (Association et al., 2011; Fitzgerald et al., 2014). Whether the cell types seen in BO match their anatomically separate but histologically similar counter-parts has only been explored in limited gene and protein expression studies (McDonald et al., 2015; Lavery et al., 2014), but shedding light on this may provide a basis for harmonising the diagnosis of BO.

1.3.3 Origins of Barrett’s oesophagus

The origin and natural progression of BO remain controversial. At its original description, BO was thought to occur as a result of tethering of the stomach within the chest (Barrett, 1950). Since then, multiple theories have been proposed, which can be broadly separated into four themes.

1.3.3.1 Gastric migration

After the concept of stomach tethering was refuted, a revised proposition of gastric migration was proposed (Allison and Johnstone, 1953; Barrett, 1957; Hayward, 1961). At the time, the presence of goblet cells was not adequately addressed and other theories gathered more momentum. More recently, a better understanding of the mechanisms and influences of gastrointestinal cell differentiation combined with more thorough investigation of gastric cardia intestinal metaplasia led to the development of a compelling mouse model for gastric migration theory (Chandrasoma et al., 2001; White et al., 2008). In one elegant study, IL-1 β was overexpressed and mice were treated with DCA, resulting in a BO phenotype which was shown to derive from LGR5+ cells which resided in gastric cardia glands (Quante et al., 2012). Currently, a similar mechanism has not been adequately demonstrated in humans.

1.3.3.2 Squamous transdifferentiation

Previous observations of columnar (KRT8) and squamous (KRT14) protein expression in the same cells during mouse oesophageal embryogenesis stimulated the concept of metaplasia resulting from a pathological phenotypic switch back to columnar epithelium (Rishniw et al., 2011). Examples of this have been shown *in vitro* by applying components of bile (typically deoxycholic acid, or other secondary bile acid cocktails)

to oesophageal squamous cell lines or oesophageal tissue then observing for upregulation of intestinal epithelium-associated TFs, such as CDX2 (Kazumori et al., 2006; Hu et al., 2007), or dysregulation of squamous markers or signalling pathways (Reveiller et al., 2012). More recent work has recapitulated some of the known squamous and columnar differentiation signalling pathways using organotypic culture of human cells (Vega et al., 2014), though no one has yet shown formal de-differentiation of mature squamous cells and subsequent differentiation into the multiple stable lineages that constitute BO in a pathological model.

1.3.3.3 Squamous transcommitment

Neatly avoiding the need for de-differentiation is the theory that GORD-induced mucosal damage could expose squamous progenitor cells which could transcommit to produce columnar epithelium, then populating the lower oesophagus (Farr et al., 2008). Data to support this is thin at present, but the critical roles of Notch in cell phenotype determination and cell fate combined with evidence of Notch signalling suppression on bile exposure lends credence to this theory (Tamagawa et al., 2012, 2016). If true, progenitor cell transcommitment could explain why such a broad range of cells can exist in BO, and help shed light on how such a progenitor cell population can remain stable. Newer mouse models and 3D culture methods may help address this theory soon (Vega et al., 2014).

1.3.3.4 Junctional progenitor cell recruitment

Some specific cell niches have also been identified as potential sources of BO at the squamo-columnar junction. p63 has important roles in squamous differentiation. Since p63 suppression has been shown with bile exposure and BO does not express p63 (Roman et al., 2007), a p63^{-/-} mouse model was developed which, while not viable, showed persistent embryonic columnar epithelium resembling BO in the usually squamous fore-stomach until death, marked by CA4 expression. CA4 expressing cells were also found to reside in adult wild type mice and humans at the squamocolumnar junction (SCJ), implicating a residual embryonic stem cell population as an origin of BO. The lack of CDX2 expression has however cast some doubt on how phenotypically accurate this model is. More recently p63⁺/KRT5⁺/KRT7⁺ cells were also identified as a BO source in a transgenic mouse model with ectopic expression of CDX2 in KRT5⁺ epithelium (Jiang et al., 2017).

1.3.3.5 Oesophageal submucosal gland origin

A native oesophageal structure proposed as a source of BO is the oesophageal submucosal glands (OSGs), which have been shown to localise at either end of the oesophagus and increase in abundance beneath BO (Van Nieuwenhove et al., 2001; Lörinc and Öberg, 2015). Certainly, the presence of squamous islands, not rarely seen in BO, would support overgrowth or transformation of a pre-existing oesophageal structure (Lörinc et al., 2015; Coad et al., 2005; von Furstenberg et al., 2017). OSGs have also been shown to share specific mutations in p16 (*CDKN2A*) with overlying Barrett’s glands giving a direct genetic link from OSG to BO (Leedham et al., 2008). Demonstrating this link from normal oesophagus has not been possible yet, and would require a better understanding of the mutational profiles of BO-primed OSGs, as targeted sequencing would be the only way to achieve the required depth to confirm genetic heredity in the rare somatic mutation context of normal tissue. Avoiding these issues, it could be possible to use a mouse model to track BO development, but rodents do not have OSGs - itself a feature against an OSG theory since BO phenotypes have been observed in rodent models.

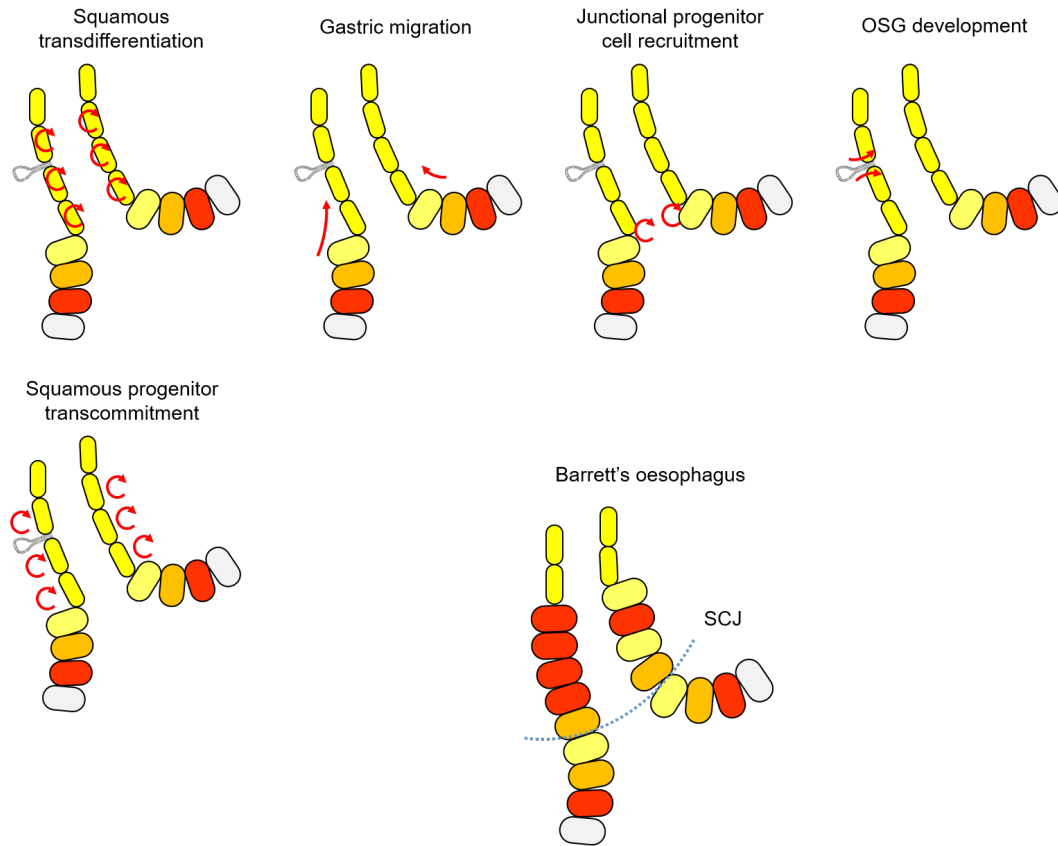


Figure 1.8: Theories of Barrett’s oesophagus development. Cartoons of the gastro-oesophageal junction (GOJ). Top row of images summarises the most recent theories of how BO develops. Lower image pictorially displays the changes seen in BO with columnar epithelium seen above the usual position of the squamocolumnar junction (SCJ), marked with a dotted line. This is endoscopically defined as the most proximal extent of gastric rugae. OSG, oesophageal submucosal gland.

Other proposed origins of BO include recruitment of circulating stem cells derived from bone marrow, shown by genotyping BO lesions from a patient who had had a bone marrow transplant from a different-sex donor, and from mice after surgically inducing GORD (Hutchinson et al., 2010). These elegant experiments were not however replicated in a similar experiment (Aikou et al., 2013), though bone marrow-derived cells have been featured in gastric carcinogenesis in separate studies (Houghton et al., 2004).

Many of the proposed BO origin theories are based on transgenic mouse studies, and the submucosal gland cell theories are based on human histopathology studies. Unfortunately, submucosal gland theories cannot be tested in mice since mice and humans have key differences in their gastrointestinal anatomy, and rodents lack oe-

sophageal glands (Macke et al., 2011). Animal models are also compromised by the methods in which metaplasia is initiated, via surgical or transgenic means, making reliable translation into human BO difficult. These difficulties argue for an unbiased and systematic genetic approach to BO characterisation in humans with all relevant control cell types as knowledge of the originating cell type of BO could be influential in improving early diagnosis and developing or validating therapies that target the causative cell populations in dysplasia and OAC. For example, ensuring mucosal treatments adequately remove the cellular source of BO, or phenotyping OAC by its BO origin to test whether some therapeutic strategies should be personalised could have a large impact on prevention and treatment of OAC respectively. Furthermore, a greater understanding of the development and progression of BO may help stratify early preventative therapies, another area of intense debate in this field with global variation in practice (Bennett et al., 2015).

1.3.4 Gastrointestinal stem cells and Barrett’s oesophagus

Considering stem cells provide the source of cells in a given structure, identifying a stem cell compartment in BO may help identify its origin. The expectation is that stem cells are rare in gastrointestinal tissues, and divide slowly (Barker et al., 2007). Accordingly, they should be negative for proliferative markers such as Ki-67.

1.3.4.1 Intestinal stem cells

Lineage tracing studies in the intestine using the proliferation marker 5-bromo-2-deoxyuridine (BrdU) showed that cells in the plus four (+4) position in the intestinal crypt were very slow cycling and retained BrdU, and hence were thought to represent the intestinal stem cell niche (Potten et al., 1974) (Figure 1.6). Marking this +4 cell position, Doublecortin Like Kinase 1 (DCLK1) was shown to be critical in tissue renewal after irradiation, supporting the +4 stem cell theory (May et al., 2008). Other markers of the +4 cell population are HOPX, LRIG1, BMI1 and TERT (Barker et al., 2012), though this is mainly based on expression localisation rather than any objective lineage studies. Doubts on whether the +4 cells represent the intestinal stem cell niche were highlighted when DCLK1 expression was shown to be unrestricted in the villus and localised to Tuft cell differentiation (Gerbe et al., 2009). Challenging the +4 stem cell theory, Leucine Rich Repeat Containing G Protein-Coupled Receptor 5 (LGR5) has been shown to mark stem cells in between Paneth cells in the intestinal

crypt base by demonstrating reconstitution of differentiated intestinal cell phenotypes from a purified LGR5+ cell population in organoid culture (Barker et al., 2007; Sato et al., 2009). Gene expression studies comparing stem cells and their progeny from mice revealed over five hundred genes that were enriched in LGR5+ cells, many of which were WNT targets, such as SOX9, AchaeteScute homologue 2 (ASCL2), EphB2, Troy and axis inhibition 2 (Muoz et al., 2012). Subsequent lineage studies validated that olfactomedin 4 (OLFM4), SPARC-related modular calcium-binding 2 (SMOC2) and ring finger 43 (RNF43) also marked LGR5+ crypt base stem cells. Unexpectedly, many +4 cell associated markers were also upregulated, casting doubt on the selectivity of some these markers. LGR5+ cells have also been shown as key players in metaplasia and cancer (Jang et al., 2015; Kobayashi et al., 2012). Further experiments have revealed increased expression of both DCLK1 and LGR5 in the intestines of tumourogenic mice, though co-expression has not been observed (Chandrakesan et al., 2017). Tying +4 and LGR5+ cells together, a recent single cell transcriptomics study found that intestinal LGR5+ cells could be separated into Mex3a^{high} and Mex3a^{low} expressing groups, with Mex3a^{high} cells located into the +3/+4 position and maintaining multipotency (Barriga et al., 2017).

1.3.4.2 Gastric stem cells

Whilst the intestines harbour a consistent structure in which to define stem cells, the stomach has more sharply demarcated regions in which the stem cell populations may differ. Consequently, gastric stem cells are less well defined. As in the intestine, LGR5 has been shown to exist in pyloric and cardiac glands, zones of pathological gastric IM, and LGR5+ cells from these regions can reconstitute differentiated gastric cell types in organoid culture (Barker et al., 2010). However, LGR5+ cells are not seen in the gastric corpus, where lineage studies have shown SOX2+ cells to develop into multiple cell types in gastric glands (Arnold et al., 2011). In the same study SOX2 was also validated as a stem cell marker in normal oesophagus. SOX2 expression is seen in the gastric pylorus but marks LGR5- cells, implying that these are separate stem cell populations. The validation of SOX2 as a stem marker in adult oesophagus presents a further intriguing observation that the dominant stem cell compartments of the gut do not match epithelial or embryological gut divisions, with the pyloric region representing an ill-defined watershed region between SOX2 and LGR5. Other proposed gastric stem or progenitor cell markers include Villin (VIL1, VIL1+ cells are usually quiescent but replicate in response to cytokine stimulation), Troy (Troy+

chief cells can repopulate multiple gastric cell types in response to epithelial injury) and TFF2 (TFF2+ cells have been lineage traced into mucus neck, parietal and chief cells) (Qiao et al., 2007; Stange et al., 2013; Quante et al., 2010; Means et al., 2008). These multiple propositions of gastric stem cells probably indicates either that different parts of the stomach are reliant on different stem cell niches, although transgenic mouse models with perturbations in several gastric stem markers have not demonstrated altered gastric patterning, or that there is a subtle undefined stem or progenitor cell hierarchy in the stomach.

1.3.4.3 Stem cells to identify an origin for Barrett’s oesophagus

A feature of intestinal mucosal growth which aids in identification of putative stem cell markers in the intestinal crypt base and +4 position is the unidirectional travel of differentiating cells in crypt structures (excepting Paneth cells), the so called ‘intestinal elevator’. This has allowed studies to use immunohistochemical or fluorescence *in situ* hybridisation to confirm cell locations, and even computational cell fate modelling (Lavery et al., 2014; Baker et al., 2014). Conversely, differentiation in gastric tissue is bidirectional from the mid gland isthmus and may change across gastric regions, making confirmatory experiments more difficult. Although IM in the gastric cardia has not been associated with any increased cancer risk over IM in other sites (Reddy et al., 2016), LGR5 has been shown in the gastric cardia and lineage traced to IM beyond the SCJ in mice, supporting the gastric progenitor migration hypothesis of BO and suggesting LGR5 might mark the stem cell niche in BO (Quante et al., 2012). In BO, LGR5 tends to localise in the gland neck, creating a similar conundrum with mapping stem cell fates as seen in gastric epithelium due to more complex differentiation dynamics (Lavery et al., 2014), although some studies have conflicted over where LGR5 is expressed, with co-expression of LGR5 with other intestinal stem cell markers (ASCL2, OLFM4) also shown in the BO gland base (Jang et al., 2015). This conflict is probably due to the differing extents to which BO can be ‘intestinalised’, hence producing a heterogeneous BO gland phenotype that can mimic stem cell dynamics from different gastrointestinal regions (McDonald et al., 2015) - increased expression of gastric stem cell markers in some but not all patients with BO would support this (Wang et al., 2014b). An unanswered question which undermines the gastric progenitor migration theory is why BO can occur after total gastrectomy in humans or oesophagoduodenostomy in rodents, where the gastric LGR5+ stem cell pool is abolished. This implies that BO could either derive from a greater range of

stem cell sources, such as those native to the oesophagus, or intestinal LGR5+ cells could also migrate, which has not been shown.

1.3.5 Molecular signalling in Barrett's

Several molecular pathways have been shown to act in response to GORD in the development of BO, such as such as BMP, Notch, Wnt, SHH, KLF4 and NF- κ B. Most of these molecular complexes parallel the signalling pathways seen in normal tissue development in related epithelia, but here are acting in a pathological context through induction or suppression by bile exposure and inflammation (Figure 1.9). A full understanding of how these signalling pathways interact and are controlled in BO development is not known, but roles for DNA methylation and microRNA in the regulation of genes important in BO and OAC initiation have been demonstrated (Zou et al., 2005; Clark et al., 2018a).

1.3.5.1 Notch

Notch signalling controls key cell fate determination in intestinal epithelium, between secretory and absorptive phenotypes, and has fundamental roles in cell proliferation and apoptosis across most tissues (Artavanis-Tsakonas et al., 1999). Notch has also been shown to affect CDX2 expression, an important TF required for the intestinalisation of BO, which acts via induction of ATOH1 and thereby affects the differentiation fate of stem cells. *In vitro* experiments show that bile acids reduce Notch signalling, which results in upregulation of the intestinal TFs CDX2 and KLF4 via unsuppressed ATOH1 activity (Figure 1.6) (Tamagawa et al., 2012; Vega et al., 2014; Tamagawa et al., 2016). These critical roles of Notch, and evidence of bile acid Notch modulation, support the BO origin theory of transcommitment of squamous progenitors (Krishnadath and Wang, 2015).

1.3.5.2 Bone morphogenetic protein

BMP signalling, part of the TGF β family, has wide ranging roles in cell proliferation, migration and differentiation, demonstrated in early embryogenesis. BMP signalling works by activating a BMP receptor heterodimeric complex to phosphorylate SMAD targets (SMAD1/5/8) which then complex with SMAD4 before being transported in to the nucleus to transcribe target genes such as Inhibitor Of DNA Binding 2 (ID2),

which has roles in numerous cellular processes (Wang et al., 2014a). Activation of BMP4 has been shown in GORD-affected oesophagus, while BMP4 is usually quiescent in normal squamous epithelium. The rationale for BMP mediated metaplasia in BO comes from observations that bile and acid exposure *in vitro* causes a BMP4 signalling increase with subsequent upregulation of CDX2 via SMAD1 activation, and that this can be rescued by inhibiting BMP with Noggin (Zhou et al., 2009; Castillo et al., 2012).

1.3.5.3 Wingless

The critical roles for Wnt in intestinal homeostasis and endodermal patterning towards columnar differentiation have led to interest in Wnt's part in the development of BO (Krausova and Korinek, 2014). The wide ranging roles of Wnt in embryogenesis have made this difficult to unpick, though direct actions on BMP signalling could be an important feature (Miyoshi et al., 2002). Defects in the effectors of Wnt, such as β -catenin, have been shown to predispose BO to carcinogenesis, but whether this is also involved in metaplasia is less clear.

1.3.5.4 Krüppel like factors

KLFs also regulate numerous biological processes, such as proliferation, differentiation and apoptosis (Ghaleb and Yang, 2017). KLF4 is a required transcription factor in goblet cell differentiation (Katz et al., 2002). High levels of KLF4 expression have been seen in BO, which is not surprising considering its intestinal characteristics. More interestingly, exposing cultured squamous epithelium to bile acids caused increased NF- κ B-controlled KLF4 promoter activity, which may point to a more direct action of GORD on KLF4 in the development of BO (Kazumori et al., 2011).

1.3.5.5 Hedgehog

SOX9 upregulation via SHH (a protein of the Hh family) induced BMP4 activation has been shown to promote columnar differentiation in oesophageal epithelium after bile acid induced injury (Wang et al., 2010). In health, SOX9 is expressed in the basal epithelium of the developing oesophagus, and this is lost once the organ matures. Observed high levels of SHH signalling and SOX9 expression in BO and OAC have led to proposals that SHH induced SOX9 could be an initiating event in BO, which

could be consistent with the residual embryonic stem cell BO origin theory (Clemons et al., 2012; Wang et al., 2011).

1.3.5.6 Mechanisms and signalling pathways in oesophageal adenocarcinoma

Inflammation in response to GORD results in inflammatory cell infiltration. In the acute phase this consists of neutrophils and monocytes whereas chronic inflammation is associated more with lymphocytes and plasma cells. These cells produce cytokines (chiefly IL-6, TNF- α and IL-1 β), reactive oxygen species (ROS) and reactive nitrogen species (RNS), all of which have been implicated in oesophageal carcinogenesis (Dvorakova et al., 2004; Dvorak et al., 2007b,a). IL-6, TNF- α and IL-1 β activate numerous signalling pathways known to drive tumour growth by stimulating angiogenesis and cell proliferation, and inhibiting apoptosis. Direct damage of DNA by ROS/RNS causing mutations and epigenetic changes leads to activation of other inflammatory pathways, further perpetuating tumour growth conditions. Important signalling pathways activated by extrinsic and intrinsic inflammation and associated with cancer progression are NF- κ B and IL-6/STAT3 (Dvorak et al., 2007a). In Barrett's carcinogenesis (OAC arising from BO) it seems that both of these signalling pathways are activated by bile exposure, as shown with *in vitro* cell line studies and demonstrated *in vivo* with increasing levels of IL-6 in a stepwise fashion from BO, to BO with dysplasia, to OAC (Dvorakova et al., 2004; Abdel-Latif et al., 2004; Dvorak et al., 2007a; Goldman et al., 2011). This indicates a mechanism by which intrinsic inflammation is activated, and correlates with the oxidative stress induced DNA damage seen in OAC (Dvorak et al., 2007b) (Figure 1.9).

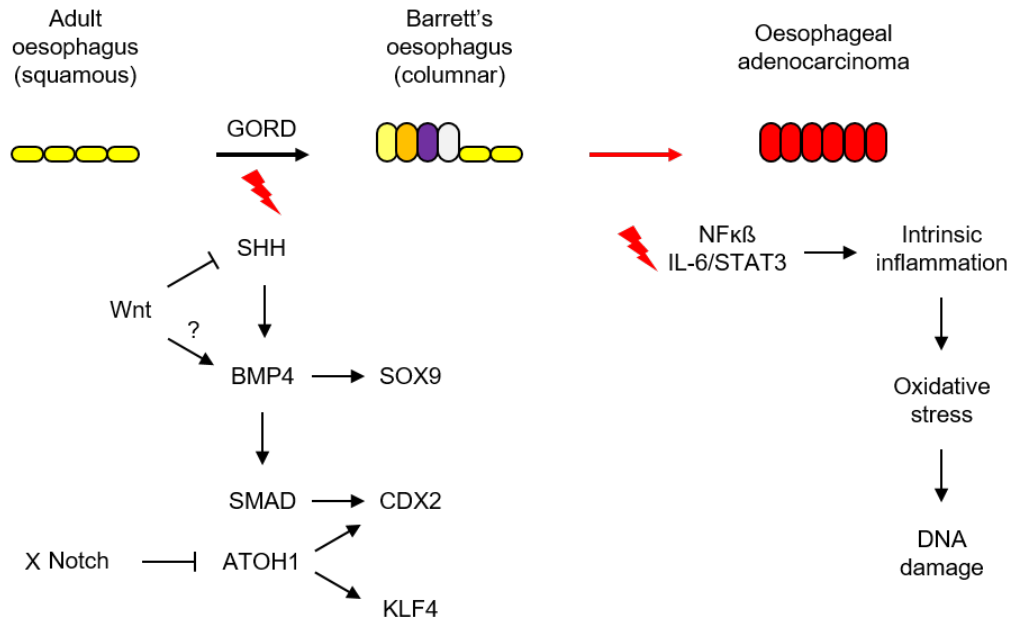


Figure 1.9: Molecular signalling in the development of oesophageal adenocarcinoma. The cartoons depict the transition from normal squamous epithelium (left image) in the adult oesophagus, to columnar BO (middle) in response to GORD/bile exposure, and OAC (right image). Beneath, some known interactions between key players in the development of BO (left side) and OAC (right side) are shown.

1.3.6 Modelling gastro-oesophageal reflux and Barrett's oesophagus

While *in vitro* 2D cell line culture of relevant cell types is useful to explore the molecular mechanisms of BO, proof demands replication of findings in a more realistic milieu, such as in animal models or 3D culture models, the most frequently used confirmatory experimental systems.

1.3.6.1 Animal models

Ideally, an animal model of BO or OAC should have genetic relevance to humans, a human-like GOJ, and naturally occurring pathophysiological GORD (Attwood et al., 2008). Numerous mammalian species have been tested as models of BO. Most GORD-BO experiments have used canine, rodent and primate models, but some investigators have also tried modelling GORD in opossums, pigs, ferrets, guinea pigs and cats. The lack of naturally occurring GORD in most species, especially in rodents (Horn et al., 2013), led to the development of several surgical approaches to simulate GORD.

Interestingly, baboons have genetic homology to humans, naturally have GORD and uniformly developed BO in a small study, but OAC was never observed (Glover et al., 2008). Refluxogenic procedures were first tried in rats, due to availability and ease of husbandry, resulting in oesophageal lesions that resembled BO and in some cases progressed to OAC (Levrat et al., 1962; Attwood et al., 1992), but concerns over mixed tumour phenotypes and unreliable reproducibility have limited their appeal. More recently, similar refluxogenic approaches have been used successfully in mice, notably in a study which showed a 60% incidence of BO in forty BALBc mice at 16 weeks post oesophago-jejunosomy, however a procedural mortality of 17% was reported (Raggi et al., 2010). A magnet based system to create an oesophago-jejunal fistula was shown to have far better survival, and implicated active matrix metalloproteases in the BO-OAC sequence (Davelaar et al., 2015). Despite these developments, there are major concerns in rodent use for simulating human BO due to key differences between rodent and human gastrointestinal anatomy (Horn et al., 2013). Rodent oesophagus does not have oesophageal submucosal glands or papillae, but does have a clearly defined layer of keratinisation, analogous to the stratum corneum of the skin, which is absent in the human oesophagus. Also, the rodent stomach contains the epithelial squamo-columnar junction (SCJ), which divides the organ into a forestomach and distal stomach, whereas the SCJ approximates the GOJ in humans. This difference in SCJ location is very important when modelling BO as demonstrating a change in SCJ position (a hallmark of diagnosing BO) and replicating oesophageal injury consistent with that seen in humans is difficult to achieve or prove in rodents.

Availability of transgenic mice led to studies combining refluxogenic procedures with genetic alterations to influence BO and OAC development. First, $p53^{-/-}$ with surgically induced GORD was shown to increase the rate of OAC development over controls, though few animals survived (Fein et al., 1999). Conflicting these observations, a larger study, which also introduced *CDKN2A* mutations, found little phenotypic effect (Hao et al., 2009). $p27^{-/-}$ models have also been investigated, showing a significantly higher incidence of BO and OAC in transgenic mice, although all of the mice survived to the study endpoint at 20 weeks (Ellis et al., 2001).

Most current BO origin theories have been based on purely genetic mouse models, without resorting to surgically induced GORD. For example, mice with defective thrombospondin-1, Hh pathway activation (using an *in vivo* SHH transgenic transplant culture system), p63 deficiency, ectopic expression of CDX2 in KRT5+ epithelium and IL1 β over-expression have all been shown to develop columnar epithelium

or intestinal features, though this is normally described within the stomach, adjacent to the SCJ (Crawford et al., 1998; Wang et al., 2010, 2011; Jiang et al., 2017; Quante et al., 2012). The latter of these models is particularly exciting as bile salts were supplemented into drinking water, demonstrating that a mechanism underlying a potential source of BO was compatible with more practical approach to simulating reflux, although the actual incidence of OAC from BO was low. Transgenic models are often criticised for oversimplifying the complex interactions of aberrant molecular mechanisms implicated in BO genesis, and several key genes in this process are lethal when mutated or deleted preventing direct transgenic study.

After rodents, canine studies have also been frequently used to mimic human BO-OAC. Canine oesophago-gastric anatomy is very similar to human, with OSGs and non-keratinised oesophageal epithelium, but the difficulties in performing refluxogenic surgery and the long time it takes for pathology to appear have led to declining use. A recent study used pigs to study OSGs by causing oesophageal injury with radio-frequency ablation, sampling from the affected area and generating a cell culture (organoid) model (von Furstenberg et al., 2017). Pigs recapitulate human oesophageal anatomy well, and this *ex vivo* study showed that OSGs were capable of producing two organoid phenotypes, a solid squamous (p63+) type and a hollow columnar (KRT7+) type leading the authors to propose OSGs as a BO source.

1.3.6.2 *In vitro* and *ex vivo* models

In view of the weaknesses in 2D culture and animal models, several human based 3D culture systems have been used in studies of upper gastrointestinal epithelium and BO. The main drawbacks of these models are the limited duration of culture which can be achieved, the lack of durable epithelial homeostasis, absence of functioning immune system, lack of heterogeneous cell interactions and limited recapitulation of mesenchymal interaction. 3D culture models are broadly split into two types, organoid and organotypic culture.

Organoid culture involves isolating progenitor cells from tissue, then culturing these in a scaffolding gel with a cocktail of additive factors to simulate the stem cell niche and producing a stable and long lived culture which can undergo cell differentiation to produce characteristic morphology, depending on the tissue (Sato et al., 2009, 2011). Organoid models are best suited to investigate growth morphology and cell differentiation fate over a limited time period, before cultures require splitting. Fragility of the

3D structure, and the requirement to encase the cells in gel limits the environmental conditions which culture can be exposed to, for instance organoid culture has not yet proven robust to bile acid exposure in our hands, but drug screening with colorectal cancer organoids has proven successful (van de Wetering et al., 2015). Additionally, gene editing techniques such as CRISPR/Cas9 can provide useful morphological and differentiation outputs in the context of targeted genetic defects (Drost et al., 2017). Organoid culture in BO has been limited, with one report of long term BO culture demonstrating the feasibility of the approach (Sato et al., 2011). OAC organoid culture has been established and used for drug screening and clonal studies (Li et al., 2018). Unfortunately, OAC has evaded targeted therapies thus far, and promising new immunotherapeutic strategies thought to be applicable to OAC require a functioning immune system, which is not a feature of organoid culture (Secrier et al., 2016).

Organotypic culture (OTC, also known as raft culture) aims to achieve physiological conditions by plating a fibroblast feeder layer with supplemented culture medium under an inert scaffold, on which epithelial cells are placed, usually on a thin supportive collagen-gel layer. Crucially the epithelial layer is exposed, producing a polarised cell sheet on maturity (and hence another related culture system - air/liquid interface culture). OTC has been widely successful in recapitulating an organised pseudo-epithelium with oesophageal cell lines and primary keratinocytes, gaining insights into interactions between fibroblasts and oesophageal cells in both benign and malignant settings (Underwood et al., 2010, 2015; Garcia et al., 2016). Simulating pathological conditions in OTC has been more difficult, presumably due to lack of physiological response mechanisms, though one group were able to recapitulate the T-helper cell type 1 acute inflammatory response found in GORD with addition of human peripheral blood mononuclear cells and exogenous cytokines (Laczka et al., 2017). Primary air-liquid interface culture of stem cells from BO and other upper gastrointestinal control tissues has also been demonstrated (Yamamoto et al., 2016). In this study, whole-genome expression analysis of stem cells and air-liquid interface cultures from BO, stomach and intestine showed that BO had unique characteristics, suggesting the origin and progression of Barrett's depends on a currently undefined discrete stem cell population.

1.4 Genetic features of Barrett’s oesophagus

1.4.1 Genomic landscape features of Barrett’s oesophagus and oesophageal adenocarcinoma

Substantial sequencing efforts have taken place in OAC, mostly by the The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), with whole genome sequencing (WGS) of 72 and 129 OACs, respectively (Cancer Genome Atlas Research Network, 2017; Secrier et al., 2016). Both studies characterise OAC as a highly mutated cancer, with single nucleotide variation per megabase (SNV/Mb) rates approaching 10, not far short of other cancers well known to be highly mutated, such as lung cancer or melanoma (Lawrence et al., 2013). Unlike melanoma, OAC has so far been quite resistant to detailed clinically useful molecular sub classifications, though this may change with further validation of the ICGC results, and a lack of identified driver events in commonly mutated genes in OAC (*TP53*, *ARID1A* and *SMAD4* (Weaver et al., 2014)) has frustrated attempts at targeted therapies.

Single nucleotide variants (SNVs) have been described within a flanking base context and identified causes in many cancer types (Alexandrov et al., 2013, 2015). In OAC, several mutational patterns have been observed in WGS and whole exome sequencing and, considering the relationship of GORD/BO and OAC, acid exposure has been hypothesized to relate to these, but this has so far not been validated (Dulak et al., 2013; Nones et al., 2014). In the largest WGS study of OAC, three main subgroups of OAC were described, including DNA damage repair, C>A/T dominant, and mutagenic (highly mutated) (Secrier et al., 2016). It will be exciting to see how well these groups can be recapitulated with clinically feasible techniques, and whether personalisation of therapy to each classification provides durable benefit.

In contrast to OAC, Barrett’s oesophagus has a significantly lower SNV/Mb rate (6), though it still approaches that of some invasive cancers, such as breast carcinoma and multiple myeloma (Ross-Innes et al., 2015a). As expected, the SNV/Mb rate has been shown to increase in dysplastic BO and OAC, though this result was not replicated in a more recent study (Stachler et al., 2015), perhaps due the issues with pathological standardisation of low grade dysplasia in between countries, a subject of ongoing debate (Kerkhof et al., 2007) (Figure 1.10).

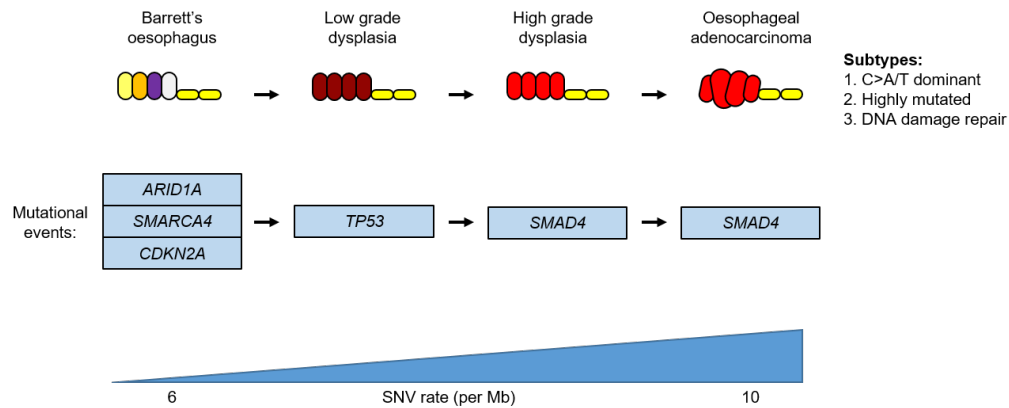


Figure 1.10: Genomic events in the progression of Barrett's oesophagus to cancer. Summary of stages of progression of BO to OAC with genomic subtypes, key mutations at each stage and effect on SNV rate during disease progression.

The persistence of selected mutations suggests that clonal competition results in an advantageous mutation 'fixing' throughout the disease, termed a 'selective sweep' (Maley et al., 2004). In BO, it has been proposed that loss of *CDKN2A* initiates this process (Wong et al., 2001). Other studies have disputed the 'selective sweep' hypothesis as significant clonal heterogeneity persists in BO (Secrier et al., 2016; Gregson et al., 2016; Stachler et al., 2015; Ross-Innes et al., 2015a; Findlay et al., 2016), and detailed genetic studies have shown that BO and OAC consists of numerous spatially separated competing clones (Leedham et al., 2008; Martinez et al., 2016). This clonal structure, and the tendency for variability in sampling protocols between studies, may account for some differences seen in mutational studies in the BO-dysplasia axis. Currently, clonality has been observed down to individual crypts, but even though these data use small amounts of tissue or cells from histological slides or brushings rather than bulk sampling, the advent of single cell sequencing would suggest a further level of resolution should be possible to ensure a true picture of the clonal landscape of BO is observed at a cellular level. Understanding the mutational architecture of BO and OAC is particularly important in light of previous results showing that genetic diversity between clones could predict neoplastic progression (Maley et al., 2006).

The metaplasia-dysplasia-neoplasia sequence has been studied with interest in the oesophagus, as genetic clues in early stages may justify preventative therapeutic approaches. Early data showing clonal *CDKN2A* and *TP53* mutations in pre-dysplastic BO suggests that mutation of such tumour suppressor genes leads to growth advantages (Barrett et al., 1999). In one study, a custom gene panel derived from WGS

was sequenced repeatedly over several years of follow up in patients, this revealed that mutations in *TP53* and *SMAD4* occurred at the bounds of normal and dysplastic BO, and OAC (Ross-Innes et al., 2015a). Further studies on TP53 have had varied results with different descriptions of mutations or aberrant protein expression in normal oesophagus or BO adjacent to OAC, though there is clearly an association which may be predictive (Kastelein et al., 2013a; Stachler et al., 2018). While there is probably further work to be done on the nature of *TP53* mutations in different grades of BO, current evidence has been deemed sufficient to recommend its use in tissue immunohistochemistry as an adjunct to H&E in clinical practice for the identification of dysplasia (Fitzgerald et al., 2014).

Copy number and somatic chromosomal alterations have also been investigated in progression of BO to OAC, but while copy number shows marked changes between BO and OAC (Ross-Innes et al., 2015a), it remained fairly consistent between non-dysplastic and dysplastic BO, and was not associated with progression to OAC until just prior to malignancy (Li et al., 2014; Stachler et al., 2018). Longitudinal studies have associated some somatic chromosomal alterations with progression to BO, but this has not yet been replicated and histological detail was scant due to the long time frame investigated (Li et al., 2015). The complicating issue with tracking changes in BO-OAC progressors is the vast number of samples needed to catch relatively few progressors, due to the overall low conversion rate (0.6% per year), meaning an unsatisfactory case-control methodology has to be relied on which biases towards tertiary care and thus loses generalisability. To overcome this, formalin fixed paraffin embedded samples have been used, which gives access to a vast historical archive of tissue, but data quality is suboptimal, with one study failing to identify expected *CDKN2A* mutations in any samples (Del Portillo et al., 2015). Due to the high prevalence of BO, and relatively low risk of OAC, it is important that any predictive test for those at highest risk of OAC can evolve with the latest available screening tools, such as minimally invasive sponge cytology and circulating DNA (Paterson et al., 2017; Rumiato et al., 2017).

1.4.2 Gene expression analyses in Barrett’s oesophagus

Whilst there are many studies exploring gene expression profiles in different contexts in oesophageal squamous cell carcinoma, studies in OAC are not common, and certainly far more limited than genomic studies in BO and OAC. Microarray data has

compared BO with adjacent normal oesophagus and shown a multitude of differentially regulated genes, but this is to be expected as these are different epithelial types (Lao-Sirieix et al., 2009; Hyland et al., 2014). However, trefoil factor 3 (*TFF3*) has been shown to upregulate in BO when compared to gastric or oesophageal controls. This feature has made (*TFF3*) useful in identifying BO using minimally invasive endoscopic tests, as being able to differentiate between gastric and BO columnar epithelium is important when defining precise anatomical location with respect to the GOJ is not possible (Ross-Innes et al., 2015b). In a study by the same group, a small but well curated set of non dysplastic BO, dysplastic BO and OAC samples were subjected to microarray analysis (Varghese et al., 2015). A 90 gene signature panel was identified that could identify high and low risk dysplasia, validated in a separate patient cohort. Ontological analysis showed that genes in a pathway regulated by retinoic acid-regulated nuclear protein were the main contributor to the 90 gene panel.

Most RNA-seq studies have been integrated with genomic characterisations and are not designed to compare BO and normal tissues, focussing instead on OAC. Transcriptional comparisons have been drawn between OAC, oesophageal squamous cell carcinoma and gastric cancer showing uniquely dysregulated genes which can define the cancer types, though in the final analysis this was integrated with other genomic features (Cancer Genome Atlas Research Network, 2017). In the other main genomic study of OAC, RNA-seq was used to validate genomic findings and in neoantigen discovery in OAC (Secrier et al., 2016). One study designed to look for differences in non-dysplastic BO, dysplastic BO and OAC shows that gene expression profiles can separate OAC from BO and normal oesophagus, but that identifying the nuance in BO pathological subtypes (non-dysplastic and low grade dysplasia) was lacking (Maag et al., 2017). Overall only 6 genes were differentially expressed between non-dysplastic BO and BO with low grade dysplasia, and no dysregulated pathways were identified. Unfortunately this study was limited in participants, hence a larger and more focussed experiment looking specifically at risk of carcinogenesis in BO would be valuable, but difficult to achieve.

In identifying the origin of BO, shared mutations in BO and a proposed source could prove the origin of BO, as has been shown within diseased tissue using whole genome amplification and targeted analysis (Leedham et al., 2008), but achieving this in the normal tissues around BO is confounded by a low mutation frequency. This low mutation frequency has specifically been observed in BO progenitors, suggesting that

mutational analysis cannot be reliably used to track the origin of BO in normal tissues. However, transcriptional analysis was able to distinguish BO stem cells from other upper gastrointestinal stem cell populations in specialised primary culture systems (Yamamoto et al., 2016). Single cell RNA-sequencing (scRNA-seq) combined with computational methods for functional clustering of cell types provides an approach to understand cellular heterogeneity (Patel et al., 2014). Given the highly heterogeneous nature of BO, scRNA-seq might clarify the phenotypic relationships between cells in normal tissues and BO, and indicate whether there are specialised cells in BO with similar functions to cells elsewhere in the gastrointestinal tract. The excitement around scRNA-seq has led to a huge range of analytical software development. Several of these packages are designed specifically to investigate changes in cell type between states or lineage (Trapnell et al., 2014; Bendall et al., 2014; Grün et al., 2015, 2016; Street et al., 2018). Since some marker genes and proteins have already been validated in stem cells of the upper gastrointestinal tract, such as LGR5 in intestine (Barker et al., 2007; van der Flier et al., 2009; Jang et al., 2015; Sato et al., 2009), it would be interesting to apply this technology in BO and in appropriate controls encompassing the range of proposed BO origins.

1.5 Single cell sequencing

1.5.1 Single cell RNA-seq

Investigating cell heterogeneity in complex organisms and tissues has been historically difficult due to limitations on the number of probes or antibodies that can be simultaneously applied in high throughput assays, such as FACS. Advances in FACS have enabled around 15 proteins to be simultaneously characterised, and when combined with mass spectrometry and metal ion labelling, this can be further increased to approximately 35-40 markers (Bandura et al., 2009). However, these approaches require pre selection of proteins of interest and are thus inherently biased, albeit relatively cost effective ways, to assay hundreds of thousands of cells quickly. Beyond this, gene specific primers have been used for single cell quantitative polymerase chain reaction on single cells enabling up to 96 primer pairs to be analysed, however the aforementioned bias persists, preventing new marker discovery (Ottesen et al., 2006).

To obtain whole transcriptome detail in a single cell requires RNA to be amplified, since a typical cell contains only about 10pg of total RNA. Early amplification meth-

ods (such as linear amplification of cDNA using T7 RNA polymerase or oligo d(T) primers conjugated to adapter sequences for reverse transcription and selective amplification of polyadenylated mRNA by PCR (Van Gelder et al., 1990; Tang et al., 2009)) were highly 3' biased. A technique that amplifies only full-length mRNA transcripts using an Moloney Murine Leukaemia Virus reverse transcriptase was later developed, substantially reducing (though not removing) 3' bias (Ramskld et al., 2012). This approach was further developed into the Smart-seq2 methodology summarised later. Another source of bias was identified due to the large amount of amplification needed in scRNA-seq, introducing duplicates of RNA species which could impact on downstream molecule quantification. To reduce this bias, RNA was labelled before amplification with unique molecular indexes, allowing identification of separate RNA species (Islam et al., 2014), although the value of this has been debatable as duplication also arises from sampling and fragmentation bias (Parekh et al., 2016).

Classically, cell selection methods have made scRNA-seq experiments difficult, with early studies relying on micro pipetting methods (Tang et al., 2009). FACS and accurate well targeting made scRNA-seq more achievable and higher throughput, although limitations persisted due to the requirement for cells to be processed manually as individual samples once sorted. Throughput was accelerated using robotic systems for sample preparation (Muraro et al., 2016), or microfluidic chip scRNA-seq (Treutlein et al., 2014). Initially, commercial microfluidic approaches were limited by the number of reaction chambers on a chip (as in the Fluidigm C1®), as samples were still kept individually, but drastic throughput increases have been realised more recently through real time labelling of RNA. This has been achieved using microfluidic encapsulation with barcoded primers (in microparticles or hydrogels, inDrop and Drop-seq) and in multiwell plates or slides with preloaded barcodes and primers (tag-based sequencing) (Klein et al., 2015; Macosko et al., 2015; Jaitin et al., 2014; Fan et al., 2015; Han et al., 2018b). Excitingly, high throughputs have also been achieved with standard laboratory equipment using a multiple combinatorial labelling approach, where cell nuclei are repeatedly re-pooled and labelled to generate a unique barcode for each cell (Rosenberg et al., 2018).

These increases in throughput do have disadvantages. Microfluidic encapsulation techniques have extremely low cell selection efficiency (7-12%) and so vast numbers of cells need to be processed through the chip to capture a meaningful approximation of the true cellular heterogeneity, especially where rare cells are concerned. As sequencing costs would be generally prohibitive with tag-based methods due to the

large number of samples, sequencing read depth requirements were many fold reduced by selecting only the 3' end of mRNA using oligonucleotide baits and sacrificing the full length transcriptome. In some systems 3' sequencing may actually produce better data for differential gene expression analysis than in full length because there is only a single fragment per transcript, reducing the over expression bias due to long transcripts producing more fragments per transcript (Tandonnet and Torres, 2017). However, 3' sequencing does not allow analysis of transcript splicing, severely restricts mutational analysis, tends to lower gene detection rates compared to full length methods and spike-in controls are not currently supported, which can be useful for normalisation and testing clustering stability (Ding et al., 2015; Severson et al., 2018). Recent comparisons of common scRNA-seq methods show that Smart-seq2 detected the most genes per cell with Drop-seq and MARS-seq detecting nearly 50% fewer genes per cell, and that Drop-seq is the most cost-effective scRNA-seq method, closely followed by MARS-seq and Smart-seq2, although Drop-seq costs are likely to be the most underestimated due to the high number of unusable cells and reduced flexibility to generate a set number of libraries (Svensson et al., 2017; Ziegenhain et al., 2017) (Table 1.2).

Method	Approach	Transcript	Cell number	Speed	Difficulty	Cost	Read depth	Benefits
C1	Microfluidic	Full	10^2 - 10^3	++	+	+++	10^6	Automated
Smart-seq2	Plate	Full	10^2 - 10^3	+	+++	+++*	10^6	Highly sensitive
MATQ-seq	Plate	Full	10^2 - 10^3	+	+++	+++	10^6	Total RNA coverage
MARS-seq	Plate	3'	10^2 - 10^3	+	+++	++	10^4 - 10^5	Early multiplexing
CEL-seq	Plate	3'	10^2 - 10^3	+	+++	++	10^4 - 10^5	Early multiplexing
Drop-seq	Microfluidic	3'	10^3 - 10^4	+++	++	+	10^4 - 10^5	Fast and high throughput
InDrop	Microfluidic	3'	10^3 - 10^4	+++	++	+	10^4 - 10^5	Fast and high throughput
10X	Microfluidic	3'	10^3 - 10^4	+++	+	+	10^4 - 10^5	Simultaneous samples
SEQ-well	Nanowell	3'	10^3 - 10^4	+++	+	+	10^4 - 10^5	Minimal equipment
SPLIT-seq	Plate	3'	10^3 - 10^5	+++	++	+	10^4	Benchmark preparation

Table 1.2: Summary of approaches to single cell RNA-seq. *Costs of Smart-seq2 are reduced 10 fold with release of Tn5 transposase protocol (Picelli et al., 2014). Difficulty and speed refer to time sent in laboratory based preparation and sample preparation, not including base equipment set up. Cost is on per cell basis with allowance made for sequencing depth requirements, though efficiency and experimental power are not factored in.

While several thousand genes per cell can be detected by scRNA-seq methods, the loss of the cellular context can be difficult to overcome in some experiments where spatial detail is required. Initially fluorescence-*in situ* hybridization allowed several genes to be probed, but more recently larger scale transcriptional information can be gained from cells without loss of context. *In situ* sequencing approaches have been used, but suffer from a lack of sensitivity and low gene detection rates. Multiplexed error-robust fluorescence *in situ* hybridization (MER-FISH) is an exciting method able to

Method	Tissue	Organism	Cells	Key features	Reference
Smart-seq2 and C1	Lung	Mouse	198	Early application of scRNA-seq	Treutlein et al. (2014)
MARS-seq	Spleen	Mouse	4468	Early cell barcoding and pooling	Jaitin et al. (2014)
CEL-seq	Duodenum (organoid)	Mouse	238	Developed RaceID	Grün et al. (2015)
Drop-seq	Retina	Mouse	44808	Single cell droplet encapsulation	Macosko et al. (2015)
STRT-seq and C1	Neural cortex	Mouse	3005	Developed BackSPIN	Zeisel et al. (2015)
CEL-seq	Pancreas	Human	3000+	Developed RaceID2 and StemID	Grün et al. (2016)
CEL-seq2	Pancreas	Human	3000+	Robotic sample preparation	Muraro et al. (2016)
Smart-seq2 and C1	Pancreas	Human	3386		Segerstolpe et al. (2016)
SPLiT-seq	Brain	Mouse	156049	Re-pooling for unique barcodes	Rosenberg et al. (2018)
Microwell-seq	Multiple (50)	Mouse	400000+	Large scale cell atlas	Han et al. (2018b)

Table 1.3: Summary of relevant single cell RNA-seq experiments. These single cell RNA-seq experiments were designed to characterise tissues. Details of sequencing approaches used and the developments each study brought to single cell sequencing are included (Key features column).

characterise 10^3 - 10^4 RNA types by combinatorial fluorescence labelling of individual RNA over multiple analysis and imaging cycles, however this is a slow approach and prone to high background fluorescence which can prevent RNA detection, therefore its use in complex tissues will require significant optimisation (Chen et al., 2015; Moffitt et al., 2016).

1.5.2 Characterisation of the human upper gastrointestinal tract with single cell RNA-seq

RNA-seq has been phenomenally useful in characterising cells and tissues, provides highly accurate measurements of RNA composition, and is relatively unbiased in comparison to qPCR and microarrays. Early RNA-seq protocols required large amounts of cDNA, which in turn required millions of cells from which to extract the necessary RNA. The averaging effects over all sequenced cells in the final expression data meant that cell heterogeneity could not be explored, and may even confound findings in complex tissues or highly heterogeneous cell populations (Patel et al., 2014). Over time, increasingly low input RNA-seq methodologies eventually resulted in scRNA-seq. as summarised in Table 1.2, early scRNA-seq protocols were laborious and expensive, limiting experiments to a few tens of cells (Qiu et al., 2012), however improvements in the sample processing chemistry, and protocol releases to the academic community, allowed laboratories to generate their own reagents (Smart-seq2, non-commercial), improving data quality and substantially reducing costs (Picelli et al., 2013, 2014). While Smart-seq2 is still used (Segerstolpe et al., 2016), many alternatives have been used for tissue characterisation experiments highlighting the technical developments that have occurred alongside biological studies in this field (Table 1.3).

Human upper gastrointestinal tissue characterisation has been limited to the pancreas and the developing gut from 6-25 weeks (oesophagus, stomach, small intestine and large intestine) (Grün et al., 2016; Muraro et al., 2016; Segerstolpe et al., 2016; Gao et al., 2018). In contrast, mouse tissues have been extensively characterised with scRNA-seq (Han et al., 2018b). In general, approaches to characterising tissues with scRNA-seq apply clustering techniques to group cells into similar subsets based on RNA composition, then attribute a phenotype to each cluster through marker gene identification, either by looking for specific markers from known biology, or using differential expression tools to identify unique cluster characteristics. Considering the interest in scRNA-seq, a multitude of clustering methods have been developed and used. A recent analysis of 12 popular clustering methods showed that Seurat and SC3 were the best performing (Satija et al., 2015; Kiselev et al., 2017; Du et al., 2018).

Other possibilities with scRNA-seq data include stem or progenitor cell identification, differentiation trajectory mapping and mutation identification (discussed below). RaceID2 (also linked to StemID (Grün et al., 2016)), which is a k-medoids based clustering tool designed to isolate a candidate stem cell population. To do this, StemID draws links between defined clusters, on the assumption that a stem-like population should link to more clusters than a differentiated cell type, and calculates entropy about each cluster mean, on the basis that higher cluster entropy should be seen in graded transcriptional changes as stem cells leave their niche to differentiate. StemID then multiplies the two values to give a stem index. This method was able to identify *LGR5* as a stem gene in mouse duodenal organoids, though contradictory results were produced when comparing pancreatic ductal candidate stem cells and lineage tracing experiments in the same tissue. Differentiation trajectory algorithms, such as Monocle, Wanderlust and embeddR (Trapnell et al., 2014; Bendall et al., 2014), order cells by perturbations in gene expression on the assumption that cells with different RNA composition profiles are likely to represent cell differentiation states. By linking cells in this way, chronology or 'pseudotime' can be applied to cell differentiation and thus an origin can be traced. These methods have been especially useful where an experimental factor or known biology can identify expected nodes in pseudotime trajectories. In a naive setting, these methods have been less useful as pseudotime does not represent a chronological direction, and they struggle to manage pluripotent differentiation trajectories and are therefore difficult to apply to scRNA-seq data from complex tissue structures.

A further important consideration is batch effects when embarking on a tissue characterisation experiment. Batch effects have been shown to undermine many published scRNA-seq studies, as controls have not been included to cope with variation between batches (Hicks et al., 2017). Theoretically, if multiple batches are required, ensuring samples cross-over or are re-permuted between batches could control for these effects.

Strikingly, most human-based scRNA-seq experiments are designed to understand normal tissues and cellular processes. Appetite to apply scRNA-seq to pathological samples has been more limited, perhaps due to noisy data restricting meaningful analysis and the difficulties in batch controlling experiments where sample acquisition is unpredictable or unreliable, such as in major resections and clinical trials. An example of this is the ambitious scRNA-seq results in melanoma where cell drop-outs in some patients and batch effects have restricted the implications of the findings (Tirosh et al., 2016). This makes scRNA-seq ideal for the study of BO, as it represents a step towards analysing more complex pathologies, such as cancer, but is rooted in easily available normal upper gastrointestinal tissues which can provide vital cell controls, ensuring that BO cellular phenotypes can be identified correctly.

1.5.3 Other single cell sequencing methods

In parallel with scRNA-seq, other single cell sequencing modalities have also developed, including some approaches which have enabled multiple genetic analyses to be applied to a single cell.

1.5.3.1 Single cell genomics

Interest in single cell genomics (scG) developed from a similar ambition to scRNA-seq, in that averaging mutations across cells in bulk samples might make identification of tumour architecture difficult to achieve. Unlike scRNA-seq, identifying the contributions of distinct cellular populations to a multicellular tissue or organism is possible from bulk sequencing, but this requires either extremely deep sequencing (especially in normal tissues where somatic mutations are rare), which is expensive to achieve, or a targeted gene approach which requires knowledge of the expected mutations. In ideal circumstances, scG could only require 200 cells to detect 1% of tumour-mass clones (Gawad et al., 2014), and the spacial separation of cells belonging to different

sub-clones can also facilitate ordering of mutations, allowing a better representation of a tumours heterogeneity and a more accurate phylogenetic tree representing a given cancer’s evolution (Ortmann et al., 2015; McPherson et al., 2016). Together, these aspects of scG have allowed for the identification of tumour vulnerabilities at specific stages, helping both the type and timing of targeted therapies (McPherson et al., 2016).

As in all single cell sequencing techniques, overcoming noisy data produced from massive amplification of starting material is a huge technical challenge. Elimination of contaminants is a significant issue which has improved with the advent of scG microfluidics (Blainey and Quake, 2011). The choice of scG approach has diverged into three main amplification techniques, each of which tend to suit different analyses.

- Degenerate oligonucleotide primed polymerase chain reaction (DOP-PCR)
- Multiple displacement amplification (MDA)
- Multiple annealing and looping-based amplification cycles (MALBEC)

DOP-PCR is a well established technique to genome amplification which uses primers that can potentially bind to many sequences, in contrast to the highly specific primers in standard PCR, improving the recovery of the genome to about 10%. MDA uses random primers to bind the genome and a polymerase that displaces other fragments on the same template, forming branching structures allowing for massive DNA amplification. MALBEC uses a similar protocol, but with common sequences and temperature cycling to promote looping of the amplicons to reduce uncontrolled amplification before PCR. Comparisons between these methods have shown that MDA has better coverage and hence higher SNV detection rates, but MALBEC and DOP-PCR have more uniform coverage allowing a better sensitivity and specificity to detect copy number variations (CNVs) (Hou et al., 2015; Huang et al., 2015). MALBEC has also been shown to have a lower allelic drop out rate than MDA (Huang et al., 2015). scG has yet to be used in BO or OAC.

1.5.3.2 Single cell epigenomics

Epigenetics is the study of the mechanisms that allow genomes to give rise to phenotypes. Therefore recognising the diversity of individual cells is fundamental to epigenetics. Similar to the themes in developing other single cell sequencing approaches,

technical constraints have previously limited epigenetic study to average behaviours over larger cell numbers, as seen in the comprehensive mapping of homogeneous cell populations and tissues used to catalogue functional elements in the human genome (Thurman et al., 2012; Kundaje et al., 2015). Several approaches to single cell epigenetic study have been developed and these are summarised below (Table 1.4) (Kelsey et al., 2017). The value of these approaches in human tissues and disease is more related to understanding transcriptional control, which requires matched phenotypic detail, ideally at the same single cell resolution. This has led to other experimental approaches to combine analyses (see below, Table 1.5).

Application	Method	Features	Issues	Reference
DNA modification	scBS-seq	Used to identify 5mC sites	40% maximum coverage, single chromosome observations	Smallwood et al. (2014); Farlik et al. (2015)
	scAba-seq	Generates recognition sites for 5hmC positions using restriction endonuclease AhaS1	Unknown false negative rate	Mooijman et al. (2016)
	CLEVER-seq	Detects 5fC by labeling the sites with specific reactivity of malonitrile and analyzing its C-T conversion		Zhu et al. (2017)
DNA accessibility	scNOME-seq	Maps nucleosome positioning and chromatin accessibility	No selectivity, so high depth required to cover areas of interest	Pott (2017)
	scATAC-seq	Uses Tn5 transposase to insert sequencing adaptors into accessible regions of the genome	Sparse data	Buenrostro et al. (2015)
	scDNase-seq	Uses deoxyribonuclease I to identify open chromatin	Low throughput, high FDR	Cooper et al. (2017)
Chromosomal organization	scHIC	Measures proximity of DNA sequences in 3-D space	Insufficient resolution for identifying specific promotor-enhancer interactions	Stevens et al. (2017)
Histone modification	scCHIP-seq	Maps post-translational modifications of histones that correlate with chromatin activity states	Antibody based sensitivity and specificity issues, sparse data	Rotem et al. (2015)

Table 1.4: Summary of epigenomic approaches applied to single cells. BS=bisulfite; FDR=false discovery rate

1.5.3.3 Combined methods sequencing in single cells

Technologies which allow defining genotype, epigenetic mechanisms, phenotype and proteome in a single cell have generated much excitement due to the mechanistic insight gained above defining phenotype. Currently, several methods exist to sequence various combinations of elements (DNA, RNA and protein) in single cells (summarised

in Table 1.5). In general, these approaches are complicated to apply resulting in loss or decay of one or more elements of interest, and hence a reduction in complexity compared to a single element technique. Consequently, single cell multi-omic led discoveries in human tissues and disease are difficult to achieve. In identifying a source cell in BO, a multi-omic approach combining the power of genomically linking related cells with epigenetic mechanism and transcriptional or proteomic phenotype would be ideal, however achieving such an experiment is extremely difficult with current techniques.

Method	Cells used	Element	Approach	Key features	Reference
DR-seq	mESC (E14) and breast cancer cell line (SK-BR-3)	G, T	gDNA amplification and cDNA synthesis without separation then split for scWGS and scRNA-seq	Genome copy number could drive transcriptome variability	Dey et al. (2015)
G&T-seq	HCC38, HCC38-BL, iPSC	G, T	gDNA and polyA mRNA magnetic bead separated for scWGS and scRNA-seq	Transcriptional consequences of aneuploidies	Macaulay et al. (2015)
scM&T-seq	E14	DNAm, T	gDNA and polyA mRNA magnetic bead separated for scBS-seq and scRNA-seq	Non-CGI promotor methylation and transcription are negatively correlated	Angermueller et al. (2016)
scMT-seq	Mouse dorsal root ganglion	DNAm, T	gDNA and polyA mRNA separated for scRRBS and scRNA-seq	Methylation of non-CGI promoters better anti-correlated with gene transcription	Hu et al. (2016)
scGEM-seq	Fibroblasts hPSC, hESC, NSCLC	G, DNAm, T	C1 chip based method	Chronology of DNAm and transcription	Cheow et al. (2016)
SMMTP-seq	GBM	P, T	Cells lysed, proteins and RNA detected by PEA and qPCR	RNA and protein were complementary in defining cell state	Darmanis et al. (2016)
scTrio-seq	HepG2, mESC, hepatocellular carcinoma	DNAm, T	DNA and RNA separated by centrifuge for scRRBS and scRNA-seq	Related the measured datasets to identify subpopulations	Hou et al. (2016)
scCOOL-seq	Mouse embryos	Cs, DNAm, CNV	Chromatin treated by GpC methyltransferase, scBS-seq of treated DNA; chromatin accessibility inferred by GpC methylation level	DNAm differs between paternal and maternal alleles but chromatin accessibility is similar	Guo et al. (2017)
REAP-seq	Human lymphocytes	P, T	10X platform, protein detected by oligo-labelled antibody	Assesses co-stimulation of CD27-agonist on CD8+ lymphocytes, identifies an unknown cell type	Peterson et al. (2017)
CITE-seq	Cord blood	P, T	10X platform, protein detected by oligo-labelled antibody	Identified phenotypes not seen in scRNA-seq alone	Stoeckius et al. (2017)
SIDR-seq	Human lung and breast cancer, MCF7, HCC827, SKBR3	G, T	Nucleus and cytosol separated with magnetic beads. mRNA and gDNA sequenced separately	CNVs correlated with corresponding gene expression	Han et al. (2018a)
scNMT-seq	mESC	Ns, DNAm, T	As in scM&T-seq with addition of GpC methyltransferase to DNA	Identifies dynamic coupling between epigenomic layers in differentiation	Clark et al. (2018b)

Table 1.5: Combination sequencing methods in single cells. Summary of the main single cell combination multiomic methods. m=mouse; h=human; g=genomic; ESC=embryonic stem cell; G=genome; T=transcriptome; DNAm=DNA methylation; P=proteome; Cs=chromatin state; BS=bisulfite; CNV=copy number variation; Ns=nucleosome; RRBS=reduced representation bisulfite sequencing; PEA=proximity extension assay.

1.5.4 Identifying single nucleotide variants in single cell RNA-seq

Whole genome sequencing of single cells has allowed some types of mutations to be detected in single cell data, though low coverage, allelic drop out and degraded DNA all pose challenges (Yuan et al., 2017). Calling mutations in next generation sequencing data such as RNA-seq has generated much interest, as these approaches are often cheaper to apply and provide robust gene expression data allowing genotype-phenotype matching. Enthusiasm for mutation calling techniques in scRNA-seq has been muted, perhaps due to the challenge of coping with noisy data and inherent issues in calling mutations from mRNA sequencing, such as variable allele frequency and gene drop-out effects between samples (O’Brien et al., 2015).

Single cell RNA-seq variant calling has been applied in colorectal cancer cell-lines to explore the relationships between cells by using software designed to call SNVs in data obtained from sequencing bulk samples, but significant variation between different callers raises some doubts as to the accuracy of the final results, as a single SNV caller was eventually used (Chen et al., 2016). Other packages designed for calling mutations in single cell sequencing data have been specific to DNA-seq, designed for use with very deep scRNA-seq, or make assumptions that may be void in aneuploidy and are therefore not ideal for use in highly heterogeneous or cancer samples (Zafar et al., 2016; Edsgrd et al., 2016). No specialised tool currently exists for single cell variant calling in scRNA-seq data, though using synthetic spike-ins, which should be devoid of any substitutions, has been used to establish a false discovery rate in subsequent analysis of SNVs called by a heuristic SNV caller (HaplotypeCaller) (Engel et al., 2017). Since a wealth of SNV callers are available and validated in RNA-seq data (Sandmann et al., 2017), an alternative approach may be to use a highly sensitive caller, such as VarScan2 (Koboldt et al., 2012), and then apply a post-calling algorithm to filter variant calls, rather than relying on inbuilt filtering techniques which are not built with scRNA-seq in mind.

Since the truth of a variant call is binary, many existing machine learning approaches are ideally suited to improving SNV call accuracy over baseline heuristic calling methodologies. This has been previously shown in bulk tumour-normal paired RNA-seq data by using deep sequencing to produce a set of true and false positives, and then extracting SNV metadata to generate SNV classifiers which are then used to construct a model for SNV prediction. To achieve this in scRNA-seq, known true

and false variant calls must first be established as a prior. As discussed above, ERCC spike-ins can serve as a false positive, and stringent measures can be applied to identify true positive mutations, such as SNP reference databases (dbSNP) and matched bulk sequencing (Sherry et al., 2001). Certainly, if variant calling in scRNA-seq can be reliably achieved, rarer cell populations in BO which would evade detection in genomic sequencing, and would be averaged out of bulk RNA-seq, could be linked to genotype or mutational signatures, potentially allowing identification of the molecular basis of how BO and OAC are associated.

1.6 Immunotherapy and treatment of oesophageal cancer

Current treatment strategies for oesophageal cancer (OC) are broadly separated into:

- Local - endoscopic mucosal resection or radiofrequency ablation for early lesions (Haidry et al., 2015)
- Definitive surgical - resection with curative intent, with neoadjuvant chemotherapy or chemoradiotherapy (Urschel and Vasan, 2003; Shapiro et al., 2015)
- Definitive non surgical - radiotherapy or chemoradiotherapy (Conroy et al., 2014)
- Palliative - symptom control, chemotherapy or chemoradiotherapy (Ajani et al., 2010)

Since the majority of OC is not suitable for curative resection at diagnosis and staging, and surgery is associated with a poor quality of life and high recurrence rate, interest in new drug treatments is high (Crosby et al., 2004; Blazeby et al., 2000). Currently, OC chemotherapy includes a combination of platinum, pyrimidine analogues (anti-metabolites), taxanes or anthracyclines. Over recent years, improved understanding of how cancers develop and thrive has resulted in therapies which treat specific tumour vulnerabilities, so called targeted therapies (Abdo et al., 2017). In OC, targeted therapies are mainly aimed at mammalian target of rapamycin (mTOR), topoisomerase enzymes and subfamilies of receptor tyrosine kinases, including epidermal growth factor (EGFR and HER2), vascular endothelial growth factor (VEGF), and hepatocyte growth factor (HGF) pathways (Figure 1.11). However, despite these

agents bringing significant benefits to other cancers, the effects in OC have been less exciting, with trials in mTOR inhibition, anti-VEGF, anti-EGFR and anti-HGFR all being negative (Ohtsu et al., 2013, 2011; Waddell et al., 2013; Lordick et al., 2013; Catenacci et al., 2017).

Some of these failures can at least partly be attributed how targeted agents are applied in OC, due a lack of good biomarkers, lack of use of existing biomarkers, and no universally consistent scheme for biomarker interpretation in relevant tissues (examples in Figure 1.11, adapted from (Abdo et al., 2017)). For example, EGFR amplification only occurs in approximately 15% of GOe cancers, leading to disappointing trial outcomes when EGFR status is not screened (Lordick et al., 2013), while Trastuzumab, a monoclonal antibody directed against HER2, showed benefit in GOe cancers with HER2 expression graded as +3 on immunohistochemistry (Bang et al., 2010). Ambitious projects have characterised the genomic landscape of OC, helping address our understanding of which therapies might be applicable to which patients with OC (Secrier et al., 2016; Cancer Genome Atlas Research Network, 2017), but it is important to remember that biomarker usage is not limited to targeted therapies, and more focussed patient selection using prognostic biomarkers for conventional treatments may give incremental benefits sooner (Abdo et al., 2017).

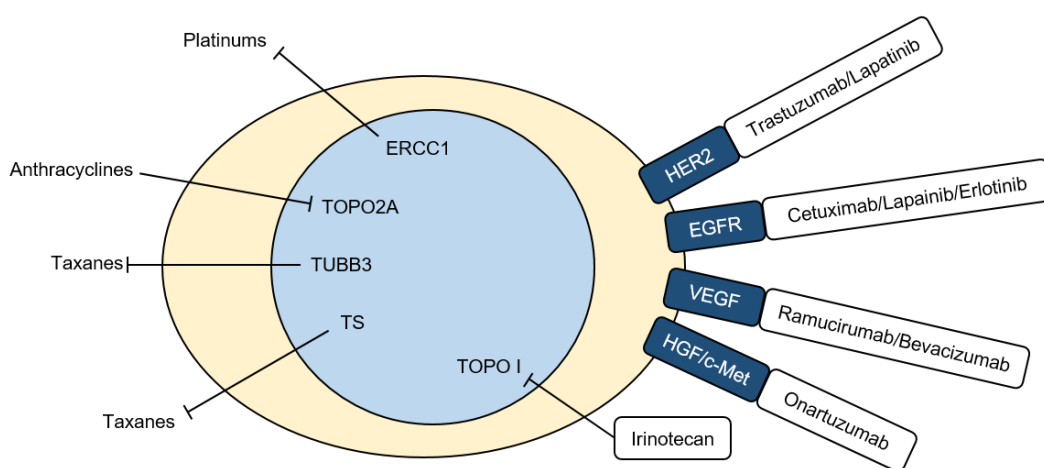


Figure 1.11: Commonly used and targeted therapies in oesophageal cancer. Image shows a representation of a cancer cell with commonly used chemotherapeutic agents listed on the left side (unboxed). Links are shown to biomarkers that have been shown to be prognostic in OC therapy for each agent. The right side shows targeted therapies (boxed) interacting with their named molecular targets. ERCC1=excision repair cross complementation group 1; TOPO2A=topoisomerase 2 α ; TUBB3=tubulin β 3; TS=thymidylate synthase; TOPO I=type I topoisomerase.

1.6.1 Immunotherapy

Immune activating therapies (immunotherapies) have shown transformational changes in the care of some cancers, such as melanoma (Wolchok et al., 2017), and have thus generated much interest in notoriously difficult to treat cancers, such as OC. Immunotherapeutic approaches in OC broadly separate into vaccine-based, adoptive T-cell transfer and checkpoint inhibition.

1.6.1.1 Cancer vaccines

Cancer vaccines are designed to directly kill cancer cells and activate immune system components, or trigger an immune system response to specific cancer cell antigens. Whether these approaches could work in OC is unknown, as currently only preliminary phase 0/I studies have been completed. So far, these trials have shown some encouraging results using anti-MAGE-A4, anti-NY-ESO1 and multi-epitope peptide vaccines, but more work needs to be done to establish vaccination as a useful approach in OC (Saito et al., 2014; Kageyama et al., 2013; Iinuma et al., 2014).

1.6.1.2 Adoptive T-cell transfer

Since the advent of tools to stably alter genetic information, it has been possible to modify T-cell specificity. The best example of this is chimeric antigen receptor (CAR) T-cell therapy, which has shown dramatic results in small numbers of children with therapy-resistant acute lymphoblastic leukaemia (Maude et al., 2014). CAR T-cell therapy involves isolation of a patient's T-cells and *ex vivo* modification to introduce a synthetic gene encoding for a special receptor (CAR) designed to bind to a certain cancer cell and cause cell death. This type of therapy is in its infancy in solid organ tumours, and only a single study has been carried out in OC using a related T-cell treatment which involves modification of the T-cell receptor (TCR), but this showed little effect in patients with recurrent OC, though the modified T-cells did show durable activity (Kageyama et al., 2015). The results of other currently recruiting adoptive T-cell trials in OC are eagerly awaited.

1.6.1.3 Immune checkpoint inhibition

In human immune regulation, numerous molecules interact to form a regulatory network of activating and inhibiting checkpoints. These checkpoint pathways serve to facilitate anti-pathogen response while simultaneously protecting cells and tissues from the effects of autoimmunity or persistent immune activation. T-cells trigger adaptive immune responses to neoplastic cells and pathogens, and are thus effectors of the immune regulating network. Malignant cells can evade immune detection through perturbations of checkpoint pathways, inhibiting the T-cell response to target neoplastic cells. This inhibition of tumour surveillance and recognition results in unrestrained propagation of malignant cells. Checkpoint inhibition consists of antibodies directed against immune checkpoints, such as programmed cell death-1 (PD-1) and one of its ligands PD-L1, and cytotoxic T lymphocyte-associated antigen 4 (CTLA-4), thereby maintaining immune surveillance and preventing evasion by malignant cells (Pardoll, 2012).

PD-1 is a member of the immunoglobulin superfamily of molecules involved in regulation of T-cell activation and has two defined ligands, PD-L1 and PD-L2. PD-L1 is expressed in many cancer cells and acts as a key mediator of PD-1-modulated immunosuppression (Brahmer et al., 2010).

CTLA-4 is expressed by T lymphocytes within the cytoplasm, and is expressed at the cell surface upon antigen activation. CTLA-4 engages B7-1 (CD80) and B7-2 (CD86) causing down regulation of the stimulation of T-cell receptors. Blocking CTLA-4 limits this negative signal, supporting T-cell activation and expansion (Ji et al., 2012).

Since anti CTLA-4 treatments have been associated with significant side effects, focus has shifted to anti PD-1 and anti PD-L1 in current OC trials. In GOe cancers, limited studies have shown variable tumour cell PD-L1 expression, from 40% to 63% in gastric adenocarcinoma, oesophageal squamous cell carcinoma (OSCC) and OAC (Ohigashi et al., 2005; Wu et al., 2006; Kollmann et al., 2018). Interestingly, melanoma trials have shown durable clinical response without tumour expression of PD-L1 expression (Daud et al., 2016), prompting alternative metrics to be used for immune responsiveness and therapy response prediction. Tumour infiltrating lymphocytes (TILs) are one such surrogate measure of immune response and multiple studies have shown increased numbers of TILs as a positive prognostic marker, including in the gut (Nosho et al., 2010). In sarcoma, TILs have been inversely associated with cancer stage,

suggesting that less immunogenic cancer cell clones are selected during cancer progression (DuPage et al., 2012). TILs have been shown to frequently express PD-1 in gastric cancer (compared to TILs in normal mucosa) suggesting that high TIL levels might be a useful biomarker for anti PD-L1 therapy. In gastric and oesophageal cancer, PD-L1 expression is associated with worse clinical outcomes, such as nodal metastases and advanced stage (Ohigashi et al., 2005; Wu et al., 2006).

In addition to TIL levels and PD-L1 expression, mutational burden, as seen in tumours with high microsatellite instability (MSI), and immune and epithelial cell heterogeneity have been shown to associate with checkpoint inhibition response. Work is ongoing to develop more sophisticated predictive approaches to incorporate these observations (Rizvi et al., 2015; Nakamura et al., 2017; Goodman et al., 2017; Morrison et al., 2018).

In GOe cancers, only three main clinical trials have been published investigating the effects of checkpoint inhibition. All of these have been in patients with advanced malignancy, frequently having progressive disease resistant to other therapies, making assessment of treatment durability impossible. Also, as only one trial has been at phase III in patients who had exhausted standard therapy, no comparisons can be made with conventional treatments.

- CheckMate-032 (phase I/II (Janjigian et al., 2018)) investigated the effects of nivolumab (anti PD-1) monotherapy, or plus ipilimumab (anti CTLA-4, in two dosage regimes) in 160 patients from the US and Europe with inoperable chemotherapy refractory GOe cancer. Patients had been heavily pretreated, with 79% receiving two or more previous therapies. Results in the three treatment cohorts showed partial clinical response in 12% (95% confidence interval (CI), 5% to 23%), 24% (95% CI, 13% to 39%), and 8% (95% CI, 2% to 19%), respectively. Significant side effects were encountered in all groups (17%, 47%, and 27%, grade 3-5), most marked in the anti CTLA-4 treatment groups.
- KEYNOTE-059 (phase II (Fuchs et al., 2018)) investigated the effects of pembrolizumab (anti PD-1) in 259 patients with inoperable gastric cancer who had failed at least two lines of therapy, from 17 countries. Partial response rate was 11.6% (95% CI, 8.0%-16.1%), with complete response in 2.3% (95% CI, 0.9%-5.0%). 17.8% of patients experienced significant side effects.
- ATTRACTION-2 (randomised, double-blind, placebo-controlled phase III (Kang et al., 2017)) randomly assigned 493 patients, from 49 sites in Asia, with in-

operable gastric or GOJ cancer who had failed at least two lines of therapy to nivolumab (n=330) or placebo (n=163). A significant but small overall and 12 month survival benefit was seen in the nivolumab group (median overall survival was 5.26 months (95% CI 4.606.37) in the nivolumab group and 4.14 months (3.424.86) in the placebo group, $p<00001$). Significant side effects were seen in 10% of patients who received nivolumab and 4% of patients who received placebo.

Interestingly, no significant variation in response between patients with different PD-L1 expression levels in tumours or MSI-high and MSI-low tumours was seen in Checkmate-032, though this could only be a hypothesis generating subgroup analysis in this trial. In KEYNOTE-059, partial response seemed to associate with PD-L1 positive tumours, however benefit was also seen in PD-L1 negative tumours, and half of the complete responders did not express PD-L1.

These trials have paved the way for a wave of registered trials looking at checkpoint inhibitors in early stage and non-treatment refractory GOe cancers. The hope is that the effects observed are more substantial, due to as yet undefined mechanisms such as the increased TIL levels seen in earlier cancers, and will allow assessment of response durability. Whilst GOe checkpoint inhibitor trials to date have not demonstrated the importance of known biomarkers generated from trials in other cancers, it is crucially important to continue gathering these data to help inform future therapeutic strategy.

1.6.2 LUD2015-005 - a phase I/II study of anti PD-L1 in combination with chemo(radio)therapy for oesophageal cancer

LUD2015-005 is a phase I/II study to evaluate anti PD-L1 (durvalumab) $\pm$ anti CTLA-4 (tremelimumab) + chemo(radio)therapy in patients with oesophageal cancer. Durvalumab is a human IgG1 mAb that blocks PD-L1 binding to PD-1, allowing T-cells to recognize and kill tumour cells. Preclinical unpublished studies have shown durvalumab to block the inhibitory effect of PD-L1 on primary human T-cells, and have demonstrated T-cell-dependent antitumour activity in vitro.

As mentioned previously, both BO and OAC have a high mutational burden (Lawrence et al., 2013; Ross-Innes et al., 2015a; Stachler et al., 2015), and OAC has a poor prognosis (Cunningham et al., 2010). Highly mutated cancers such as OAC, and

in particular the molecularly characterised mutagenic subtype (Secrier et al., 2016), have been shown to respond to immune activating therapies (Rizvi et al., 2015; Goodman et al., 2017; Morrison et al., 2018), thus there is a clinical need to apply these therapies to oesophageal cancer. Since immune and epithelial cell heterogeneity have been implicated in therapeutic response (Luksza et al., 2017; Morris et al., 2016), there is keen interest in using techniques such as scRNA-seq to better understand intra-tumoural and immune cell heterogeneity, and how this changes with systemic immunotherapy.

Embarking on a plate based scRNA-seq experiment in a clinical trial context is challenging due to problems with consistent sampling, sample processing, costs and throughput. With this in mind, a series of modifications were made to a standard plate based single cell work flow to increase its feasibility in a translational context.

- Tissues preserved for later processing
- Robotic sample preparation to increase throughput
- Miniaturising sequencing reactions to reduce costs
- Sequencing depth reduction to reduce costs and reduce batch effects

Each of these modifications is aimed at addressing some of the issues encountered in adapting plate based scRNA-seq to bedside science. Further introductory details are in Chapter 5. Some of these modifications are then put into effect in a pilot experiment to generate data from a trial patient undergoing multi-time point scRNA-seq to test whether expected biological signal is captured (Figure 1.12). Additional WGS, RNA-seq, circulating DNA and immune cell clonality data have been collected, but are not analysed as part of this thesis.

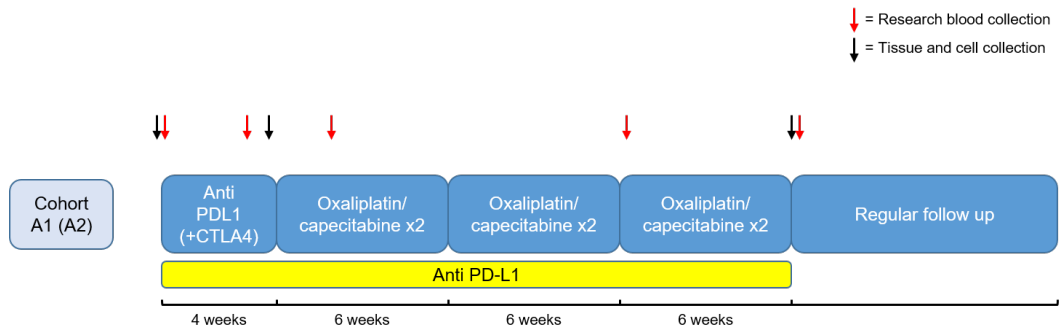


Figure 1.12: Summary of cohort A in LUD2015-005 trial. Outline of cohorts A1 and A2 in LUD2015-005 with tissue, cell and blood sampling schedule. Samples are taken at baseline, after immunotherapy but before any conventional therapy, and after conventional therapy. Later cohorts include patients with resectable disease and neoadjuvant therapy with anti PD-L1 monotherapy.

The ultimate aim of this project is to capture cancer cell and immune cell functional and genomic heterogeneity in patients before and after PD-L1 therapy (with and without other therapeutic strategies in later cohorts) to identify molecular cues which could help stratify patients to immunotherapy in the future. Details on additional recruiting cohorts for which cell and tissue storage is ongoing is available (www.oncology.ox.ac.uk/trial/lud2015-005).

Chapter 2

Materials and Methods

2.1 Human tissue sampling

Patients attending routine endoscopic surveillance of Barrett’s oesophagus and patients with mild reflux symptoms undergoing gastroscopy for diagnostic purposes gave written informed consent and provided samples (patients A-F and I-Q, South Central - Oxford C Research Ethics Committee: 09/H0606/5+5; patients G-H Yorkshire & The Humber - Sheffield Research Ethics Committee: 16/YH/0247). Patient numbers were chosen to provide suitable biological replicates, and cells sequenced to provide balanced sample sizes at sequencing input. Double bite quadrantic 2mm biopsies were obtained endoscopically using standard biopsy forceps (Radial Jaw 4 Standard Capacity, Boston Scientific, Natick, USA) from a central region of the BO segment avoiding the proximal BO margin and gastro-oesophageal junction (GOJ). Control samples were taken from the second part of the duodenum, the stomach 20mm distal to the GOJ and the normal oesophageal squamous epithelium 20mm clear of the most proximal extent of BO. Each sample was fragmented and then pooled to ensure all sampling sites were represented in each investigative modality. Fragments pools were divided into three groups for histological verification, single cell RNA-seq, and whole-tissue RNA-seq where used (figure 2.1).

Patient	Age	Sex	Smoking (cigs/day)	Alcohol (units/week)	BMI	Prague	scRNA-seq	Bulk RNA-seq
A	78	M	0	21	26.2	C3M5	Yes	Yes
B	77	M	0	4	25.7	C0M3	Yes	Yes
C	47	F	0	16	22.2	C0M2	Yes	Yes
D	90	F	0	6	23.7	C9M11	Yes	Yes
E	82	F	0	0	27.1	na	Yes	No
F	68	F	0	0	25.8	na	Yes	No
G	76	M	10	8	27.5	C1M3	Yes	No
H	78	M	0	22	29.4	C2M4	Yes	No
I	87	M	ex 10/day	8	21.9	C8M10	No	Yes
J	75	F	0	2	35.7	C1M3	No	Yes
K	74	M	ex 1/day	24	25.5	C2M3	No	Yes
L	63	M	ex 15/day	10	22.8	C0M3	No	Yes
M	70	F	0	4	32.6	C5M7	No	Yes
N	59	M	0	15	41.0	C0M1	No	Yes
O	67	F	0	0	31.7	C13M15	No	Yes
P	65	F	30/day	3	27.3	C0M1	No	Yes
Q	49	M	20/day	0	42.9	C3M4	No	Yes

Table 2.1: Patient metadata summary. Details of patients’ samples having single cell and bulk RNA-sequencing. BMI, body mass index; PPI, protein pump inhibitor treatment. Prague indicates Prague classification for measuring the length of Barrett’s oesophagus, where C is proximal extent of circumferential lesion and M is maximal proximal extent (na for patients without BO). scRNA-seq and Bulk RNA-seq indicate whether samples from each patient were used for single cell or bulk RNA-seq, respectively, total n=17.

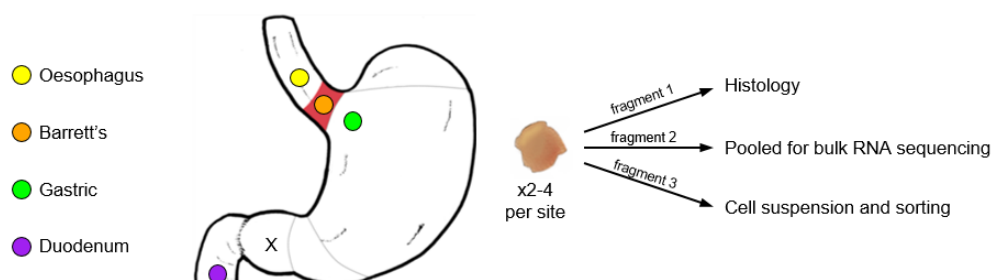


Figure 2.1: Upper gastrointestinal sampling. Approach for sampling upper gastrointestinal samples from human patients. Multiple biopsies were obtained endoscopically from each patient then fragmented, pooled and split into groups depending on the number and type of downstream uses. Samples were taken from the pylorus (marked with black X) in an additional later experiment.

Patients were selected based on their previously known pathological features (Table 3.1). Patients without BO described 0-2 reflux episodes per week with normal endoscopic appearances of the upper gastrointestinal tract on endoscopic examination, and no histological evidence of oesophagitis in the processed samples.

Oesophageal cancer samples were taken as part of the LUD2015-005 trial (EudraCT number 2015-005298-19). Sampling in LUD2015-005 differed from other upper gas-

trointestinal sampling in that multiple spatially separated samples were taken from the tumour. This restricted the number of control samples that could be taken, and typically the duodenal samples were omitted, then gastric samples, depending on how many tumour samples were required. Additionally, in cases of endoscopically impassible using standard endoscopic equipment, sampling was restricted to tumour and normal oesophagus (Figure 2.2).

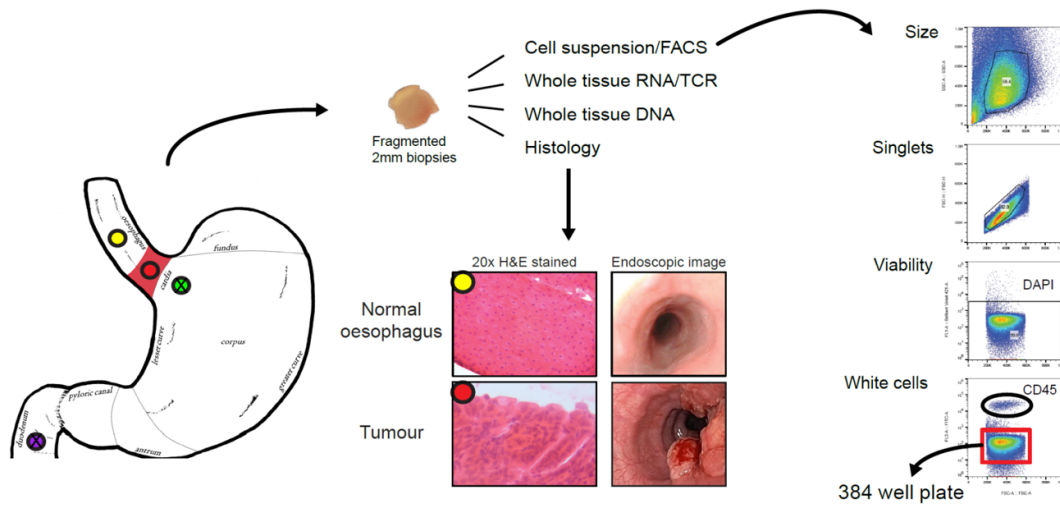


Figure 2.2: Upper gastrointestinal sampling in LUD2015-005 trial. Approach for sampling upper gastrointestinal samples from patients in LUD2015-005 trial. Multiple biopsies were obtained endoscopically from each patient from the indicated sites, though not consistently from stomach or duodenum (green and purple, marked with black cross). Tumour sampling was more extensive than in Barrett’s oesophagus, with up to ten spatially separated samples taken across the lesion. Immune cell filtering with CD45 was not applied in the pilot data shown later.

2.2 Cell isolation

Sample fragments were placed directly into a digestion solution (all buffers listed below in section) and gently oscillated at 4°C for 60 minutes. Samples were then further fragmented with scissors and briefly manually triturated with a p1000 pipette. Fragments were allowed to settle and the cell-containing supernatant filtered (Sysmex Celltrics® 100 micron) into a 15ml Falcon tube. This process was repeated 3 times and the product centrifuged at 300xg for 20 minutes at 4°C to create a cell pellet which was resuspended in sorting buffer. A small amount of each sample was pooled for labelling controls. Pre-conjugated CD45-FITC (1:10, mouse monoclonal, cat. 130-080-202, Miltenyi Biotec) (Kurian et al., 2013) and EpCAM-PE (1:10, mouse

monoclonal, cat. 130-110-999, Miltenyi Biotec) (Metsuyanin et al., 2009) antibodies were added to cell suspensions to help identify epithelial and immune cells, respectively, and they were incubated/washed according to manufacturer’s advice. DAPI (1:2000, Sigma-Aldrich®) was added to cell suspensions immediately prior to sort. Fluorescence activated cell sorting (FACS) was carried out using a BD Biosciences FACS Aria IIIu platform with 70 μ m nozzle in the case of the first four patients and the additional squamous samples, and a Sony SH800S Cell Sorter with 100 μ m chip in the second batch of two patients including the pyloric samples and LUD2015-005 trial samples (Figure 2.2). Cells were selected based on size and singlet gating to saturate cell output while minimising debris passed to subsequent gates. Size and singlet gating were then adjusted to capture EpCAM+ cells, on the basis that these would represent a range of epithelial cells and minimise debris selection. Resultant cells were sorted directly into 96 well plates (Life Technologies™ MicroAmp® Optical 96-well Reaction Plate) pre-prepared with cell lysis reagent (0.2 μ l 0.2% Triton™ X-100 (Sigma-Aldrich®) and RNase inhibitor (Takara Recombinant RNase Inhibitor) at 19:1) and then immediately frozen on dry ice. Trial samples were sorted into 384 well plates (FrameStar® 384 Well Skirted PCR Plate, 4titude®), pre-prepared with 0.1 μ l cell lysis reagent (Echo® Acoustic Liquid Handling, LabCyte™). To confirm sorting accuracy in 384 well plates, a TMB-HRP reporter assay (see next section) was used to confirm accurate droplet placement (a positive result causes the well to turn blue). Accuracy was typically >95% (Figure 2.3, sort as used in scRNA-seq data from trial patient, 98.2% of well are blue).

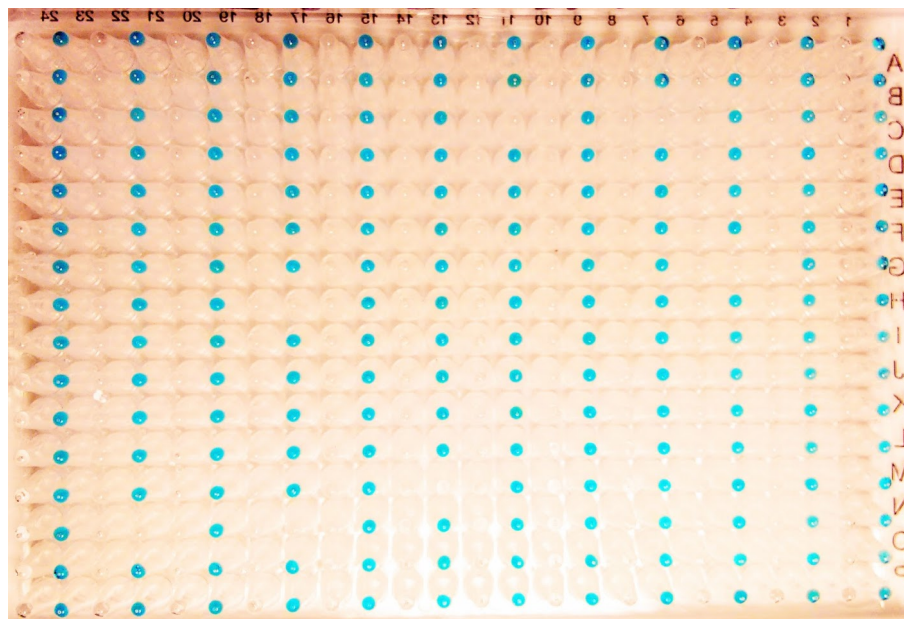


Figure 2.3: Strategy used to confirm accurate FACS droplet sorting. Results from sorting into a 384 well plate using the Sony SH800S Cell Sorter. A portion of the same cell suspension then used for scRNA-seq was separated and HRP added before sorting into a 384 well plate pre-prepared with TMB. Accurate droplet placement causes 'blue-ing' of the well.

To confirm spectral accuracy, compensation bead controls and pooled cell suspensions were used for fluorescence-minus-one controls where possible. Each plate was re-permuted to avoid batch effects at the next stages of preparation, with no single plate containing cells from only a single patient or tissue type (aside from the clinical trial patient). Variable patterns of 6 blank wells were also prepared in each plate, 3 of which had a 10pg of brain total RNA (Agilent Technologies) added as a positive control. A single 100 cell pool was also sorted in experiments using the Sony SH800S Cell Sorter to provide a bulk control as whole tissue RNA-seq was not performed in these patients. To check for bias in cell selection, index sorting was carried out in most experiments to analyse expression of antibodies in relation to tissue type and subsequent data quality. Using the input metrics available up to the point of sequencing, logistic regression was also undertaken to see if higher quality cell data could be predicted before sequencing. While the length of the experiment tended towards having an effect on data quality, recorded metrics at FACS could not accurately predict whether a cell would meet a read count threshold (see chapter 5).

2.3 TMB-HRP reporter assay

3,3',5,5'-Tetramethylbenzidine (TMB) and Horseradish Peroxidase (HRP) were used to assess the sorting efficiency of the Sony SH800S cell sorter in 384 well plates. Aliquots of 0.1 μ l TMB (BioLegend, kept at 4°C) were placed into all wells of a 384 well plate (Echo[®] Acoustic Liquid Handling, LabCyte[™]). Cells from the same disaggregation (and subsequently used for scRNA-seq from the same sorting session) were centrifuged at 500xg for 4 minutes and resuspended with 500 μ l PBS, then centrifuged again at 500xg for 4 minutes. The sample was then resuspended with 375 μ l PBS and 125 μ l HRP (Life Technologies, kept at -20°C), making a 1:4 dilution of HRP. After setting up the sorter and checking alignment onto a covered plate, HRP sample was sorted into every other column of the plate containing TMB then the plate was centrifuged briefly (Fisherbrand[™] Mini Plate Spinner Centrifuge, ThermoFisher). The plate was then kept at 4°C for 10 minutes before counting the number of wells which have turned blue (the sorting efficiency).

2.4 Single cell RNA-seq

Transcriptome libraries were prepared using a Biomek FX liquid handling instrument (Beckman Coulter) with the published smart-seq2 method (Picelli et al., 2013, 2014), and Nextera XT (Illumina[®]) methodology (as per the manufacturer's guidance) with custom, unique index primers (provided by the Wellcome Trust Centre for Human Genetics) after tagmentation and ERCC spike-in at a dilution of 1:100,000 (as described in (Salit et al., 2011)). Libraries were sequenced using the Illumina[®] HiSeq 4000 platform, aiming for 3.5x10⁵ reads per cell at 75 base pair (bp) paired end.

2.5 Bulk RNA-seq

Tissue fragments were processed using the mirVana[™] miRNA Isolation Kit (ThermoFisher) according to manufacturer's guidance. Total RNA was enriched using ribodepletion (Ribo-Zero, Illumina[®]) prior to cDNA conversion. Second strand DNA synthesis incorporated dUTP. cDNA was end repaired, A-tailed and adaptor ligated. Samples then underwent uridine digestion. The prepared libraries were size selected and multiplexed before 75bp paired end sequencing using the Illumina[®] HiSeq 4000 platform.

2.5.1 RNA-seq pre-processing and quality control filtering

All sequencing data were mapped using STAR (release 2.5.2a, <https://github.com/alexdobin/STAR>, (Dobin et al., 2013)) to the hg19 version of the human genome with transcriptome annotations from Gencode (release 25). Duplicated reads were marked using Picard (version 1.9, <https://github.com/broadinstitute/picard>). Counts tables were made with HTSeq (version 0.5, https://htseq.readthedocs.io/en/release_0.11.1/, (Anders et al., 2015)). For phenotype analysis, cells were excluded that did not meet a threshold set to exclude all negative controls and outliers, and includes all remaining positive controls (Chapter 3). For example, this was fewer than 25,119 fragments mapping to the transcriptome in the first experiment (chapter 3, patients A-D), set at a level to exclude negative and include positive (diluted neural RNA) controls. To check the agreement between single cell and bulk RNA-sequencing, counts from the most abundant cell population from a single patient and tissue were summed by each detected gene and then correlated against the same genes in bulk RNA-seq expression (Pearson correlation) to confirm similar gene expression profiles. Before final quantification of gene expression, counts were normalised (by trimmed mean of M-values (TMM) in edgeR (version 3.18.1, <https://bioconductor.org/packages/release/bioc/html/edgeR.html>, (Robinson et al., 2010))) and fragments per kilobase million (FPKM) values were calculated in edgeR, to reduce bias from gene length and sequencing depth between samples. Genes with less than 4 FPKM in at least 3 cells were filtered out as these would not influence the biological relevance of findings. In view of the high numbers of genes in which no expression was observed in many samples, counts were then re-normalised with the reduced gene set, so that the TMM normalisation could be more effective for single cell data. Expression values were then converted to transcripts per kilobase million (TPM) by dividing each FPKM value by the sum of FPKM values in that sample, and multiplying by 10^6 . A further gene filtering step was included to remove highly expressed genes with low variability across all samples (cells in the top decile for mean expression and below the fifth centile for coefficient of variation), again to exclude genes of limited biological relevance or interest.

2.5.2 Analysis of processed RNA-seq data

SC3 (version 1.4.2, <http://bioconductor.org/packages/release/bioc/html/SC3.html>, (Kiselev et al., 2017)) was used to provide cell cluster information by inputting gene

expression data in TPM format. Cluster robustness to experimental technical variation was tested using BEARscc, which models technical noise from ERCC spike-in measurements (version 1.2, <https://github.com/seversond12/BEARscc>, (Severson et al., 2018)). Cluster number, k , was chosen manually using the distribution of cluster-wise mean silhouette widths across clusters in all 250 simulated technical replicates for each cluster number k (2 to 8 for individual tissue and 1 to 15 for all tissues). t-SNE data were generated using the Barnes-Hut implementation of t-SNE in R (version 0.13, (Van Der Maaten, 2014)). Differential expression analysis was carried out between cell groups using edgeR (see above for version) from normalized counts according to the package manual. P values used were determined by permutation test at 5% (250-1000 permutations) to allow for multiple comparisons or, in cases of heavily unbalanced sample numbers, converted to false discovery rates (FDR or q values) by the Benjamini-Hochberg procedure as implemented in edgeR. Pathway analysis was performed using goseq to identify over or under represented ontological terms to an FDR of less than 0.05 (version 1.28, <https://bioconductor.org/packages/release/bioc/html/goseq.html>, (Young et al., 2010)). Identification of stem-like cells was performed using RaceID2 and StemID, please see <https://github.com/dgrun/StemID> for more details (Grün et al., 2015, 2016). Where gene expression is described in binary terms, the threshold was set to include or exclude 90% of cells with the highest expression of a given gene, to allow for biological noise.

2.5.3 Whole genome sequencing

The whole genome for K562 cells was obtained from the Sequencing Read Archive (from the human ENCODE project, accession number SRX285595). For this data, sequencing libraries were prepared from DNA using the Illumina[®] TruSeq DNA Preparation Kit (v5) according to manufacturer’s protocol. Single end (50bp) sequencing was performed using an Illumina[®] HiSeq 2000. Raw data were mapped with BWA (version 0.7.11, <http://bio-bwa.sourceforge.net>, (Li and Durbin, 2009)) to the hg19 version of the human genome (as used in RNA-seq experiments) and duplicates marked with Picard (version as above) before SNV calling (see below).

2.5.4 Single nucleotide variant calling

Variant calling was performed using the following software:

SAMtools (version 1.1, <http://samtools.sourceforge.net/>, (Li, 2011))

GeMs (version 1.0, <https://github.com/cui-lab/gems>, (Murillo et al., 2016))

freebayes (version 1.1.0, <https://github.com/ekg/freebayes>, (Garrison E, 2012))

VarScan2 (version 2.3.8, <http://varscan.sourceforge.net/>, (Koboldt et al., 2012))

HaplotypeCaller (following GATK best practices up to the variant filtering stage, (Poplin et al., 2017; Van der Auwera et al., 2013))

Default stringencies for each package were used in calls. Classifiers were called for each variant using nuseq (available at bitbucket.org/arothon85/nuseq), a python based package derived from Ding *et al* (Ding et al., 2012). Further classifiers were computed by normalising nuseq output where possible, then filtered by excluding classifiers where Pearson correlation coefficient $> \pm 0.75$. All variant calls were de-duplicated across all samples analysed, to avoid over-inflating the effects of classifiers when modelling variant predictions. For details of R packages used in `snv.predict`, please see Chapter 4 and Appendix D. Mutational signatures were generated from single nucleotide variants called with VarScan2 and filtered with `snv.predict` using SomaticSignatures (version 2.12.1, <http://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html>, (Gehring et al., 2015)). Hierarchical analysis of mutations was performed by creating a Euclidean distance matrix and using complete-linkage clustering in base R.

2.6 Tissue staining

Sections for pathological diagnosis by Professor Rob Goldin (Professor of pathology, Centre for Pathology, St Mary's Hospital, Imperial College, London, UK) and Dr Lai Mun Wang (Consultant pathologist, Department of Pathology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK) were haematoxylin and eosin (H&E) stained. 4 μ m (paraffin embedded) or 20 μ m (cryopreserved) sections were prepared. Samples were then dewaxed (where necessary) and rehydrated prior to H&E staining, fixing and mounting. Mouse tissues for staining were obtained from previous experiments carried out in the laboratory and approved by the University of Oxford's ethical review committee and licensed by the UK Home Office (license number PPL

30/2862) (Tordella et al., 2013). Oesophageal samples were obtained from the Oxford Radcliffe and Translational Gastroenterology Unit biobanks.

For immunostaining, sections were de-waxed, rehydrated and incubated with 3% hydrogen peroxide in methanol to block endogenous peroxidase activity (10 minutes, room temperature). Antigen retrieval was carried out using 10mM sodium citrate, pH6 at 100°C for 10 minutes. Sections were then blocked with normal goat serum (at room temperature) and incubated overnight at 4°C with a primary antibody against anti-p63 (1:1000, mouse monoclonal 4A4 cat. sc-8431, Santa Cruz (Glickman et al., 2001)), anti-KRT14 (IF, 1:1000, rabbit polyclonal cat. 905301, BioLegend[®]), anti-KRT8 (1:200, mouse monoclonal C51 cat. sc-8020, Santa-Cruz (Karthaus et al., 2014)), anti-LEFTY1 (IHC, 1:1000, rabbit polyclonal D7E3G cat. 12647, Cell Signalling), anti-SPINK4 (IF, 1:500, rabbit polyclonal cat. HPA007286, Sigma-Aldrich[®]) (Wapenaar et al., 2007) or anti-OLFM4 (IHC, 1:200, D1E4M rabbit monoclonal cat. 14369, Cell Signalling Technology[®]).

For immunohistochemical staining, samples were then treated with biotinylated secondary antibody (Vector Labs; 1:250) for 40 minutes at room temperature. The staining reaction was completed using the Vector Elite ABC kit, counterstained with haematoxylin, and imaged using a histology microscope. Protocols and antibody concentrations for immunohistochemical staining were optimised by Dr Michael White and final processing was performed using an auto-stainer (Leica Bond III).

For immunofluorescence staining, expression was detected using Alexa Fluor (1:250, Molecular Probes) for one hour. DAPI (1:2000, Sigma-Aldrich[®]) was used to stain nucleic acids during the secondary antibody incubation. Samples were observed using confocal microscopy (LSM 710, Zeiss). In all images presented, z-stack imaging was undertaken and maximum intensity images generated for use (ZEN, Zeiss).

2.7 Cell preservatives for single cell RNA-seq

RNA^{later} (QIAGEN) was used as per the manufacturer’s instructions. For slow freezing, tissue fragments were placed into a cryovial with 500 μ l of freezing solution (90% FBS/10% DMSO). Vials were then placed into an isopropanol freezing jar (Mr. Frosty[™] Freezing Container, Thermo Scientific[™]) at -80°C . For use in scRNA-seq, fragments were thawed and then processed as described above. For Dithiobis succinimidyl propionate (DSP) preservation a stock solution of 50x DSP was prepared

(50mg/ml in 100% anhydrous DMSO, stored at -80°C). This was thawed to room temperature when needed and diluted to 1x with PBS (following published methodology, to avoid precipitation) and approximately 10 volumes of DSP added to tissue fragments. For use, tissue fragments were disaggregated as described above and the RNA crosslinking reversed with DTT at 50mM final concentration and incubated at 37°C for 30 minutes. Full methodological details are available elsewhere (Attar et al., 2018).

2.8 Cell culture and staining

OE21 and OE33 cells were kindly provided by Dr Michael White, from the Underwood lab, University of Southampton, and mycoplasma screened. Cells were cultured in RPMI supplemented with L-glutamine, penicillin/streptomycin, 10% (vol/vol) fetal bovine serum at 37°C with 5% CO_2 . For staining, cell culture was carried out on coverslips in multi well plates. Once exposure conditions had taken place for each experiment, cells were fixed in 4% paraformaldehyde (PFA) for 15minutes then washed x3 with PBS. Next, cells were permeabilized with 0.5% Triton X-100 for 10min at 4°C then blocked with 5% normal goat serum in PBS, and subsequently incubated with appropriate primary antibody at room temperature. After further washes in PBS, the appropriate secondary antibodies (1:250) and DAPI (1:2000, Sigma-Aldrich®) were applied to cells for one hour. Images were taken as for tissue immunofluorescence (described above).

2.9 Immunoblotting

Cells were harvested in PBS, washed twice in PBS then lysed for 20 minutes on ice in RIPA buffer, before centrifuging at 20000xg for 30 minutes at 4°C . The supernatant was collected into a fresh tube as the soluble protein fraction. The pellet was then dissolved in x5 volumes of urea buffer, incubated at room temperature for 20 minutes, then centrifuged at 20000g for 20 minutes before collecting supernatant as insoluble protein fraction. Protein concentrations were estimated using Bradford assay and normalised using loading buffer. Normalised protein extracts were loaded onto 8% SDS-polyacrylamide gels (20ml: 9.3ml water, 5.3ml 30% acrylamide mix, 5.0ml 1.5M Tris (pH 8.8), 0.2ml 10% SDS, 0.2ml ammonium persulfate and 0.012ml TEMED) with a prestained protein ladder (iBright Prestained Protein

Ladder, ThermoFisher). Gels were transferred onto nitrocellulose membranes and then incubated with primary antibody overnight at 4°C, then with the appropriate secondary HRP-conjugated antibody (1:2000, DakoCytomation). The results were visualized by enhanced chemiluminescence detection (Amersham Biosciences) using X-ray film (Fujifilm). Primary antibodies used were anti-SOX2 (1:200, mouse monoclonal E-4 cat. sc-365823, Santa Cruz (Luo et al., 2013)), anti-CDX2 (1:200, mouse monoclonal E-1 cat. sc-393572, Santa Cruz), anti-p63 (1:1000, mouse monoclonal 4A4 cat. sc-8431, Santa Cruz (Glickman et al., 2001)), anti-KRT14 (1:2000, rabbit polyclonal cat. 905301, BioLegend®), anti-OCT3/4 (1:200, mouse monoclonal C-10 cat. sc-5279, Santa-Cruz (Duan et al., 2015)), anti-ASPP2 (1:100, mouse monoclonal DX54.10, from Lu lab, University of Oxford (Sottocornola et al., 2010)), anti-actin, 1:2000, mouse monoclonal cat. ab8226, Abcam).

2.10 Cell co-culture

Each cell line (OE21 and OE33) was labelled with lipophilic dyes (Vybrant™ Multicolor Cell-Labeling Kit, ThermoFisher) as per the product manufacturer’s instructions. Cells were quantified with an automated cell counter (TC-20™, Bio-Rad). Cells were then plated in a 50:50 ratio, in incremental amounts in different wells (from 1×10^4 - 1×10^5 cells per well), with replicates, in a multiwell plate. Cell culture was in an environment controlled timelapse fluorescence microscope, using a TE-2000e microscope (Nikon) and Orca ER camera (Hamamatsu Photonics). Auto-focussing images were taken at fixed points in bright-field, RFP and GFP spectra every 30-40 minutes, depending on number of imaging points. Metamorph (Molecular Devices) and imageJ (version 1.47, <https://imagej.nih.gov/ij/index.html>, (Schneider et al., 2012)) were used to manage imaging and generate video output.

2.11 Organoid culture

The stomachs of ASPP2 Δ 3loxP-CreER mice were used from an existing study in the laboratory (PPL 30/2862), organoid culture was carried out as previously described in the Supplemental Experimental Procedures section of Barker *et al* (Barker et al., 2010). Time lapse microscopy was performed as described above using only bright-field imaging, and imageJ used to make videos. Further data from images was generated using a macro written by Dr Mark Shipman in imageJ.

2.12 Data presentation

All plots are made in R (version 3.4.1) using ggplot2 (version 2.2.1, (Wickham, 2016)). Where box plots are used, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles), the whiskers extend from the hinge to the largest or smallest values at most 1.5 x inter-quartile range (IQR) from the hinge. Data beyond the whiskers are outliers and are plotted individually. Where bracketed p values are included in box plot data, this is obtained from a Wilcoxon signed-rank test and annotated as follows: $p < 0.05 = *$; $p < 0.01 = **$; $p < 0.001 = ***$; not significant = NS. In cases of high data ranges, frequently seen in scRNA-seq due to gene drop-out effects, jitter plots are used with random vertical and horizontal noise injected to demonstrate data density at a given point. This noise is adjusted manually to avoid distorting data interpretation. Smooth curves in scatter plots are Loess curves generated using ggplot2.

2.13 Buffers

Digestion Solution:

1x Phosphate buffered solution (GibcoTM)
2mM EDTA, 100U/ml type I collagenase (Worthington Biochemical Company[®])
Sodium phosphate (5.6mM)
Monopotassium phosphate (8mM)
Sodium chloride (96mM)
Potassium chloride (1.6mM)
Sucrose (44mM)
D-Sorbitol (55mM)
DL-Dithiotreitol (0.5mM))

Sorting buffer:

1x phosphate buffered solution (GibcoTM)
2mM EDTA
5% heat inactivated fetal bovine serum (Sigma-Aldrich[®])

RIPA buffer:

10 mM Tris.Cl (pH 8.0)

140 mM NaCl

1 mM EDTA

0.5 mM EGTA

1% Triton X-100

0.1% sodium deoxycholate

0.1% SDS

Complete Protease Inhibitor Mixture (Roche))

Urea buffer:

8M Urea

40mM NaCl

50mM Tris

2mM MgCl₂

50mM NaF

50mM b-glyceradelyde phosphate

1mM sodium orthovanadate

10mM sodium pyrophosphate

Complete Protease Inhibitor Mixture (Roche))

Chapter 3

Barrett's oesophagus and upper gastrointestinal cell populations

The origins of Barrett's oesophagus have been elusive since pathological columnar epithelium was first described in the oesophagus in 1950 by Norman Barrett, CBE. In this chapter I apply the single cell RNA-seq to normal and metaplastic tissues in the upper gastrointestinal tract to characterise normal and abnormal cell populations and investigate the phenotypic overlaps seen in normal tissues and in Barrett's oesophagus.

3.1 Single cell RNA-seq identifies subpopulations in normal upper gastrointestinal epithelia

To characterise the cell populations in BO, samples were taken from 13 BO patients (A-D, I-Q) attending for routine endoscopic surveillance of non-dysplastic BO. From each patient, biopsies were taken from BO, adjacent macroscopically normal oesophagus (20mm proximal to BO), stomach (20mm distal to the gastro-oesophageal junction) and duodenum (as described in Materials and Methods).

Individual 2mm biopsies were divided to provide tissue for single cell RNA-seq, bulk tissue RNA-seq and histology in 4 out of 13 patients, and bulk tissue RNA-seq and histology alone in the remaining 9 patients (see Materials and Methods). Single cells and histology were also prepared from normal oesophageal biopsies from two patients with gastro-oesophageal reflux disease but no previous or current diagnosis of BO (E-F) and a further two patients with BO (G-H) using an adjusted sampling process to include gastric pylorus instead of duodenum, as it became apparent later that further regional gastric controls would strengthen my conclusions. All sampled patients were taking regular acid suppression therapy and had no features of oesophageal dysplasia or malignancy (Figure 3.1 and Table 3.1, also shown in Methods).

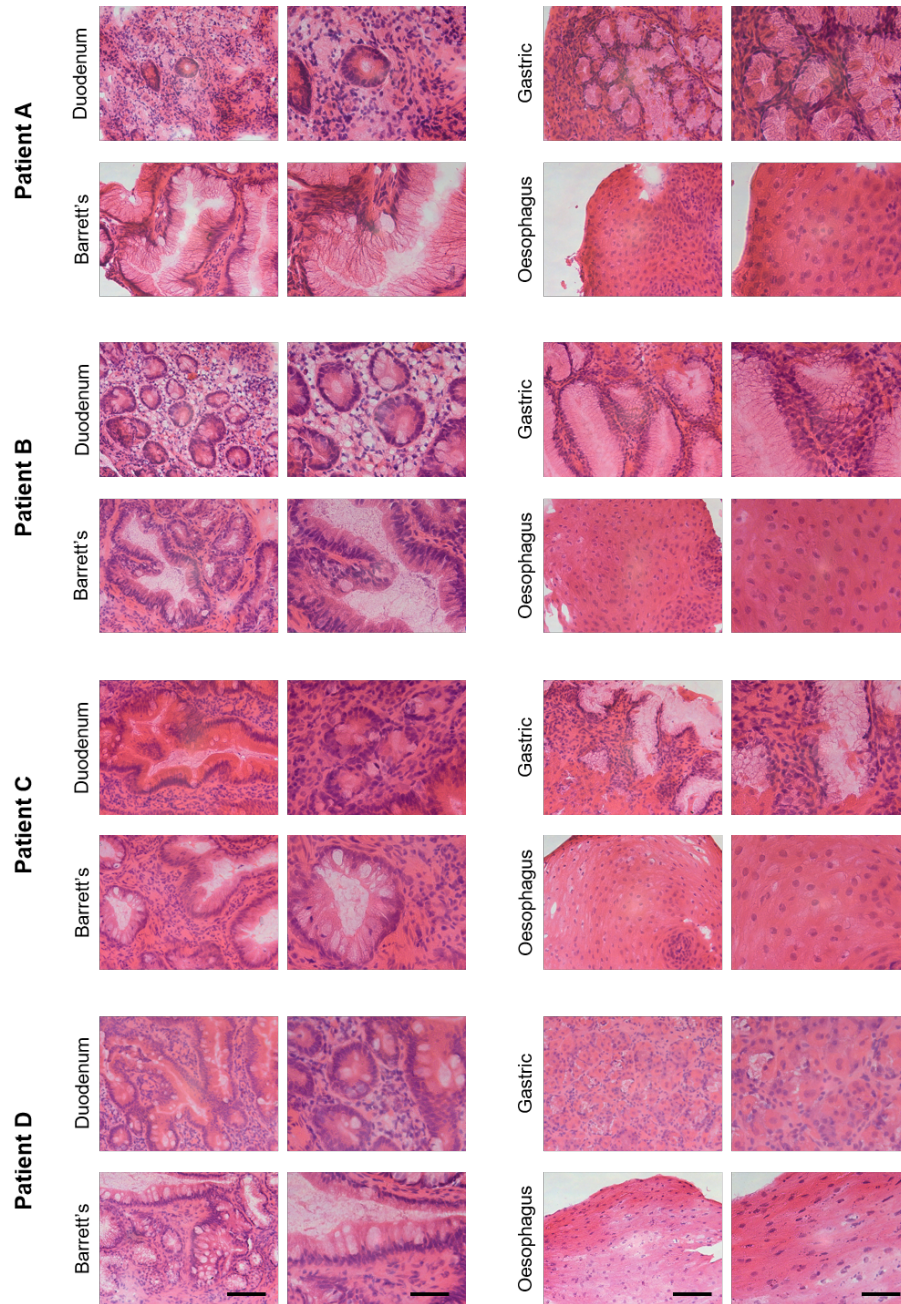


Figure 3.1: Slides from upper gastrointestinal tissues used for single cell RNA-seq. Haematoxylin and eosin staining of frozen sections taken from the biopsy fragments that were used for single cell and bulk RNA-sequencing from four patients (excluding the normal oesophageal samples taken from non-Barrett's oesophagus patients). Images on the right of each pair are a higher magnification of the image on the left. Each scale bar (on lower-most images) applies to the whole column; $100\mu\text{m}$ in first and third columns, and $50\mu\text{m}$ for the second and fourth columns.

Bulk RNA-sequencing followed by hierarchical clustering of differentially expressed genes in the duodenal, gastric, oesophageal and BO samples from 13 patients with

Patient	Age	Sex	Smoking (cigs/day)	Alcohol (units/week)	BMI	Prague	scRNA-seq	Bulk RNA-seq
A	78	M	0	21	26.2	C3M5	Yes	Yes
B	77	M	0	4	25.7	C0M3	Yes	Yes
C	47	F	0	16	22.2	C0M2	Yes	Yes
D	90	F	0	6	23.7	C9M11	Yes	Yes
E	82	F	0	0	27.1	na	Yes	No
F	68	F	0	0	25.8	na	Yes	No
G	76	M	10	8	27.5	C1M3	Yes	No
H	78	M	0	22	29.4	C2M4	Yes	No
I	87	M	ex 10/day	8	21.9	C8M10	No	Yes
J	75	F	0	2	35.7	C1M3	No	Yes
K	74	M	ex 1/day	24	25.5	C2M3	No	Yes
L	63	M	ex 15/day	10	22.8	C0M3	No	Yes
M	70	F	0	4	32.6	C5M7	No	Yes
N	59	M	0	15	41.0	C0M1	No	Yes
O	67	F	0	0	31.7	C13M15	No	Yes
P	65	F	30/day	3	27.3	C0M1	No	Yes
Q	49	M	20/day	0	42.9	C3M4	No	Yes

Table 3.1: Patient metadata summary. Details of patients’ samples having single cell and bulk RNA-sequencing. BMI, body mass index; PPI, protein pump inhibitor treatment. Prague indicates Prague classification for measuring the length of Barrett’s oesophagus, where C is proximal extent of circumferential lesion and M is maximal proximal extent (na for patients without BO). scRNA-seq and Bulk RNA-seq indicate whether samples from each patient were used for single cell or bulk RNA-seq, respectively, total n=17.

BO showed a clear distinction between squamous (i.e. normal oesophagus) and non-squamous (i.e. gastric, duodenum and BO) epithelia (Figure 3.2A). BO samples from all 13 patients had some similarities to duodenal and gastric samples. When a defined list of genes known to distinguish gastrointestinal epithelia (Lavery et al., 2014) was used in hierarchical clustering, BO samples appeared most closely related to gastric tissue, consistent with previous studies (McDonald et al., 2015) (Figure 3.2B). Samples from patients E-H only had single cell RNA-seq from fewer or different sampling sites, and thus paired bulk RNA-seq was not done explaining why these patients are not included in these heatmaps and why the patient labelling is not sequential in this figure.

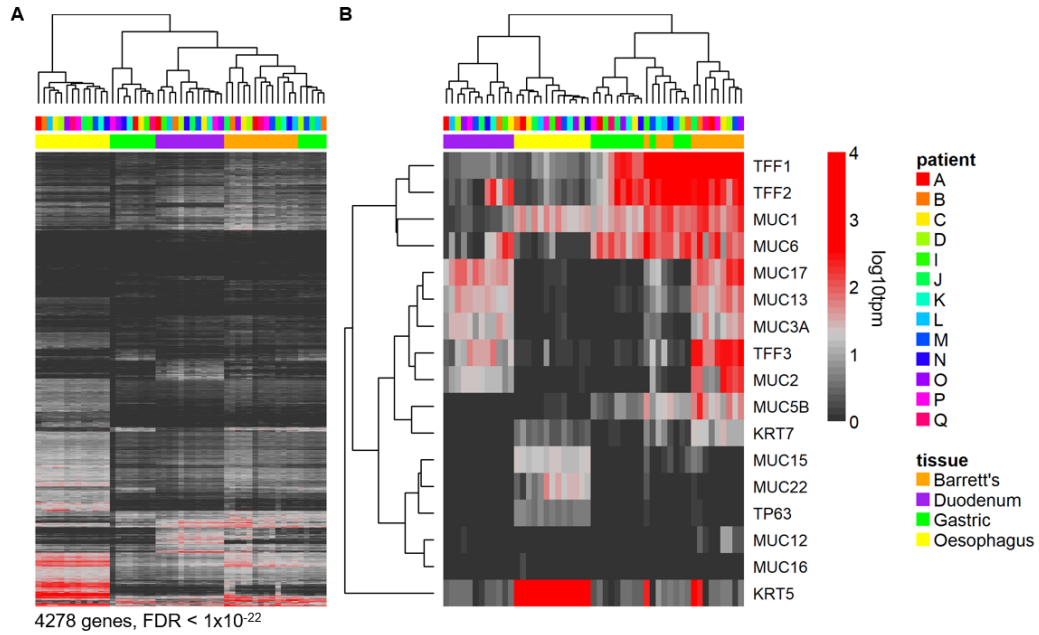


Figure 3.2: Hierarchical clustering of upper gastrointestinal samples from bulk RNA-seq. (A) From bulk RNA-seq data derived from samples from 13 patients with BO, heatmap of genes differentially expressed between any tissue type (FDR < 1e-22) with tissue hierarchy determined by nearest neighbour. Tissue indicated by colours as in Figure 2.1. One duodenal sample from patient Q failed to produce usable data and was excluded hence total n=51. As previously described, samples from patients E-H were not used for bulk RNA-seq. (B) From bulk RNA-seq data, heatmap of expression of mucin and trefoil factor genes with tissue hierarchy determined by nearest neighbour, in samples from 13 patients with BO. Gene expression units are log10 transcripts per million (TPM).

For single cell RNA-seq, a total of 4237 cells were sequenced from 8 patients (A-H) in three batches. Due to known issues with batch effects in single cell experiments (Hicks et al., 2017), analysis of cells from each batch has been kept separate where feasible and cells were permuted across plates and pooled prior to sequencing (see Materials and Methods). The first batch yielded 1040 cells (207 duodenum, 227 gastric, 371 BO and 235 oesophagus) suitable for analysis from four patients (A-D) with BO and intestinal metaplasia. A total of 214, 35, 66 and 56 BO cells were analysed from each BO patient, respectively. The second batch yielded 648 oesophagus cells suitable for analysis from two patients (E-F) with symptoms of gastro-oesophageal reflux but no identifiable oesophageal pathology. Finally, the third batch of cells yielded 194 cells (29 pylorus, 109 gastric, 32 BO and 24 oesophagus) suitable for analysis from two patients (G-H) with BO and intestinal metaplasia (Figure 3.3A). Overall, there was a mean of 1.2×10^5 counts per cell and a median of 3978 genes were detected

per cell, with at least one count per gene (Figure 3.3B, C). Summed gene counts in an ensemble of BO cells from patient A showed high correlation with matched bulk RNA-seq gene counts (Figure 3.3D).

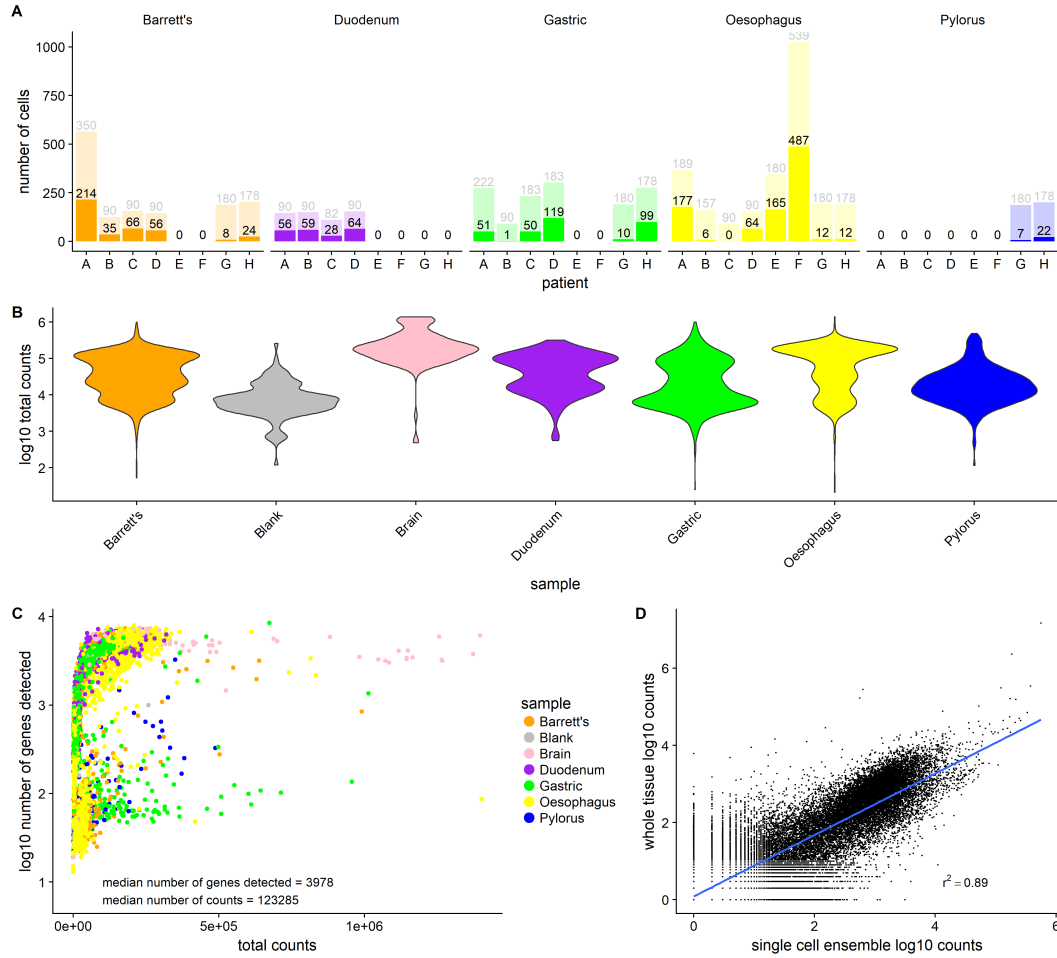


Figure 3.3: Data quality from single cell RNA-seq experiments. (A) Bar plot showing total number of cells sequenced of each tissue type by patient (A-D and G-H are patients with BO, E-F are patients without BO). Cells passing threshold for inclusion determined by controls (see Materials and Methods) are in solid colour, excluded cells are shown in faded colour. (B) Violin plots showing number of counts by each tissue type, coloured as in A, in all cells which were sequenced. (C) Plot of number of genes detected (at least one count per gene) against total counts for every sequenced cell and control. (D) Scatter plot showing the correlation between bulk tissue RNA-seq gene counts and an ensemble of single cell gene counts, taken from Barrett's samples from patient A, with Pearson correlation coefficient displayed.

To analyse cell phenotypes, cells were clustered from each normal tissue type from the BO patients by gene expression (Figure 3.4A). The eleven clusters (D1-D4, G1-G3 and O1-O4, in duodenum, gastric and oesophageal samples, respectively) were

then annotated on the basis of differentially expressed genes previously characterized in specific cell types (Figure 3.4B, a complete list of genes differentially expressed between clusters is available in Appendix A.1 to A.11). In the duodenum, these are: alkaline phosphatase, intestinal (*ALPI*)-expressing enterocytes (D1); mucin 2 (*MUC2*)-expressing goblet cells (D2); olfactomedin 4 (*OLFM4*)-expressing crypt cells (D3); and some uncharacterized cells expressing Joining Chain Of Multimeric IgA And IgM (*JCHAIN*) (D4). In the gastric samples, these are: chromogranin (*CHGA*)-expressing enteroendocrine cells (G1); gastrokinin (*GKN1*)- and trefoil factor 1 (*TFF1*)-expressing foveolar cells (G2); and mucin 6 (*MUC6*)- and *TFF1*-expressing mucus neck cells (G3).

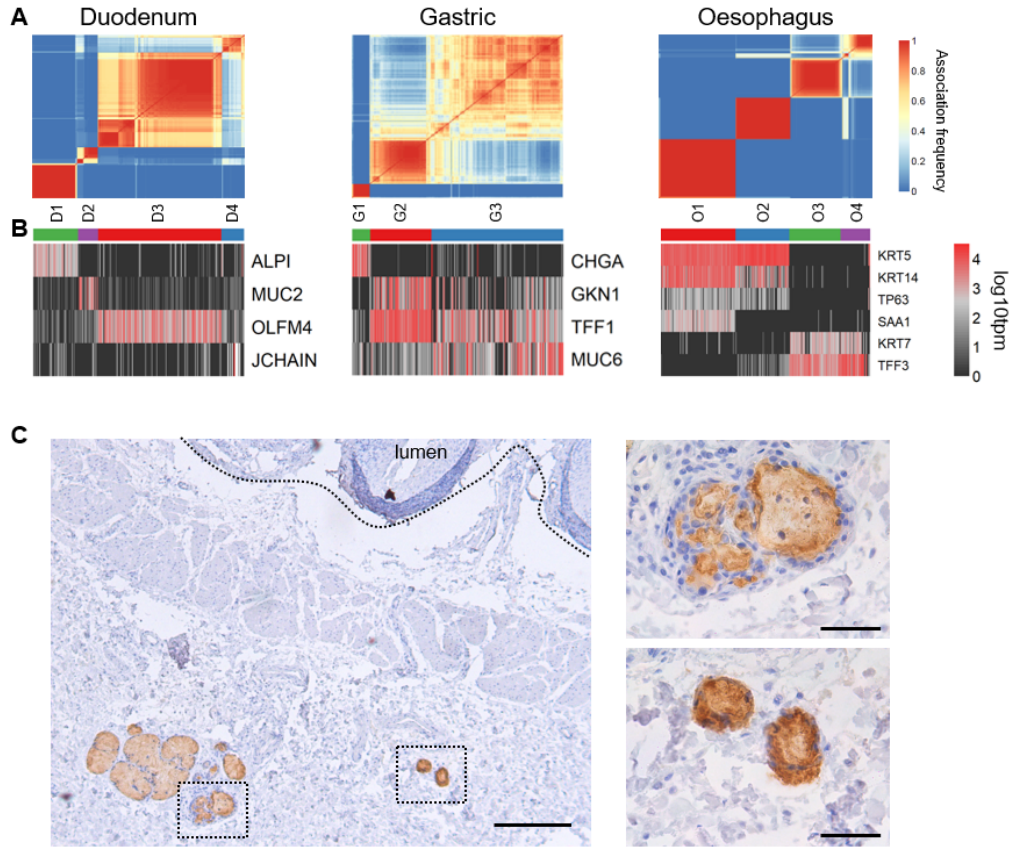


Figure 3.4: Upper gastrointestinal cell clustering by gene expression and cluster-specific gene identification. (A) Panels show the cluster consensus matrices for single cells from normal tissue sites in four BO patients. Blue-to-red colours denote the frequency with which cells are grouped together in 250 repeat clusterings of simulated technical replicates (see Materials and Methods). Cell clusters are indicated by coloured bars below the matrices. In lower panels, heatmaps show expression of known functionally relevant genes that were differentially expressed between cell clusters, corresponding to matrices above in A (>4 fold change, $FDR < 1e-5$). (C) Immunohistochemical staining of TFF3 in normal oesophagus from a resection specimen. Black dotted curve outlines deep to the basal layer of the stratified squamous epithelium, above which is the lumen as labelled. Expanded images are of OSG (top) and OSG duct (bottom), which is columnar lined proximally. Scale bars are $250\mu m$, and $50\mu m$ in expanded images.

Interestingly, four clusters were identified in the oesophageal samples. Two of these express expected squamous genes (*KRT5*, *KRT14*, *TP63*; clusters O1 and O2) and two express the columnar gene *TFF3* (clusters O3 and O4). The two squamous clusters can be distinguished by the presence (O1) or absence (O2) of acute phase response (*SAA1*) gene expression, presumably representing squamous cells in different states. The detection of *TFF3* in O3 and O4 is of great interest and is consistent with

these cells being from the columnar epithelium of oesophageal submucosal glands (OSGs) (Long and Orlando, 1999), a structure in the normal human oesophagus. To confirm this, *TFF3* stained positively in OSGs and OSG ducts in sections of normal oesophagus obtained from an OAC resection specimen (Figure 3.4C). The high number of OSG cells is unexpected, but since these cells are involved in oesophageal mucosal defence, they are likely to be more resilient to the processing conditions for single cell RNA-seq and would therefore be enriched in downstream analysis.

To confirm the association between columnar tissue and *TFF3*, keratin 7 (*KRT7*), a columnar gene used in the detection of BO, commonly co-expressed with *TFF3*, and rarely with the squamous gene *KRT5* (Figure 3.5A, top plot). Staining for another columnar gene, keratin 8 (*KRT8*), shows a distinct squamous-columnar divide in wild type murine squamo-columnar junction (Figure 3.5B). *KRT8* expression was generally low in these scRNA-seq data, but co-expression with *KRT7* can be observed, with 35.9% of oesophageal cells expressing both columnar markers ($\log_{10}\text{tpm} \geq 0.1$) (Figure 3.5A, bottom plot). *TP63* expression, while also generally low in these data, was negatively associated with *KRT7* expression as expected, though interestingly some *KRT8* cells were *TP63* positive, despite p63 staining showing a sharp demarcation of the SCJ in mice and being closely restricted to squamous type columnar cells (Figure 3.5B, bottom plot and Figure 3.4B, right sided heatmap).

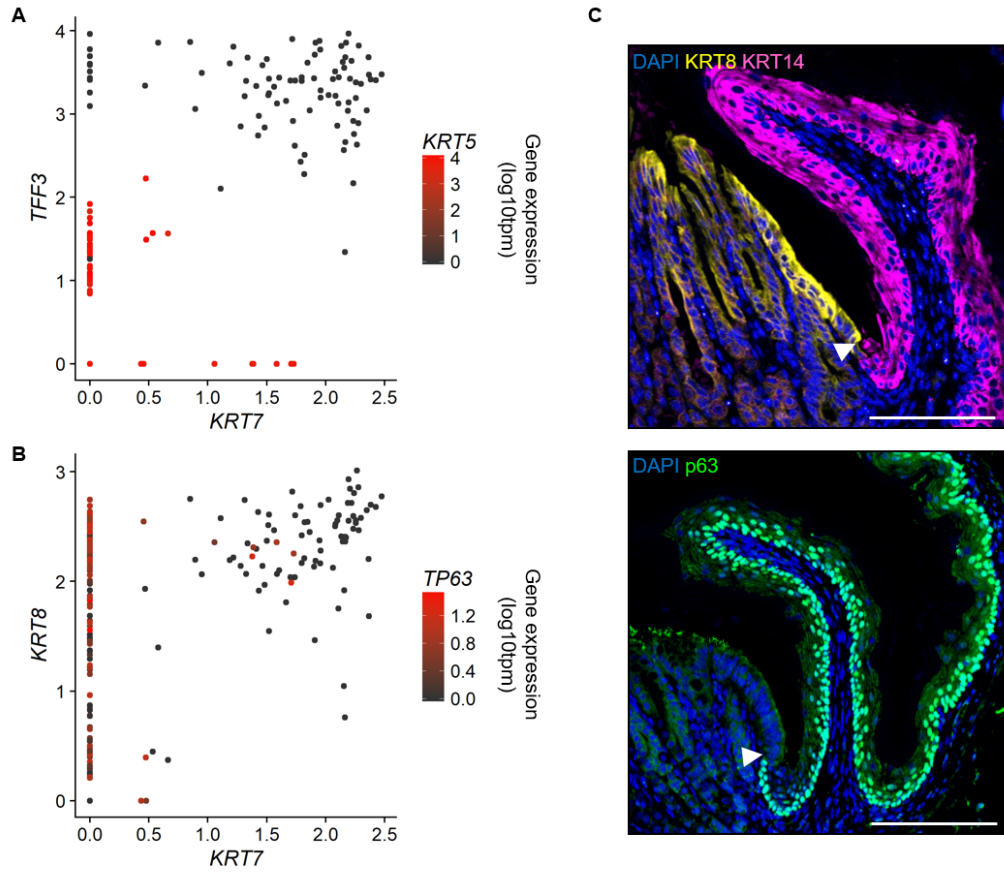


Figure 3.5: Expression of columnar and squamous markers in oesophageal cells. (A) Scatter plot of gene expression of *KRT7*, *TFF3* and *KRT5* in oesophageal cells from four patients with BO. n=248 (different to n in Figure 3.4 due to changes in gene filtering to include *KRT8*). (B) As in A, scatter plot of gene expression of *KRT7*, *TFF3* and *KRT5*. (C) Immunofluorescence staining of sections of mouse stomach including the SCJ. The top image is co-stained with KRT8 (yellow), KRT14 (pink) and DAPI (blue) and the bottom image is stained with p63 (green) and DAPI (blue). Scale bars are 250 μm. White arrow heads indicate SCJ positions.

Of note, the proton pump gene *ATP4A* and the intrinsic factor gene *GIF* were rarely detectable in gastric cells, indicating these are cardiac-type gastric samples (Figure 3.6, *ATP4A* and *GIF* top left, other genes included for general interest).

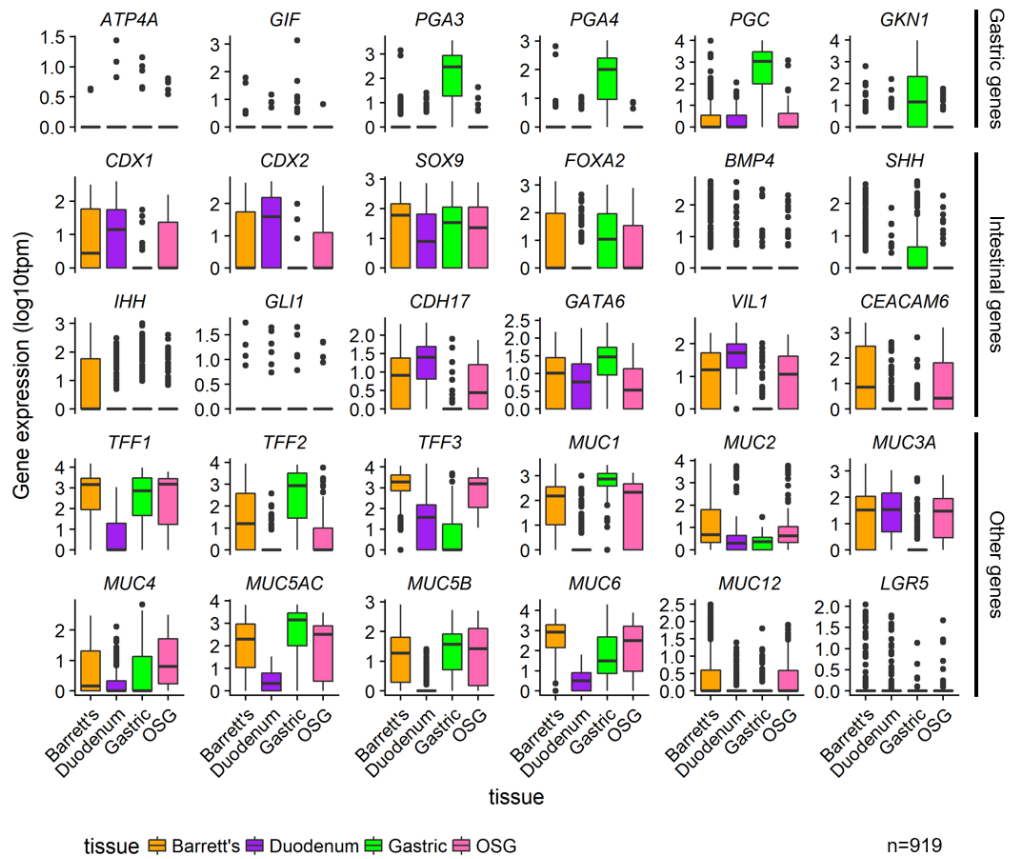


Figure 3.6: Expression of known marker genes in upper gastrointestinal cells. Boxplots showing the expression of a selection of genes used to define gastric or intestinal tissue cell types (first row of plots are gastric genes, second and third rows are intestinal genes, as indicated), and a panel of mucin and trefoil factor genes (fourth and fifth rows of plots). Cells included are from all duodenum (n=207), gastric (n=221) and Barrett's samples (n=371), columnar type oesophageal cells are also included as 'OSGs' (n=120, *TFF3* positive oesophageal cells). Total n=919.

3.2 Barrett's oesophagus is enriched for LEFTY1-expressing cells

To identify genes characteristic of distinct BO cell populations, all BO cells were clustered by gene expression (Figure 3.7A), also see Appendix A.12 to A.15 for complete list of cluster specific genes and Appendix B.2 for upregulated genes in 'Barrett's-type' cells, introduced in the next section).

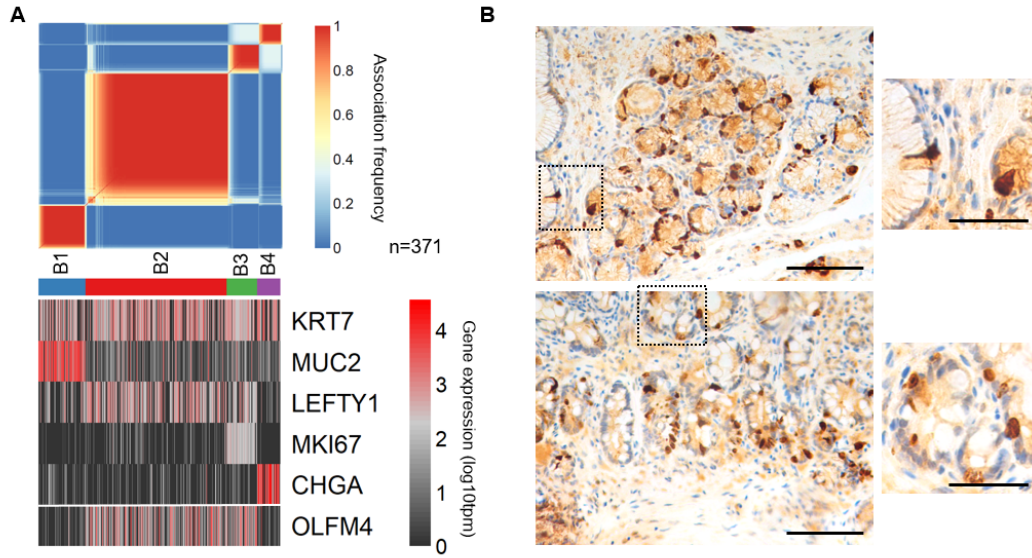


Figure 3.7: Barrett's oesophagus cell clustering by gene expression and LEFTY1 expression. (A) Upper panel, cluster consensus matrix of BO cells from 4 BO patients (n=371 cells). Blue-to-red colours denote the frequency with which cells are grouped together in 250 repeat clusterings of simulated technical replicates (see Material and Methods). Clusters (B1-B4) are indicated by the coloured bars below. Lower panel, heatmap showing expression of selected functionally relevant genes that are differentially expressed between cell clusters (>4 fold change, FDR < 1e-5). (B) Immunohistochemical staining of LEFTY1 in gastric type human BO (upper image, expanded image shows gastric type mucus neck cell) and intestinalised human BO (lower image, expanded image shows section of BO gland with goblet cells negative for LEFTY1). Scale bars are 100 μm and 50 μm in expanded image.

The clusters (B1-B4) can be distinguished by expression of *MUC2* (B1; goblet cells, 19% of BO cells); *LEFTY1* (B2 and B3, 71% of BO cells); and *CHGA* (B4; enteroendocrine cells, 9.7% of BO cells). Since all patients had intestinal metaplasia, goblet cells made up 22%, 2.9%, 29% and 7.1% of cells in patient A-D, respectively. *KRT7* is expressed similarly across all 4 clusters, consistent with it being a marker of BO (Ormsby et al., 2000; Shearer et al., 2005). The *LEFTY1*-expressing cells (B2 and B3; Figure 3.7) are divided into a larger, low proliferating (*MKI67* (Ki67) negative) cluster (B2) and a smaller, high proliferating (*MKI67* positive) cluster (B3). *LEFTY1*, a secreted protein and transforming growth factor β (TGF β) family member, is normally expressed in development, where it has roles in left-right asymmetry determination (Hamada et al., 2002), but little is known about its potential roles in adult tissues and it has not previously been associated with BO.

Cells expressing *LEFTY1* were much less numerous in other sequenced upper gastrointestinal columnar type cells when compared to BO cells (Figure 3.8). Interest-

ingly, numerous oesophageal cells did express *LEFTY1*, suggesting that there may be a link between these columnar-gene expressing oesophageal cells and cells in BO that is not shared by the other sampled columnar tissues.

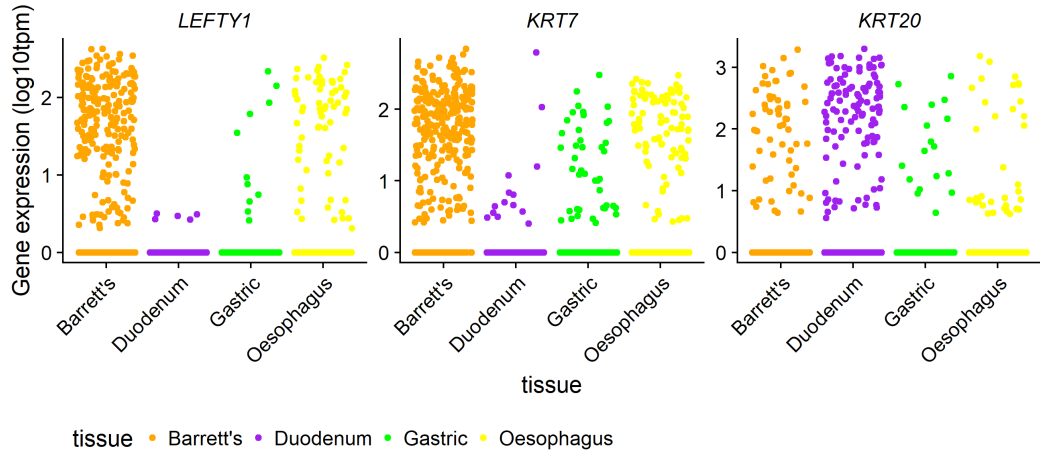


Figure 3.8: Expression of columnar markers and *LEFTY1* in upper gastrointestinal cells. Jitter plots showing the expression of *LEFTY1*, *KRT7* and *KRT20* in Barrett's, duodenum, gastric and oesophageal cells.

3.3 Oesophageal submucosal glands share an RNA composition profile with Barrett's oesophagus

To test whether some oesophageal cells did share gene expression with BO over the whole transcriptome, all cells from BO patients were clustered together (A-D). This shows that the normal tissue cells separate clearly from the BO cells based on their gene expression, with the exception of specialised cell types such as goblet or enteroendocrine cells. The majority of BO cells overlap with a sub-set of oesophageal cells, as seen when cells are clustered by gene expression (SC3 clustering with cluster stability assessed by BEARscc). Overall, cells were assigned to seven clusters (with brain controls in a separate cluster) (Figure 3.9). The annotation of clusters is done manually, based on differentially expressed genes and matching these with previous studies describing protein or gene upregulation in specific cell types.

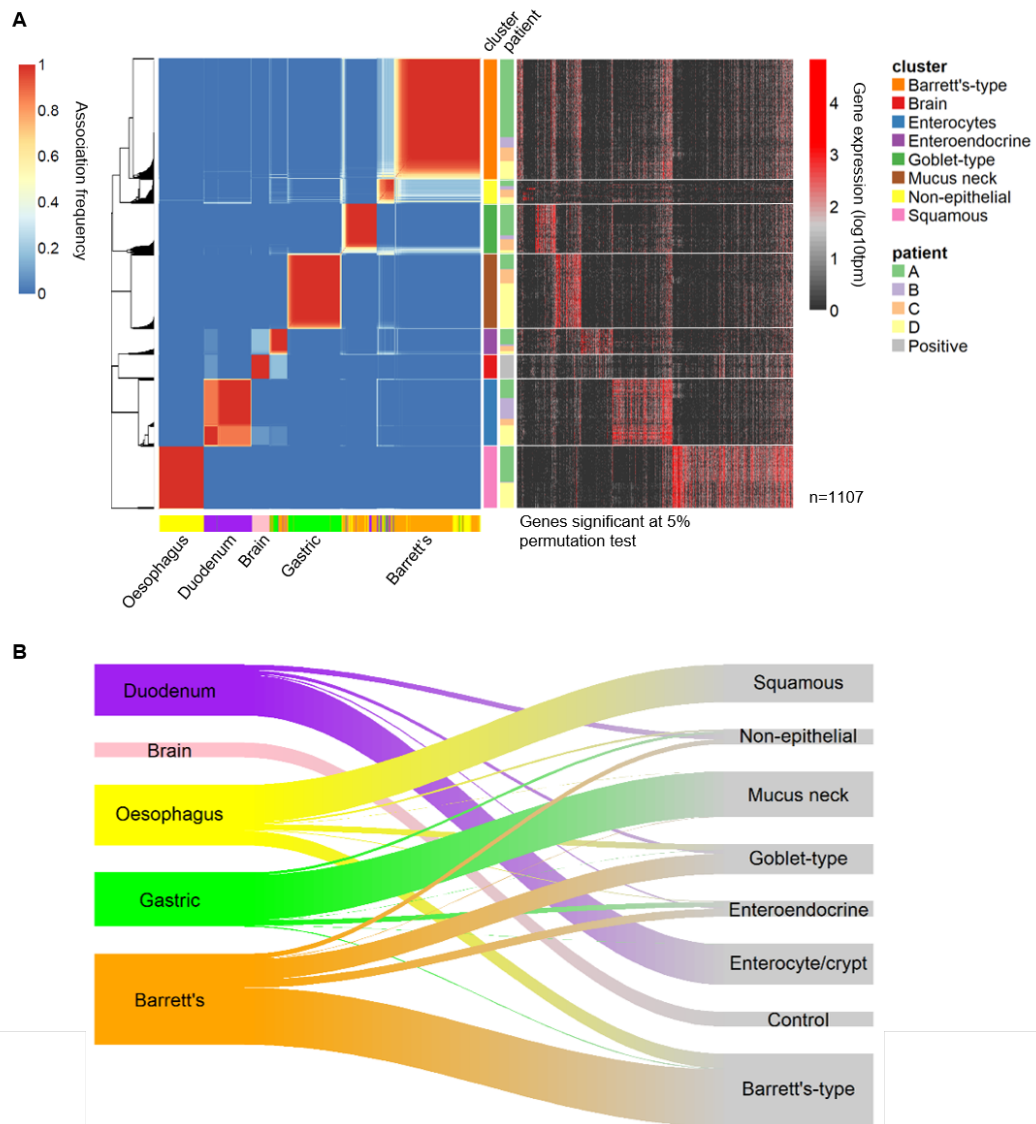


Figure 3.9: Clustering of all sampled cells from the upper gastrointestinal tract and Barrett's oesophagus. (A) The left heatmap shows a cluster consensus matrix for single cells from all tissue sites in BO patients (n=1107 including brain positive controls). Blue-to-red colours denote the frequency with which cells are grouped together in 250 repeat clusterings of simulated technical replicates. Tissue type is indicated below. Cell clusters are labelled on the right with the cell type they contain or a descriptive term if that cell type has not been previously characterised. The right side heatmap shows differentially expressed genes (>4 fold, genes significant at 5% permutation test) in each cluster from panel A. Cells from each patient are indicated by the coloured bar on the left. (B) River plot showing how each tissue type sampled contributes to the clusters shown in A. Colours and labels on the left indicate sampled tissue, as in A; labels on the right indicate cluster.

Most of these clusters are similar to those identified in the analysis of normal tissue alone (Figure 3.4) and agree with visual cluster representation using t-Distributed

Stochastic Neighbor Embedding (tsne) (Figure 3.10A, B). Clusters can be related to known cell types based on expression of previously characterised genes (Figure 3.10C), also see Appendix B for complete list of upregulated genes in each cluster).

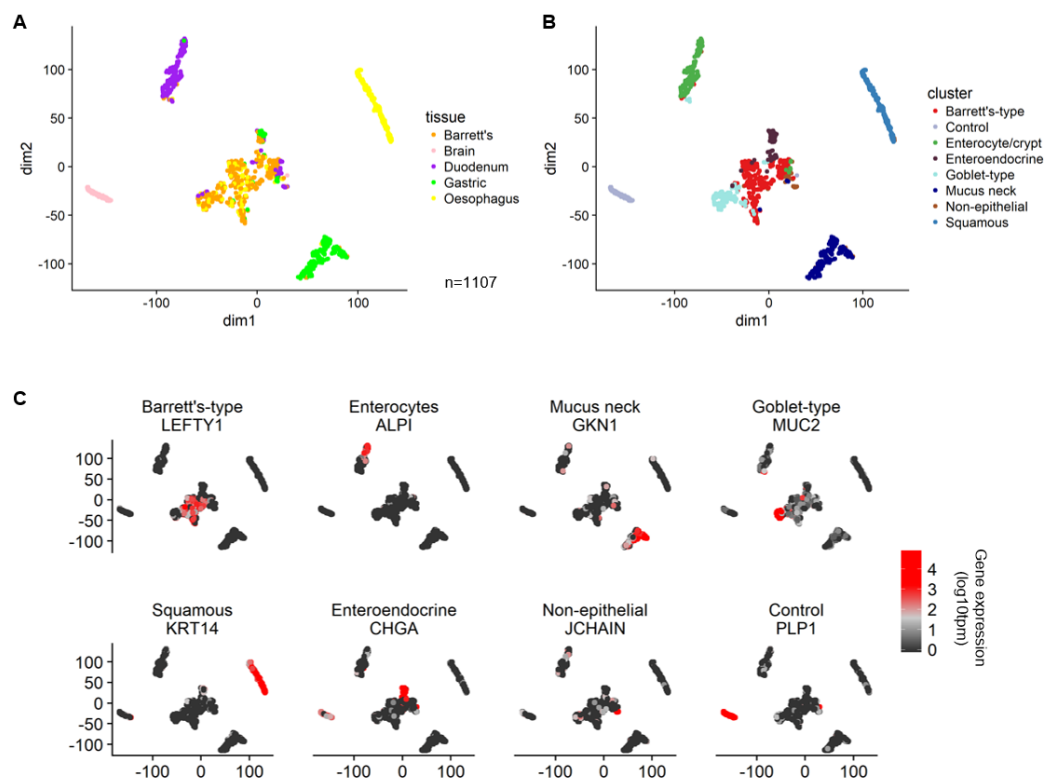


Figure 3.10: Plots of upper gastrointestinal tract cell clustering with marker gene expression in each cluster. (A) t-Distributed Stochastic Neighbour Embedding (tsne) plots of cells from all samples from four BO patients (n=1107 including brain control), showing similarity of cells in two dimensions, coloured by tissue type (yellow, oesophagus; green, gastric cardia; purple, duodenum; orange, Barrett's oesophagus; pink, brain). Brain was used as a positive control. (B) tsne plot of cells from four BO patient samples (A-D), as in A, coloured by how cells contribute to clusters generated by SC3 analysis with 250 repeat clusterings of simulated technical replicates (see Figure 3.9). Names given to the clusters are based on expression of known marker genes. (C) tsne plots of all cells from BO patients (n=1107 including positive controls), coloured in each panel by transcript level of a gene that is highly expressed in a particular cluster type.

The majority of duodenal cells fall in the cluster categorised as 'enterocytes' (similar to D1), gastric as 'mucus neck' (similar to G3), and a substantial proportion of oesophageal cells are in the 'squamous' cluster (similar to O1/O2) (Figure 3.9B). Some oesophageal cells, BO cells and a few duodenal cells fall into a 'goblet' cluster, and some gastric cells cluster with a few BO cells in the 'enteroendocrine' cluster.

The group described as 'non-epithelial' contains some endothelial cells and CD45-low immune cells (Figure 3.11).

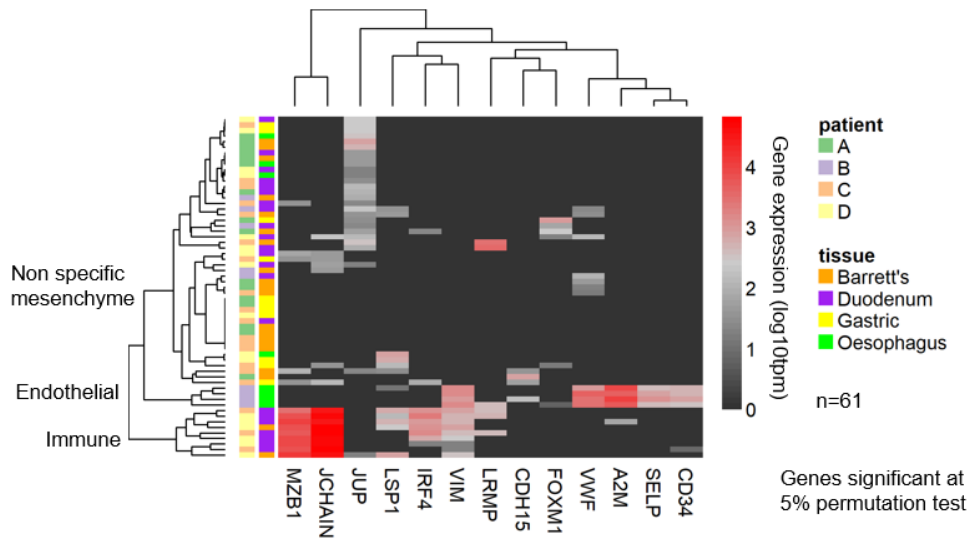


Figure 3.11: Further characterisation of 'non-epithelial' cells. Heatmap of gene expression in cells in the cluster labelled as 'non-epithelial', showing genes significantly upregulated in this cluster (>4 fold, genes significant at 5% permutation test). Genes are clustered by nearest neighbour and the dendrogram is labelled by broad cell-type category based on known functions of the expressed genes. For each cell, the tissue type and patient is indicated by the coloured bars on the left.

Notably, the majority of BO cells (63%) are in the cluster labelled as 'Barrett's-type' that also contains the subset of oesophageal cells that have a gene expression profile consistent with their being OSGs (Figure 3.9B, also see Appendix B.2 which is copied below for convenience in Table 3.2). These cells are enriched for *LEFTY1* expression.

Gene name	Log2 fold change	P value
CA1	5.041841	5.43E-47
LEFTY1	4.913848	2.44E-89
NMUR2	4.691247	3.04E-56
CXCL1	3.570928	1.80E-30
CEACAM6	3.445152	2.54E-37
NEB	3.351777	3.66E-32
MUC12	3.337496	3.92E-34
SOD3	3.240966	2.00E-30
GABRP	3.209838	1.66E-29
ELF5	3.083779	7.12E-24
GPRC5A	3.028671	1.65E-46
CEACAM5	2.987728	1.44E-29
MUC17	2.92779	6.02E-31
CT83	2.852599	3.70E-27
F5	2.834421	5.27E-26
CD55	2.704958	1.73E-32
TRIM31	2.646599	1.16E-22
HOXB6	2.498172	1.89E-23
KRT7	2.465271	3.70E-24
PHLDA2	2.322317	7.43E-32
TNFRSF11A	2.11044	3.88E-27
CD24P4	2.077556	2.51E-32

Table 3.2: Upregulated genes in the Barrett’s-type cells. Table showing genes which were upregulated in Barrett’s type cells from clustering all cells together from four patients with Barrett’s oesophagus. Genes shown had a log2 fold change of >2 and p values below a threshold determined from a 5% permutation test.

To test whether this relationship between BO and native oesophageal cells with columnar characterisation was also seen in patients without BO, all normal oesophageal cells from patients with and without BO were clustered (A, B, D and E, F, respectively). This showed that cells grouped into five clusters, three clusters (1, 2 and 4) were mainly squamous and the remaining two (3 and 5) had more columnar marker-expressing cells (Figure 3.12A, B).

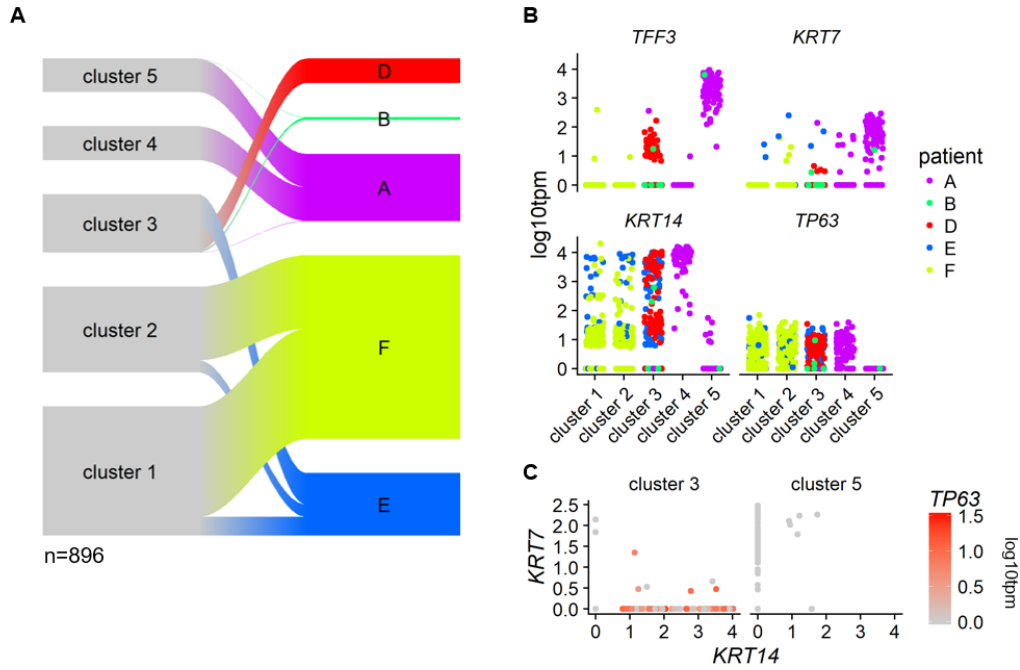


Figure 3.12: Clustering of oesophageal cells from patients with and without Barrett’s oesophagus. (A) River plot showing how oesophageal cells from patients with BO (A, B, D) and without (E and F) contribute to SC3 clusters generated using oesophageal cells only. (B) Jitter plots showing the expression of columnar (*TFF3* and *KRT7*) and squamous (*KRT14* and *TP63*) genes in cells from a coloured by patient (log10tpm) across each cluster. (C) Scatter plots showing the expression of *KRT7*, *KRT14* and *TP63* in the two clusters which contain the most *TFF3*-expressing cells (units are log10tpm).

Of the columnar clusters, cluster 5 consisted of cells from patients A and B and cluster 3 consisted of cells from patients B, D and E (patients A, B, D had BO, patients E, F had no BO) (Figure 3.12B). Although rare in these data, it is interesting that one of the clusters (cluster 3) containing *TFF3*+ cells also had four cells which were positive for the squamous genes *KRT14* (a gene pair with *KRT5*), *TP63* and *KRT7*. As p63+/KRT7+ cells have been shown to generate intestinal-like epithelial cells in organoid culture upon CDX2 over expression, it may be possible that these oesophageal cells could be related to the transitional zone progenitor cells previously observed in humans (Jiang et al., 2017) (Figure 3.12C).

To confirm whether the relationship between BO cells and OSGs was stronger than the associations with other gland-type cells, the RNA compositions of cells from other tissues was observed, i.e. gastric gland cells and BO cells that did not express *CHGA* or *MUC2* (to exclude enteroendocrine and goblet cells, respectively; see Materials

and Methods and Figure 3.13 for thresholding), and oesophageal cells that expressed *TFF3* (to exclude squamous cells). First, I used the software developed by David Severson (BEARscc, an algorithm which uses external controls to simulate technical replicates to check whether a single cell clustering method is robust to technical variability (Severson et al., 2018)). The score metric of BEARscc reflects how frequently cells within a group cluster together, as opposed to with cells from other clusters. I compared manually selected groups of 1) gastric and BO cells, 2) gastric and OSG cells, and 3) BO and OSG cells, from patients with BO (A-D). The BO and OSG cell combination had a higher score than any combination which included gastric cells, or all cells grouped together, suggesting BO and OSG cells have the most stable cell type relationship (Figure 3.13).

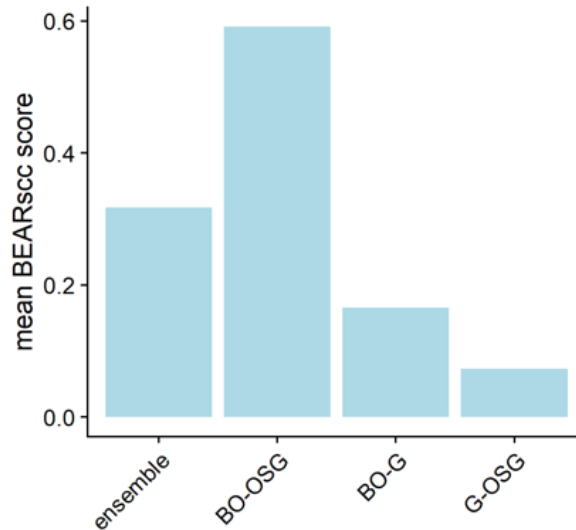


Figure 3.13: Cluster robustness scores from BEARscc in different gland-type cell groupings. Mean BEARscc score for each grouping of 'gland-like' cells (n=372), which are a sub-set of gastric (G, n=175), BO (n=78) and OSG cells (n=119): excluding gastric and BO cells that expressed *CHGA* or *MUC2* (to exclude enteroendocrine and goblet cells, respectively) and excluding oesophageal cells that did not express *TFF3* (to exclude squamous cells). 'Ensemble' refers to all cells grouped together. Thresholds were set at the tenth centile of cells in which at least one transcript was detected from each gene of interest.

Using only these manually selected gastric, BO and OSG cells with additional OSG cells from patients without BO (E-F), unbiased clustering with SC3 also confirmed the strong relationship between BO and OSG cells, with only very few gastric cells clustering with BO or OSG cells (Figure 3.14A). t-SNE, with the inclusion of duodenal cells which expressed the highest levels of *MUC6* to enrich for duodenal Brunner's gland-type cells (Bartman et al., 1998), also confirmed the strong relationship between

BO and OSG cells in tsne (Figure 3.14B). This relationship was characterised by high *LEFTY1* expression (Figure 3.14C).

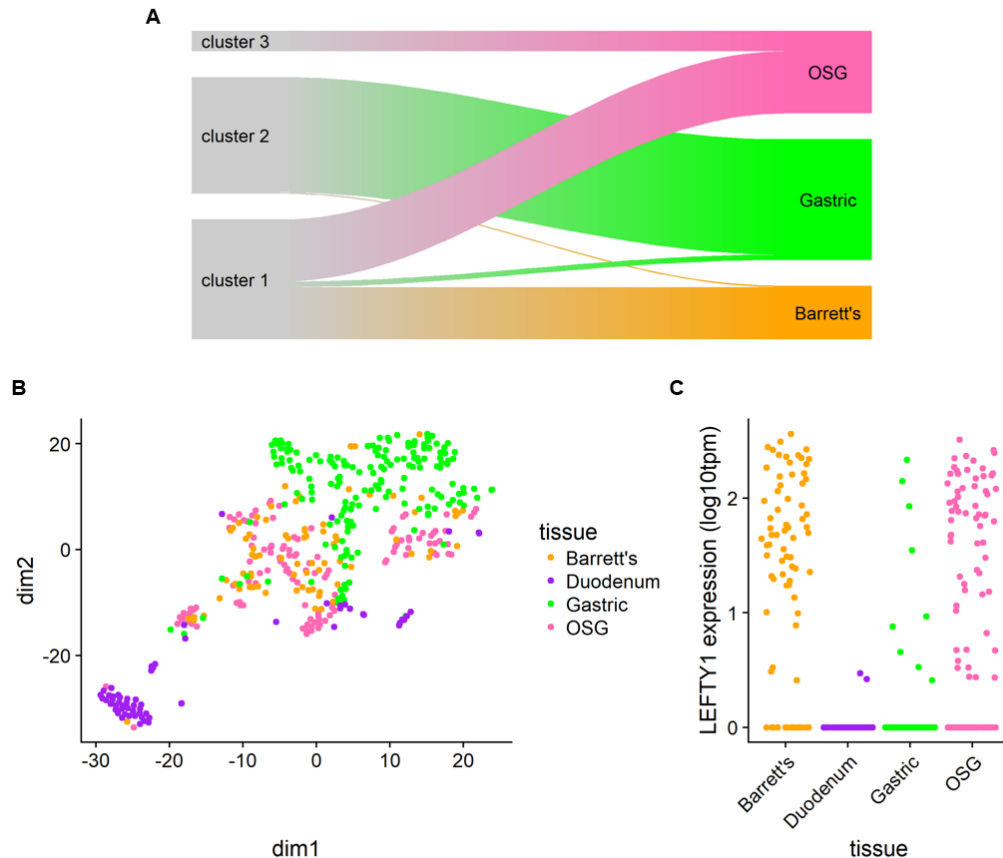


Figure 3.14: Assessment of clustering of gland-type cells from upper gastrointestinal samples. (A) River plot showing how gastric, oesophageal and BO cells contribute to SC3 clusters of 'gland-like' cells from these tissues, which are a sub-set of gastric (n=175), BO (n=78) and oesophagus cells (n=120): excluding gastric and BO cells that expressed *CHGA* or *MUC2* (to exclude enteroendocrine and goblet cells, respectively), and excluding oesophageal cells that did not express *TFF3* (to exclude squamous cells). (B) t-SNE plot of 'gland-like' cells as in A but also including duodenal cells expressing maximal *MUC6* (to enrich for Brunner's gland type cells, n=65). The threshold for including a duodenal cell was set at the fiftieth centile of *MUC6* expression, due to high expression prevalence in duodenal cells. (C) Jitter plot showing expression of *LEFTY1* in 'gland-like' cells as in C (n=438).

Finally, only a small number of mainly mitochondrial genes showed differential expression between BO and OSG cells and gene pathway enrichment analysis on these

genes did not suggest any biological processes that could mechanistically distinguish BO and OSG cells (Figure 3.15A, B).

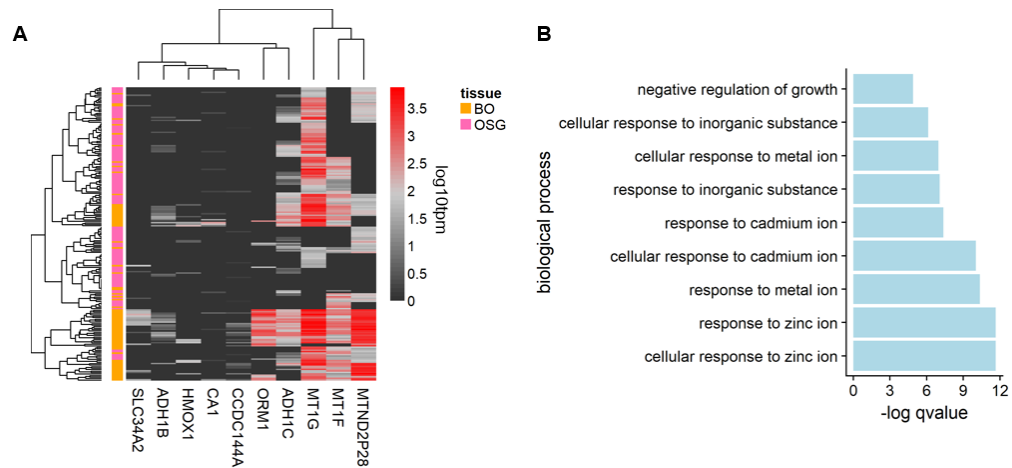


Figure 3.15: Comparison of selected Barrett's oesophagus and oesophageal sub-mucosal gland cells. (A) Heat map with hierarchical clustering of genes and samples showing the top ten most differentially upregulated genes between BO cells (n=78) and oesophagus cells (n=120): excluding gastric and BO cells that expressed *CHGA* or *MUC2* (to exclude enteroendocrine and goblet cells, respectively), and excluding oesophageal cells that did not express *TFF3* (to exclude squamous cells). (B) Enriched gene pathways (FDR < 0.01), based on the full results of A.

In view of the previously described phenotypic overlap with BO and gastric pylorus (Lavery et al., 2014), the transcriptomes of 194 cells from an additional two patients (G-H) with BO (24 oesophageal cells, 32 BO cells, 109 gastric cardia cells and 29 gastric pyloric cells) were next analysed. Unfortunately, the data produced from this batch was poor, but clustering of these cells on global and specific gene expression show that gastric cardia and pylorus exhibited similar RNA composition properties (Figure 3.16A). The BO cells also expressed several of the gastric genes, but showed differences such as increased *KRT7* and *BPIFB1* expression (Figure 3.16B).

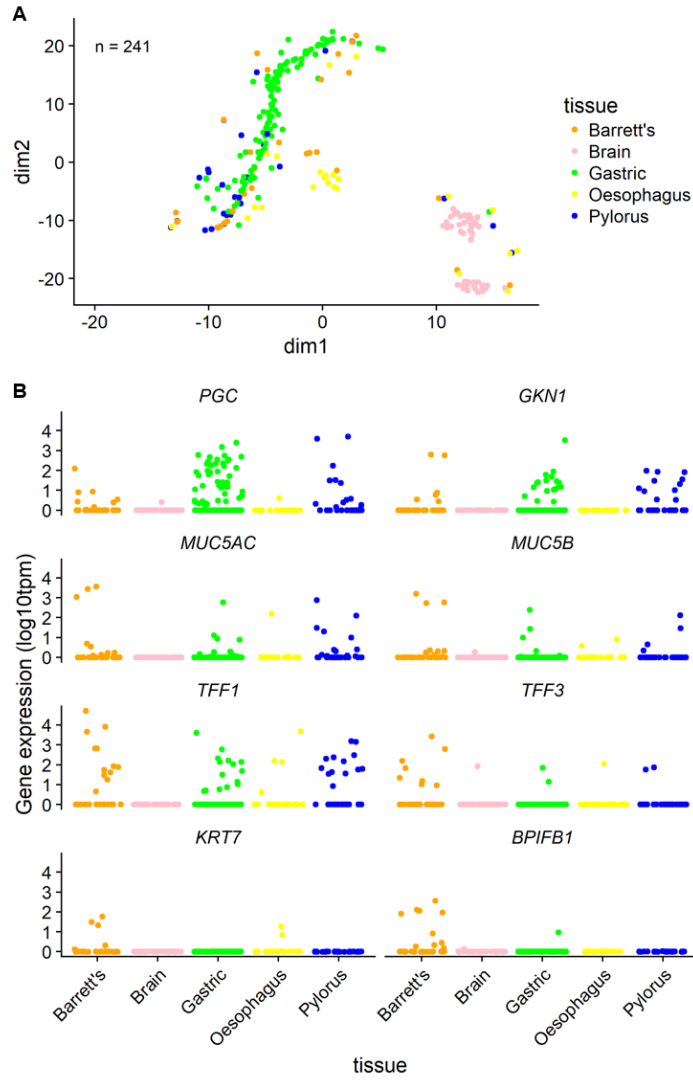


Figure 3.16: Relationships of pyloric cells with other sampling sites in the upper gastrointestinal tract. (A) t-SNE plot of gastric cardia cells (n=109), gastric pylorus cells (n=29), BO cells (n=32), oesophagus cells (n=24) and brain RNA control (n=47) showing transcriptional relationships between cells obtained from an additional two patients with BO. (B) Jitter plots showing expression of selected epithelial genes in gastric cardia, gastric pylorus, BO and oesophagus cells as in A.

Collectively, these data show that oesophageal cells expressing genes seen in OSGs, and not intestinal, gastric or squamous cells, have the greatest RNA composition similarity to BO cells.

3.4 ITLN1 and SPINK4 mark early goblet cells

In this study, 19% of BO cells were classified as goblet cells, which is consistent with the requirement in some countries, such as the US, for goblet cells to be present for the diagnosis of BO. Goblet cells are classically defined by morphological appearance and MUC2 expression, though whether their presence and abundance in BO is an important prognostic marker remains controversial. To identify other goblet cell markers, I looked for robustly differentially expressed genes in the *MUC2*+ goblet cell cluster compared to phenotypically different cell types and identified *ITLN1* and *SPINK4* among numerous others (Figure 3.17A, B and Appendix B.5). Then, applying a threshold set at the tenth centile to include 90% of cells in which at least one transcript was detected from each gene of interest (to reduce biological noise), I found that *MUC2* co-expressed with intelectin 1 (*ITLN1*) and Kazal type 4 serine peptidase inhibitor (*SPINK4*) in 61% of goblet cells from duodenum, gastric and BO samples (Figure 3.17C).

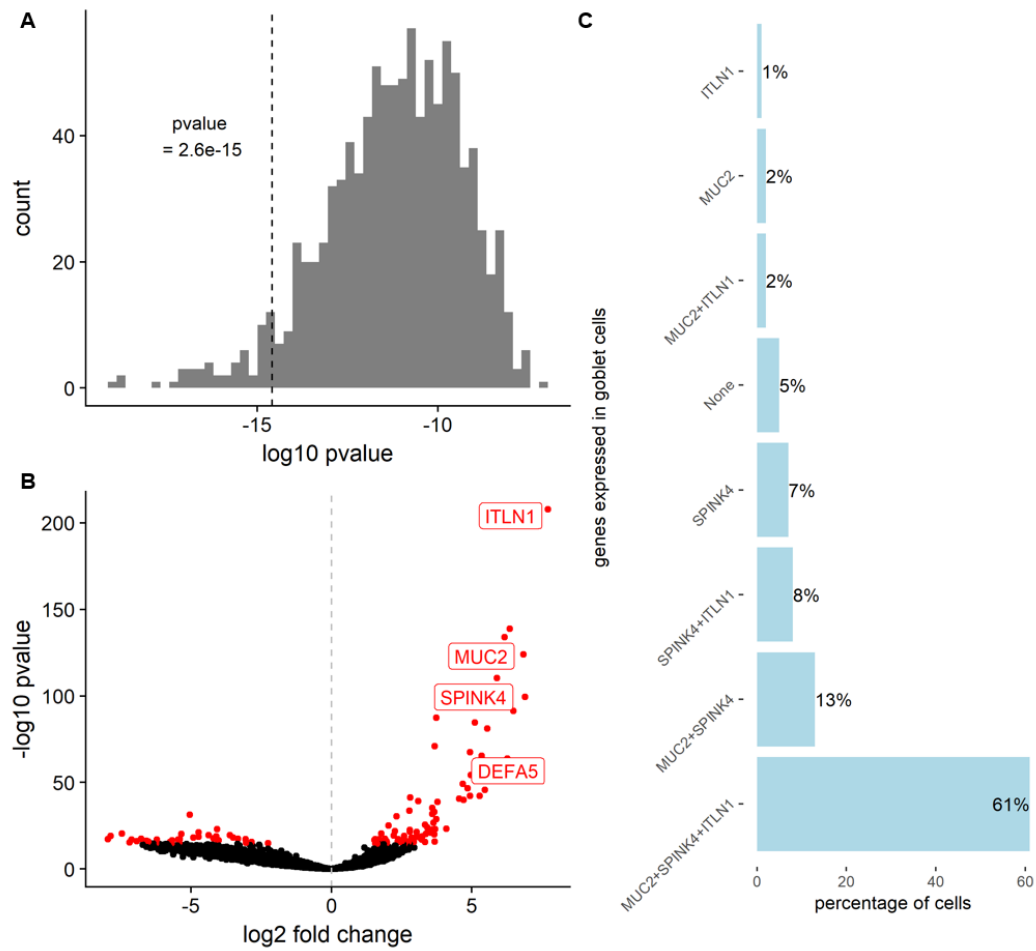


Figure 3.17: Highly expressed genes in goblet cells. (A) Histogram showing the results of a permutation test to establish a significance threshold when comparing cells from the 'goblet-type' cell cluster with all others from patients A-D. Vertical dotted line indicates the 5% cutoff from which the pvalue is taken to filter results. (B) Volcano plot showing fold change and p value of genes differentially expressed in the 'goblet-type' cell cluster as compared to all other cell clusters (see Figure 3.9). Points coloured red indicate genes significant at 5% permutation test (as in A). Selected highly significant genes are labelled. (C) Bar chart showing the percentage of cells in the 'goblet-type' cell cluster (n=98) expressing *MUC2*, *ITLN1* or *SPINK4* alone or in different combinations (thresholds set at the tenth centile to include 90% of cells in which at least one transcript was detected from each gene).

ITLN1 and *SPINK4* have been previously shown to mark goblet cells in the normal gut and some non-gastrointestinal tissues (Washimi et al., 2012; Noah et al., 2010), but we previously observed some cells that uniquely expressed *MUC2*, *ITLN1* or *SPINK4* in BO and colon (Owen et al., 2018). Therefore we hypothesized that the pattern of *SPINK4*, *ITLN1* and *MUC2* expression might mark stages of goblet cell development *in vivo*. Whilst further proof is needed to confirm these observations, we observed

that SPINK4 expression occurs in some cells which lack goblet cell morphology and tend to localise in the crypt base of intestinalised BO, where undifferentiated cells and Paneth cells reside (Figure 3.18A, red arrows). This suggests that SPINK4 expression could indicate an earlier stage of goblet cell differentiation than MUC2, or that SPINK4 marks so-called 'intermediate cells', which exhibit some mixed goblet-Paneth characteristics (Cunliffe et al., 2001). Since Paneth cells are characterised by human defensin 5 expression (*DEFA5*), co-expression of *MUC2*, *SPINK4* and *DEFA5* in goblet cells was next examined. This reveals that *DEFA5* was rarely expressed in *SPINK4* cells (Figure 3.18B), with only 9.5% of *SPINK4* positive cells expressing any *DEFA5*, and that cells which did express *DEFA5* were actually more likely to express *MUC2* (Figure 3.18C, white dotted box area). These findings suggest that cells can exhibit 'goblet-like' tendencies even in the absence of morphologically typical goblet cells, which may have implications in the clinical definition of BO in some countries, as discussed.

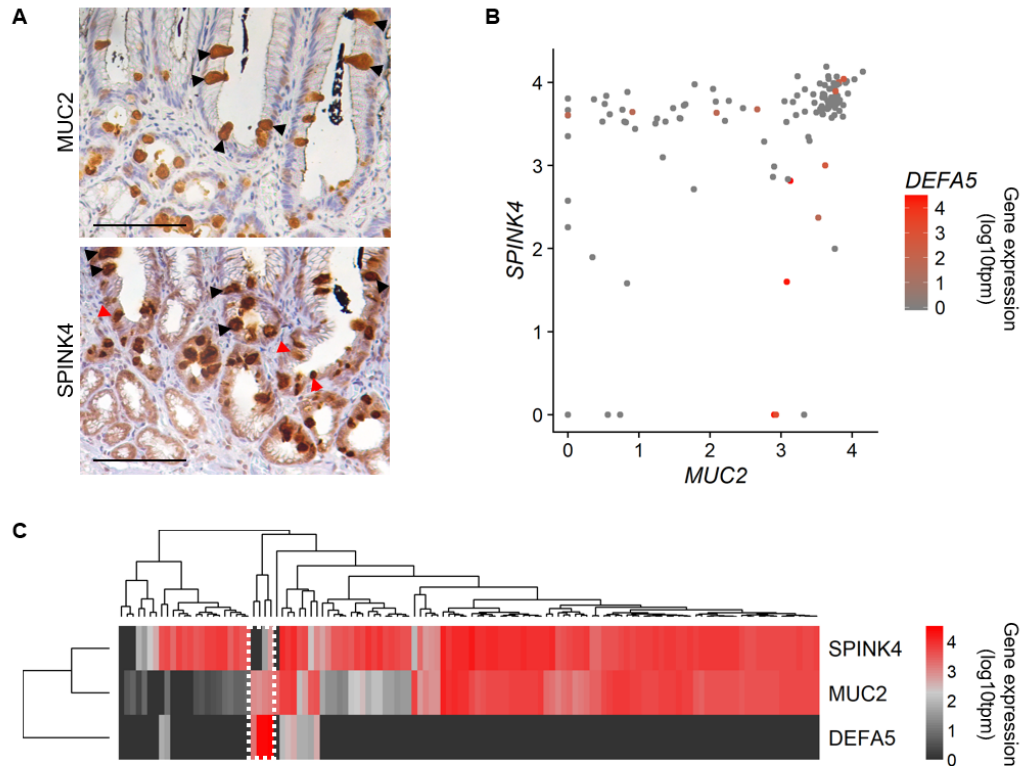


Figure 3.18: Further characterisation of *SPINK4* expressing cells that are negative for *MUC2*. (A) Immunohistochemical staining of *MUC2* (upper image) and *SPINK4* (lower image) in adjacent sections from an endoscopic mucosal resection specimen of BO. Scale bars are 100µm. Black arrows mark morphologically typical goblet cells and red arrows mark positively stained cells which do not morphologically resemble goblet cells. (B) Scatter plot showing the correlation of expression of *MUC2* and *SPINK4* in individual cells. *DEFA5* expression (a Paneth cell marker) is also shown by colour scale. (C) Heatmap showing expression of *MUC2*, *SPINK4* and *DEFA5*, white dotted box outlines cells which expressed the highest levels of *DEFA5*

3.5 *OLFM4* marks a stem-like transcript profile in Barrett’s oesophagus and oesophageal sub-mucosal gland cells

StemID is a published workflow designed to find cells with stem-like properties in single cell RNA-seq data by calculating a stem-ness score based on the entropy of cell clusters and the number of links between clusters (Grün et al., 2015, 2016). As a control I analysed duodenum cells from BO patients (A-D) and found the highest scoring cluster was enriched for *LGR5* expression, consistent with *LGR5* being a

known marker of intestinal stem cells (Barker et al., 2007; Sato et al., 2009). Applying StemID to the remaining individual tissues from the same patients did not identify any well-known stem cell markers (Figure 3.19), even though a small number of *LGR5+* cells are present in all tissues sequenced (Figure 3.6). Despite this, several highly significant genes were identified for each tissue and future corroborative work may make this data relevant in the future. The high level of significance of these markers means that the p values are near zero, and therefore an artificial ceiling is placed on the plots so that the most significant results are not excluded, this explains the odd conformation of the volcano plots from StemID seen in relevant figures (Figure 3.19 and Figure 3.20).

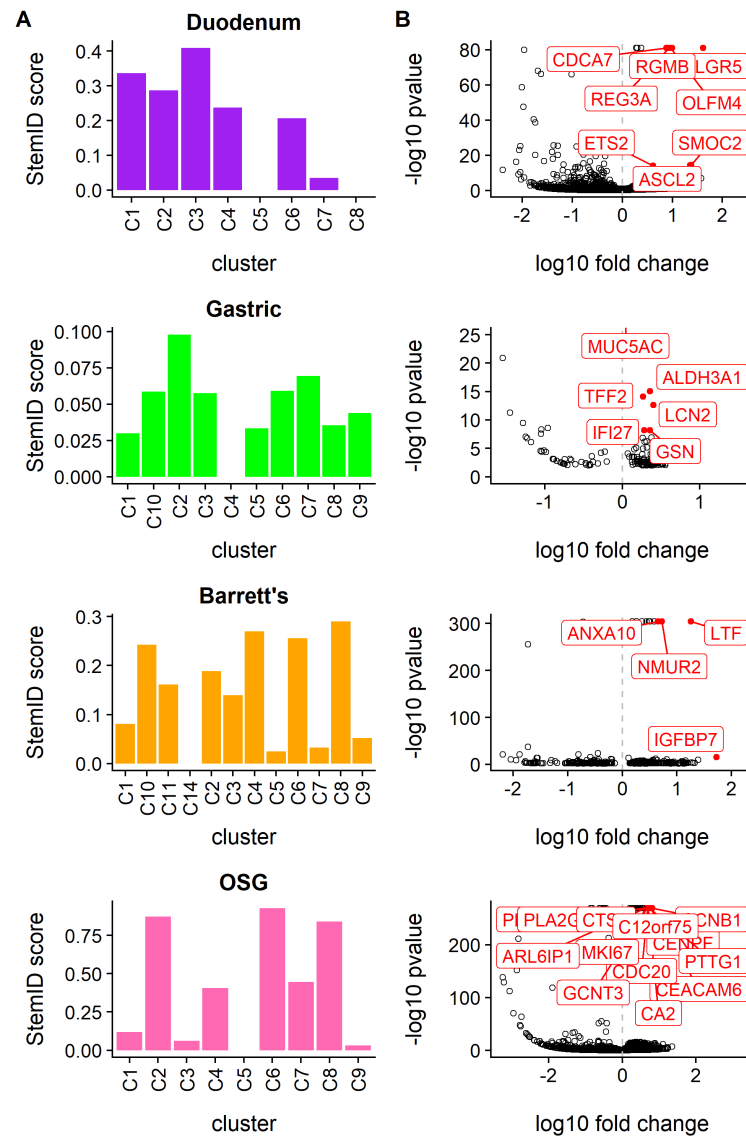


Figure 3.19: Differentially expressed genes in candidate stem cell-like clusters in upper gastrointestinal tissues. (A) StemID scores across all clusters computed for duodenum (purple), gastric (green), Barrett's oesophagus (orange) and oesophageal submucosal gland cells (pink) in bar plots from top to bottom. Scores are calculated from multiplication of the entropy (spread from the cluster mean) and the number of cluster links arising from a given cluster. Oesophageal submucosal gland cells were deduced by removing squamous (*KRT14+*) cells from all normal oesophagus cells from patients A-D (cells from patients E and F were removed to avoid comparing cells between batches). (B) Volcano plots showing significantly expressed genes in the highest scoring cluster corresponding the adjacent plot in A, from top to bottom these are duodenum, gastric, Barrett's oesophagus and oesophageal submucosal gland cells.

Since a study showed that BO contains pluripotent cells (Yamamoto et al., 2016) and in view of the striking transcript profile overlap between OSG and BO cells, I therefore

analysed all BO and OSG cells using StemID (patients A-F). Interestingly, the highest scoring cluster was enriched for the stem-cell associated gene *OLFM4* (Figure 3.20A, C, blue asterisk). BO cells from all four patients with BO (A-D) contributed to this cluster, and oesophageal cells from two patients with BO (A and B) (Figure 3.20B). The second highest scoring cell cluster (Figure 3.20A, C, red asterisk) was enriched for *LYZ*, a marker of Paneth cells, which are long-lived secretory cells found adjacent to the stem cell niche in the intestinal crypt base.

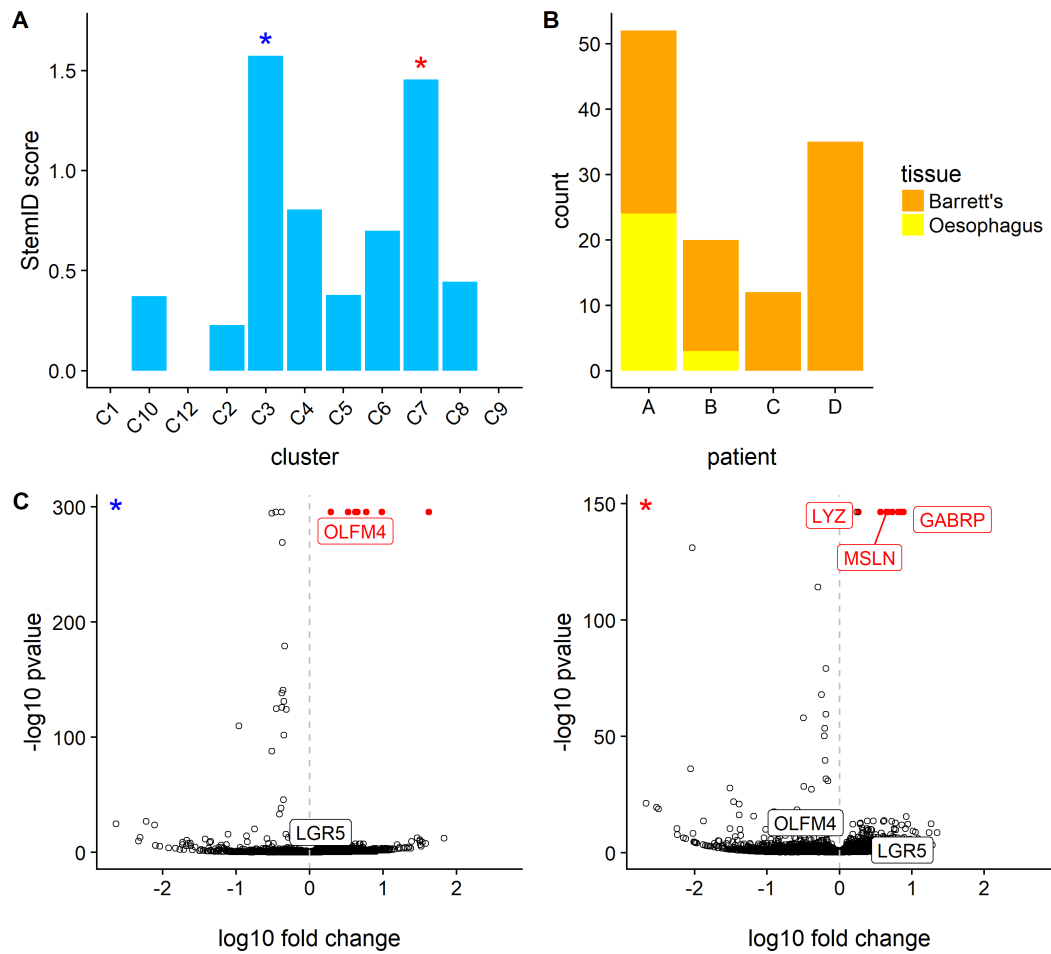


Figure 3.20: Differentially expressed genes in stem cell candidates from a combination of Barrett's oesophagus and OSG cells. (A) Bar plot shows StemID scores across all RaceID2 clusters applied to all non-squamous oesophageal cells (BO and oesophageal cells with <5 KRT14 counts to exclude squamous cells, $n=533$). Scores are calculated from multiplication of the entropy (spread from the cluster mean) and the number of cluster links arising from a given cluster. (B) Bar plot showing the number of BO and OSG cells isolated from each patient and represented in the highest scoring StemID cluster in A (cluster 3). (C) Volcano plots showing differentially expressed genes in the highest scoring cluster (C3, blue asterisk, left side) and second highest scoring cluster (C7, red asterisk, right side). Points coloured red indicate the most significant genes with a fold change > 2 . Selected highly significant genes are labelled.

OLFM4 has been shown to associate with LGR5 expression and marks stem cells in intestinal tissue in normal and metaplastic contexts (van der Flier et al., 2009; Jang et al., 2015), and has also been shown in *in situ* hybridisation assays in BO (Lavery et al., 2014). Consistent with this, immunohistochemical staining detected OLFM4 expression in human duodenal crypt bases, where stem cells are known to be located (Figure 3.21A). In BO, OLFM4 expression was less restricted to the crypt

base, especially in less intestinalised regions lacking goblet cells (Figure 3.21B), however OLFM4 was absent in gastric pyloric glands, despite the previously observed phenotypic similarities (Figure 3.21C) (Lavery et al., 2014).

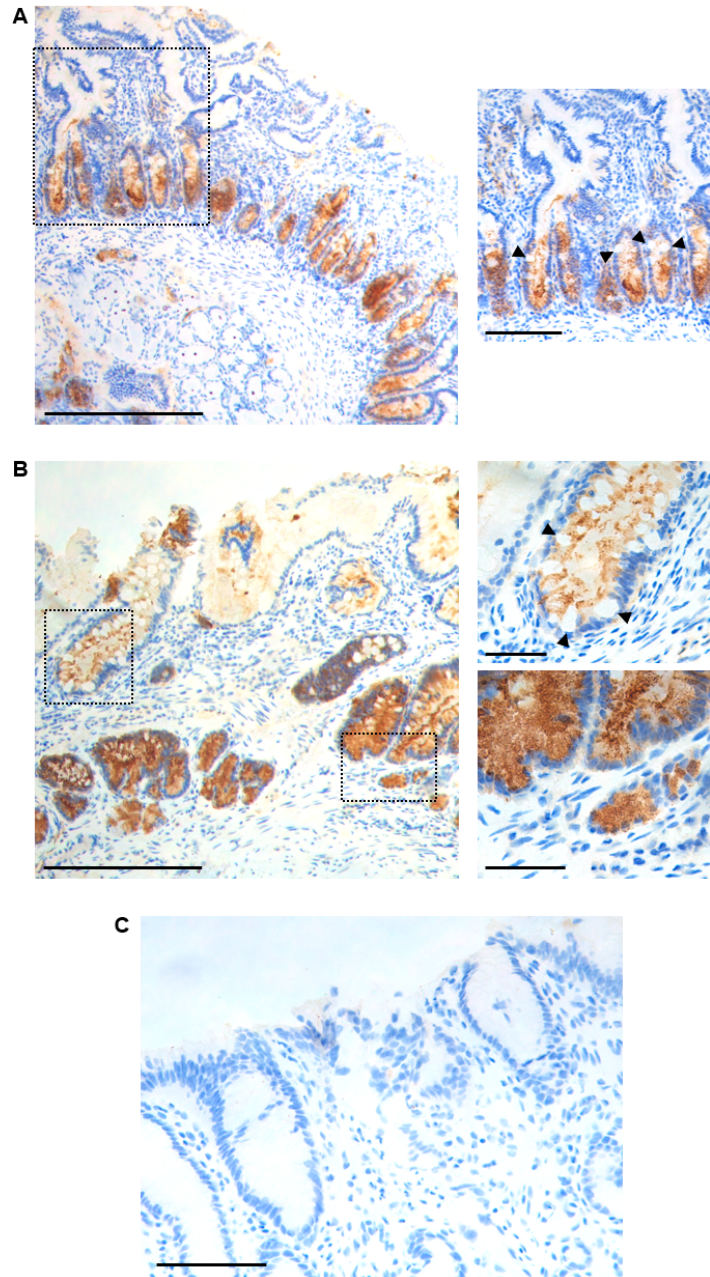


Figure 3.21: Immunohistochemical staining of OLFM4 in human upper gastrointestinal tissues. (A) Immunohistochemical staining of OLFM4 in human duodenum. Expanded image shows crypt bases, marked by the black dotted box. Black arrows highlight selected goblet cells. Scale bars are 250 μm and 100 μm in expanded image. (B) Immunohistochemical staining of OLFM4 in human BO. Expanded images shows selected areas, marked by the black dotted boxes, from superficial intestinal type BO glands containing goblet cells (marked with black arrows) and deeper glands. Scale bars are 250 μm and 100 μm in expanded image. (C) Immunohistochemical staining of OLFM4 in human gastric pylorus. Scale bar is 100 μm .

Notably, *OLFM4* has a higher mean expression in the *LEFTY1*⁺ clusters from BO cells (B2/B3) compared to the clusters expressing known markers of the differentiated goblet (*MUC2*) and enteroendocrine (*CHGA*) lineages (Figure 3.7) To examine co-expression of *OLFM4*, *LEFTY1*, *MUC2* and *CHGA* in individual cells, a threshold was applied at the tenth centile to include 90% of cells in which at least one transcript was detected from each gene of interest. Using this threshold, half of the BO cells express *LEFTY1* and *OLFM4*, alone or in combination (29% *OLFM4* and *LEFTY1*; 13% *OLFM4* only; 11% *LEFTY1* only). *LEFTY1*⁺ and *OLFM4*⁺ BO cells rarely co-expressed *MUC2* or *CHGA* (Figure 3.22) Together, these data suggest that B2/B3 represent a cell population that harbours BO progenitor cells.

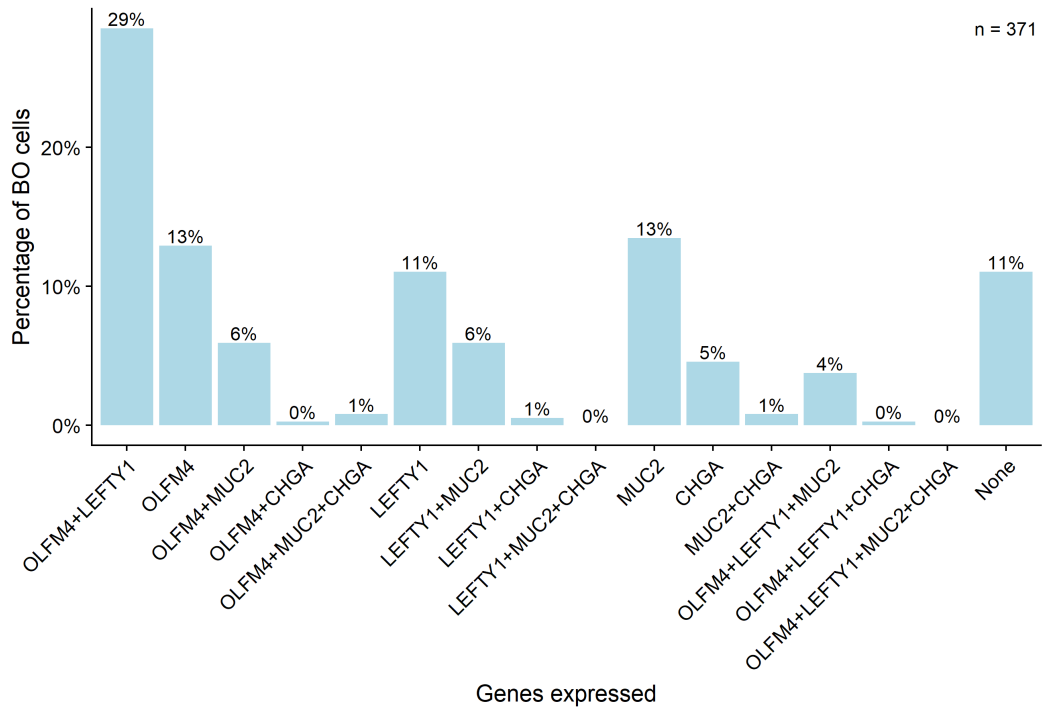


Figure 3.22: Co-expression of *OLFM4*, *LEFTY1*, *MUC2* and *CHGA* in Barrett's oesophagus cells. Single cell RNA-seq data showing percentage of Barrett's oesophagus cells from patients A-D expressing *OLFM4*, *LEFTY1*, *MUC2* or *CHGA* alone and in all combinations (thresholds for calling a gene 'expressed' were set at the tenth centile to include 90% of cells in which at least one transcript was detected from each gene).

3.6 Summary

BO is associated with an increased risk of oesophageal adenocarcinoma, but its cell of origin is unclear. Using scRNA-seq of biopsies from 6 patients and 2 unaffected subjects I have shown that cells in BO have a transcriptional profile that is strikingly similar to that of cells in OSGs, more so than any other gastro-duodenal gland or epithelial cells. This is supported by observations of similar stem-related gene expression, and upregulation of the developmental gene *LEFTY1*. I also show that *SPINK4* seems to identify an different stage of intestinal metaplasia than is marked by *MUC2* expression, as *SPINK4* is highly expressed in *MUC2*+ cells and stains cells which are morphologically atypical for goblet cells in the crypt base of intestinalised BO. Of clinical importance, these results suggest that intestinal goblet cell characteristics exist even in the absence of morphologically identifiable goblet cells, supporting the view that diagnosis of BO should not require the detection of goblet cells. Together, my findings help characterize BO in humans and demonstrate the power of single cell analysis of clinical samples to uncover biological relationships among cell types and cellular heterogeneity in healthy and diseased tissues, albeit in a limited number of patients.

Chapter 4

Calling mutations in single cell RNA-seq data

The mutational and clonal nature of BO has been previously investigated in bulk samples and in laser capture micro dissected tissues, which have limitations. In this chapter I aim to investigate the mutational landscape and clonal architecture of BO at a single cell resolution, and see how this relates to cell phenotype, specifically relating this to OSGs. To achieve this I use machine learning approaches to enhance existing variant callers and design an R package called 'snv.predict', testing it with primary cells and a cell line. Using snv.predict, I also investigate a publicly available dataset to examine how breast cancer mutational burden changes in lymphatic spread.

4.1 Single nucleotide variants in single cell RNA-seq

Using a variant caller designed to be highly sensitive to detecting mutations (VarScan2) (Koboldt et al., 2012), restricted to identifying only single nucleotide variants (SNVs), a mean of 2200 SNVs per cell were identified in 1046 upper gastrointestinal cells from four patients (Figure 4.1A).

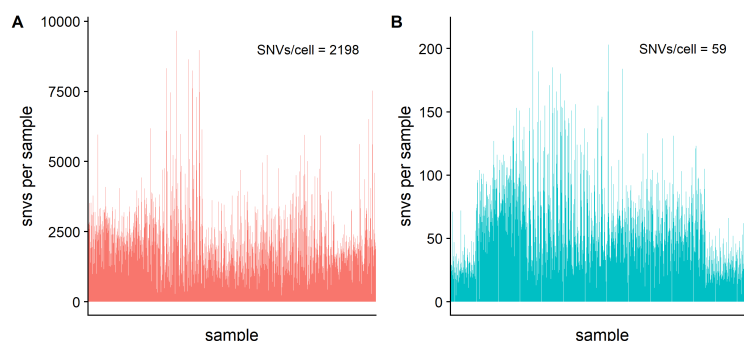


Figure 4.1: Number of single nucleotide variants detected per single cell sample. (A) Number of SNVs detected in chromosomes in each sample. (B) As in A, but restricted to SNVs called in ERCCs. Mean number of SNVs per cell are displayed in the top right of each plot.

These SNVs were spread across chromosomes and regions (Figure 4.2A). Interestingly, a mean of 59 SNVs per cell were detected in synthetic spiked in control sequences developed by the External RNA Controls Consortium (termed ERCCs (Jiang et al., 2011)) which were included in the sequencing experiments to provide additional methods of data normalisation, and were utilised to model technical noise between samples (Severson et al., 2018) (Figure 4.1B). This detection of mutations in ERCCs highlights the problems with false positives when variant calling in single cell RNA-seq.

4.1.1 Identifying ground truths in single cell RNA-seq variant calling

I hypothesised that machine learning approaches could be trained to predict true positive mutations using ERCC derived false positives and 'true positives' obtained from existing calls by applying more stringent filtering. To class as a true positive, each SNV had to meet the following criteria.

- Be present in over nine cells
- Exist in the Database of Single Nucleotide Polymorphisms (dbSNP) (Sherry et al., 2001)
- Exist in bulk RNA-seq data from the same tissue samples

Previous work has shown high precision of identifying RNA-seq mutations in matched genomic sequencing, and thus RNA editing sites represent a small minority (<1%) of mutations called (Nishikura, 2010). The most common form of RNA editing is by adenosine deaminases acting on RNA (ADAR) enabled A → I, which changes the content of the RNA molecule and is interpreted as guanosine (G) by the translational machinery (Reenan, 2005). On the antisense strand, the same process would be interpreted as a T → C mutation. A limitation of most single cell RNA-seq data is the lack of strand specificity, complicating the recognition of RNA editing sites. For the purpose of trying to reduce the number of SNVs called as a consequence of RNA editing, I also removed A → G mutations from the 'true positive' SNVs, as these have been shown as the most common in non strand specific RNA-seq data, and more common than T → C mutations (Kang et al., 2015).

Applying these conservative criteria showed that in most cells there were ample SNVs labelled as true and false positives, though the majority of calls remain unknown (Figure 4.2B).

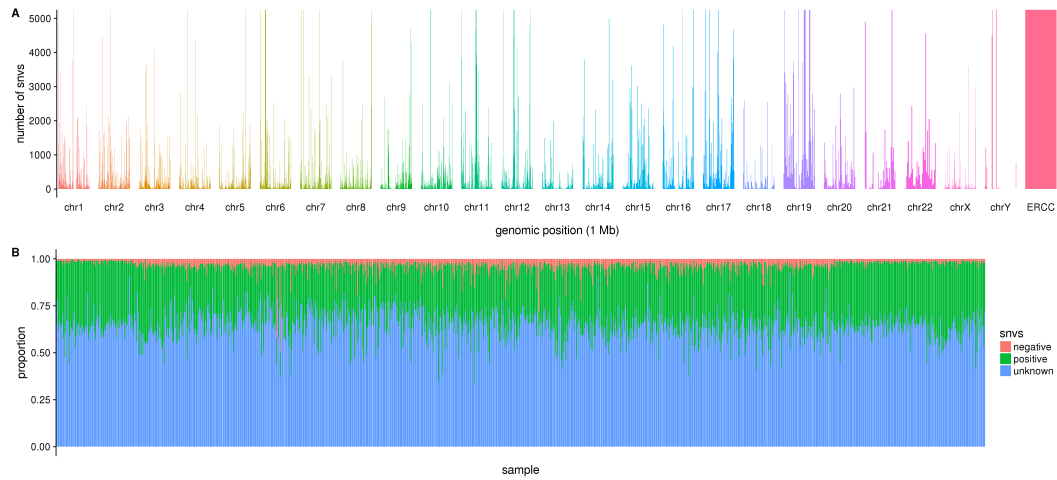


Figure 4.2: SNVs called in samples across chromosome and annotated by positive or negative control status. (A) Sum of mutations detected across chromosomes and regions in 1040 upper gastrointestinal cells and controls. (B) Proportion of mutations detected in each sample that were false positive (filled red), true positive (defined as not in an RNA editing site and present in over 9 cells and in bulk control, filled in green) and unknown (filled in blue).

Then, a software package ('nuseq', see bitbucket.org/arothe85/nuseq, derived from Ding *et al.* (Ding et al., 2012)) was used to extract classifiers for each called SNV from BAM files and the reference genome. From these, a further series of classifiers were made using `snv.bench`, a function within `snv.predict` which normalises nuseq classifiers where possible and adds the result a new SNV classification table (table of classifiers in Appendix C, `snv.predict` installable from github.com/mrrichowen/snv.predict, code available in Appendix D, excluding elastic net logistic regression as this is not enabled in the public repository installation). Principal component analysis shows that false positive SNVs cluster together, while true positive SNVs follow a much broader spread in the first three principal components (Figure 4.3). This data structure would suggest that appropriate modelling may be able to identify true positives from the large set of 'unknown' SNVs (Figure 4.2B). A more robust way to identify true positives would be to use a genome or exome sequencing with germline controls to recognise and annotate SNVs called from single cell RNA-seq data as somatic and germline mutations, but these data were not available to be paired with the single cell RNA-seq data here and hence the bulk RNA-seq data were used. A further limitation of single cell sequencing data discussed previously is allelic dropout, which could result in different genotypes in heterozygous single cells due to differential loss of alleles.

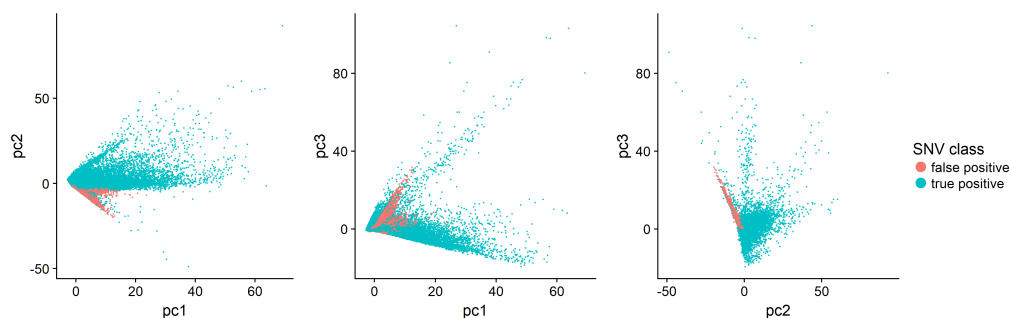


Figure 4.3: Exploring the data structure of SNVs by control status. Principal component analysis (first three principal components shown) of all true positive and false positive calls made using VarScan2 using the output of nuseq and snv.bench as variables. pc=principal component.

The total number of SNVs identified in a given cell did correlate with the total number of mapped fragments (Figure 4.4A). Some normalisation procedures were considered to account for this, but since the total number of genes detected also correlated, it was felt much of this effect would be due to the increased number of genes from which to call SNVs (similar to increased genome coverage), and that the number of cells included in this analysis represents a wide spread of depth and number of genes detected per cell, hence should limit the effect of sequencing depth on SNV calls as long as cells are not examined independently, or in too small a cluster. Looking at the effect of total mapped fragments and gene detection on the principal components defined in Figure 4.3 confirms that there is some influence of depth on principal components, however this is fairly limited for most cells (Figure 4.4B). In applications where cell numbers are more limited, normalising for sequencing depth and gene detection will need to be revisited.

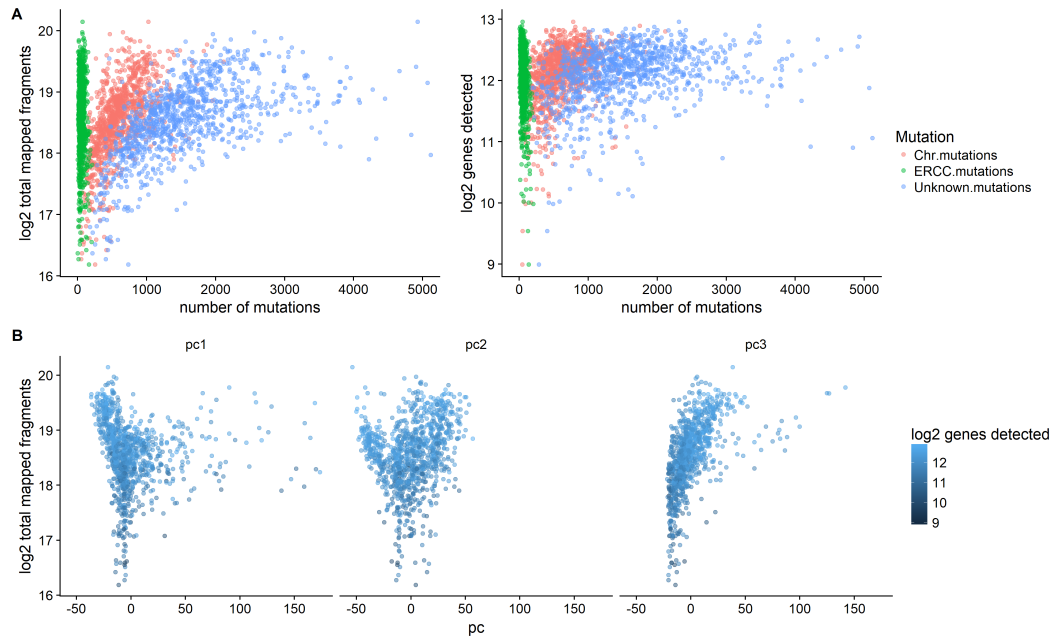


Figure 4.4: The effects of gene detection and sequencing depth on SNV calling. (A) Scatter plots showing the total number of SNVs detected against total mapped fragments (left plot) and total number of genes detected (right plot). Points are colour coded by whether SNVs were true positives (Chr.mutations, meeting conservative filtering criteria shown in list above), detected in ERCCs (ERCC.mutations) or neither (Unknown.mutations). (B) Scatter plots of total mapped fragments against the first three principal components, as calculated for Figure 4.3. Points are colour coded by coverage (gene detection). pc=principal component.

4.2 Machine learning augments SNV calling in single cell RNA-seq

Classifiers derived from SNVs called from single cell RNA-seq data in upper gastrointestinal cells and a myelogenous leukaemia cell line (K562) were first tested for correlation, and the caret package in R used to remove classifiers which to reduce pairwise correlations (Figure 4.5A, cut off set at Pearson correlation coefficient of ± 0.75). Figure 4.5B shows a schematic of how this was approached and the basic concept of a package to predict SNVs ('snv.predict') which incorporates a 'snv.performance' module to assess only true positive and false positive SNV calls to allow performance assessment and model tuning to optimise accuracy.

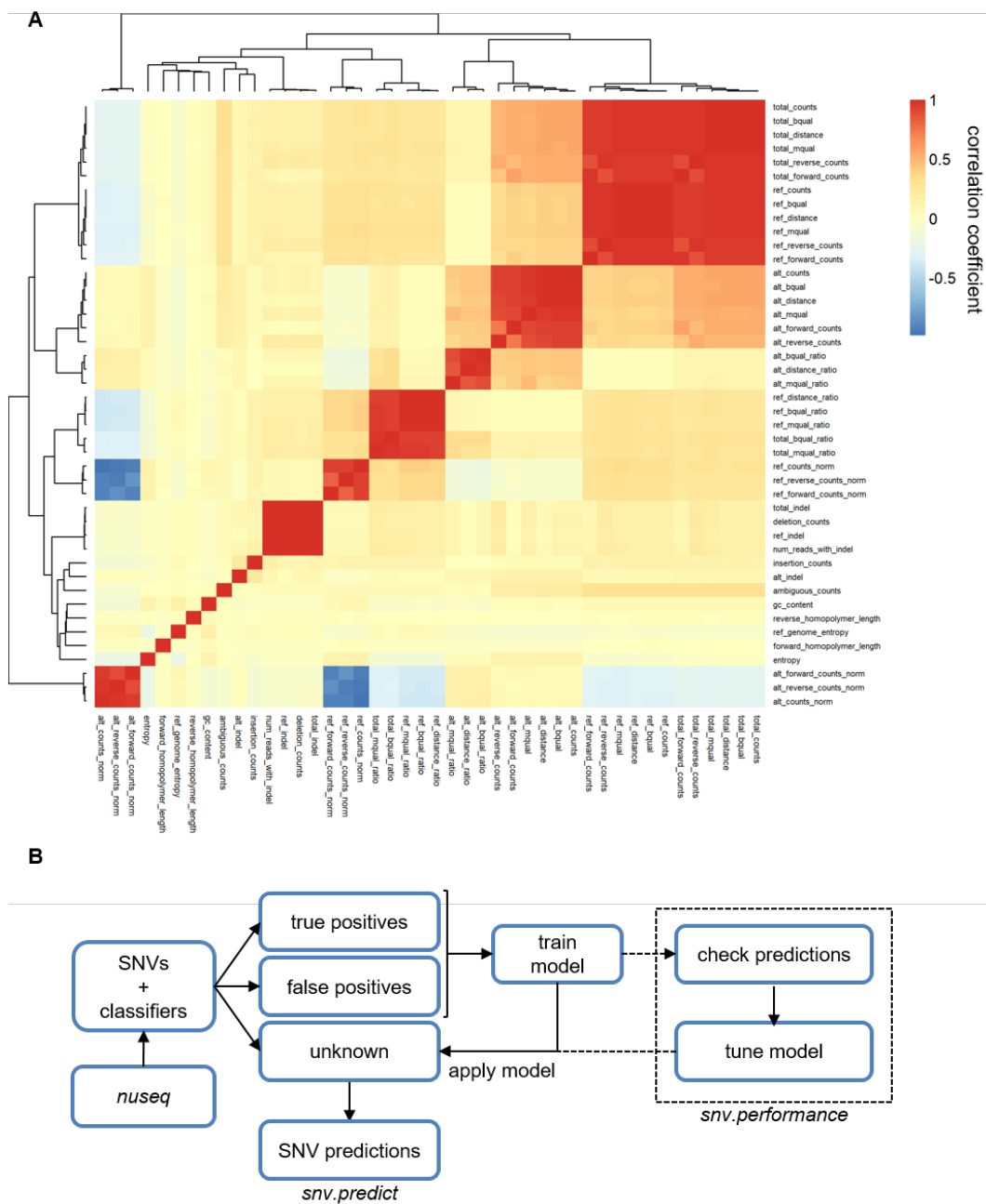


Figure 4.5: Assessment of confounding SNV metrics and schematic of `snv.predict`. (A) Heatmap showing the Pearson correlation coefficients for classifiers. Red and deep blue marked classifiers tended to correlate and were filtered out of the classifiers used in model training. (B) Schematic of how machine learning models were trained using classifiers derived from called SNVs with nuseq. 'snv.performance' is a function within the snv.predict package to allow the optimisation of machine learning methods before running predictions on all unknown SNVs.

For benchmarking machine learning approaches (using snv.performance), all primary

cell SNVs were managed as an ensemble and K562 cell line SNV calls were treated separately. This was because the positive control criteria were different for K562 as no bulk RNA-seq data were available. In K562 cells, a publicly available genome was used as an alternative to bulk RNA-seq (obtained from ENCODE, accession SRX285595). Table 4.1 summarises the datasets used and the relevant controls available.

cells	type	n	bulk control	negative control
Barrett's	human primary	371	bulk RNA-seq	ERCC
Duodenum	human primary	207	bulk RNA-seq	ERCC
Gastric	human primary	221	bulk RNA-seq	ERCC
Oesophagus	human primary	247	bulk RNA-seq	ERCC
K562	haemopoietic line	192	bulk genome	ERCC
Breast	human primary	515	na	na

Table 4.1: Cells used for SNV calling. Table summarising the cells used for calling SNVs and the controls used for snv.predict in each case. Breast cells were from (Chung et al., 2017).

SNV calling was performed using two callers, VarScan2 and HaplotypeCaller (Koboldt et al., 2012; Poplin et al., 2017). SAMtools, GeMs and Freebayes (Li, 2011; Murillo et al., 2016; Garrison E, 2012) were also used, but these either failed to run in some single cell samples, failed to produce usable data, or did not run in a reasonable time scale. Data were split for testing and training equally. For assessing performance, confusion matrices were made and receiver operating characteristics with area under curve (AUC) used. Elastic net logistic regression (alpha=0.5 to blend ridge and LASSO regularisation), random forest and support vector machine learning were all used, to encompass different methodological approaches to machine learning. From confusion matrices, the accuracy of SNV calling was substantially increased with the use of all machine learning approaches and error rate was reduced (Figure 4.6, $p=0.03$; MannWhitney U test).

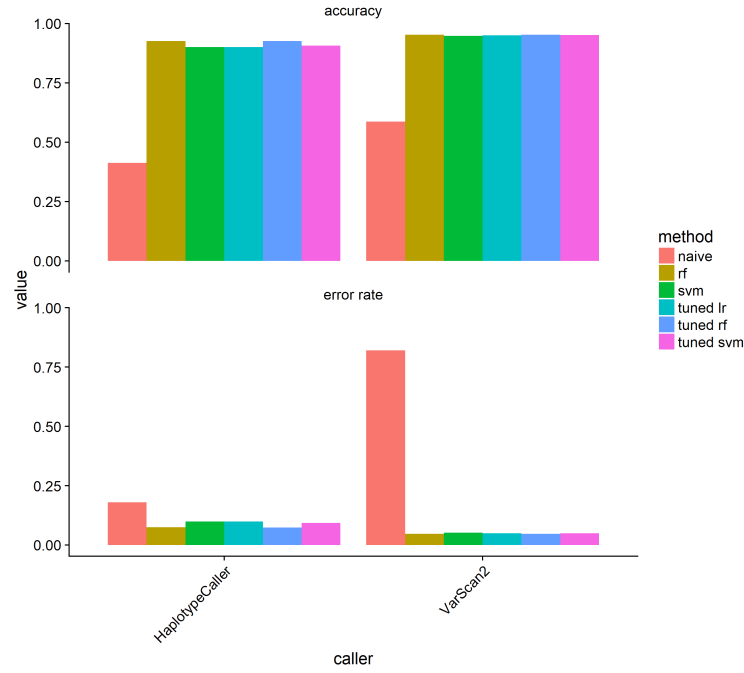


Figure 4.6: Comparison of two SNV callers which functioned in single cell RNA-seq data. Bar plots showing the accuracy (top plot) and error rate (bottom plot) of machine learning augmented SNV calling with two commonly used SNV callers (VarScan2 and HaplotypeCaller). rf=random forest; svm=support vector machine learning; lr=elastic net logistic regression.

Receiver operating characteristic plots show that, compared to naive SNV calling, all machine learning approaches resulted in significant performance improvements (Figure 4.7). For comparison to naive calling methods, a mean value for true positive and false positive rates was calculated across all samples for each caller (producing a single point, hence no AUC available). Further tuning of machine learning models resulted in additional enhancement of the AUC in each case.

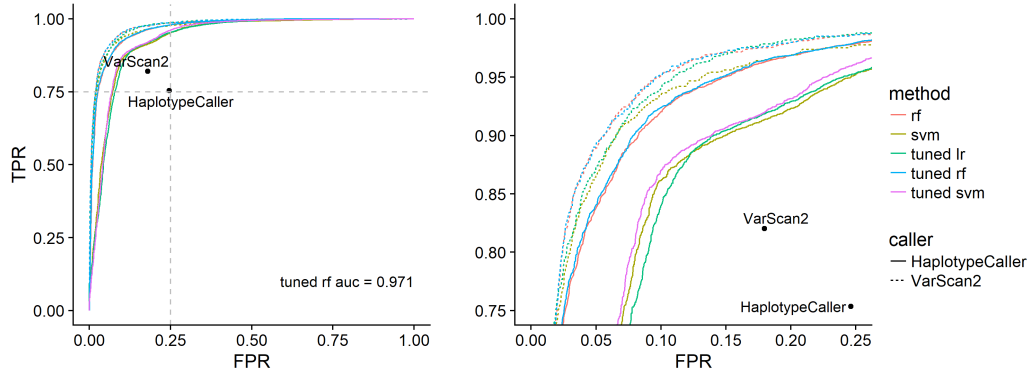


Figure 4.7: Performance of different SNV callers and machine learning approaches to SNV calling. Receiver operating characteristic (ROC) plot showing the performance of HaploTypeCaller and VarScan2 SNV calling (labelled black points, taken as mean over all 1046 upper gastrointestinal cell samples) and ROC curves showing the performance of each machine learning approach in tuned and untuned versions. The left hand plot shows full false positive rate and true positive rate (FPR and TPR) scales and right plot shows a partial scale, as indicated by the grey dotted line bounded area in the left plot. AUC for the best performing model shown. rf=random forest; svm=support vector machine learning; lr=elastic net logistic regression.

With tuned random forest machine learning, VarScan2 produced a bimodal distribution of prediction scores (Figure 4.8A). HaploTypeCaller tended to score SNVs positively (Figure 4.8B). Whether this correlates to a generally better performance of HaploTypeCaller applied naively is difficult to say in these data due to the lack of robust positive controls, but inspection of the results shows a reduced ERCC SNV calling rate in HaploTypeCaller which likely accounts for the lack of negative (<0.5) predictions (Figure 4.8C). Why HaploTypeCaller recognises less SNVs in ERCCs is not clear, but may be due to coding sections designed to recognise commonly used control data. Notably, VarScan2 is designed to be highly sensitive, and not highly specific without additional stringencies, and thus the lower total number of SNV calls made by VarScan2 casts some doubt on the accuracy of HaploTypeCaller in this context.

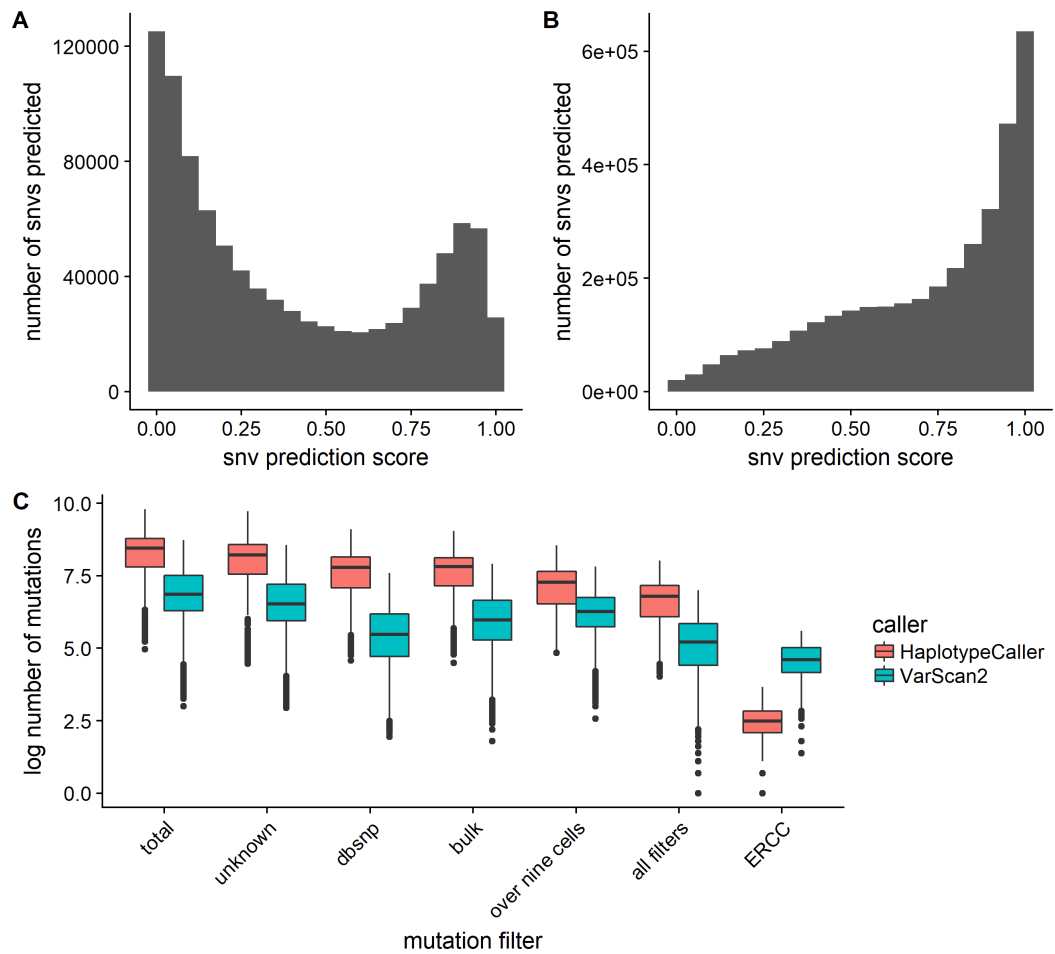


Figure 4.8: SNV call filtering effects of `snv.predict`. (A) Histogram showing distribution of prediction scores from `snv.predict` using a tuned random forest model based on calls from VarScan2. (B) Histogram showing distribution of prediction scores from `snv.predict` using a tuned random forest model based on calls from HaplotypeCaller. (C) Boxplots showing how many SNV calls were made in each category with each SNV caller.

Since `snv.performance` demonstrated enhancement of SNV calling in human primary upper gastrointestinal cells with bulk RNA-seq as part of true positive filtering, it was also applied to K562 cells in which genomic sequencing data were used to define true positives. This shows significant performance improvements except with support vector machine learning (Figure 4.9).

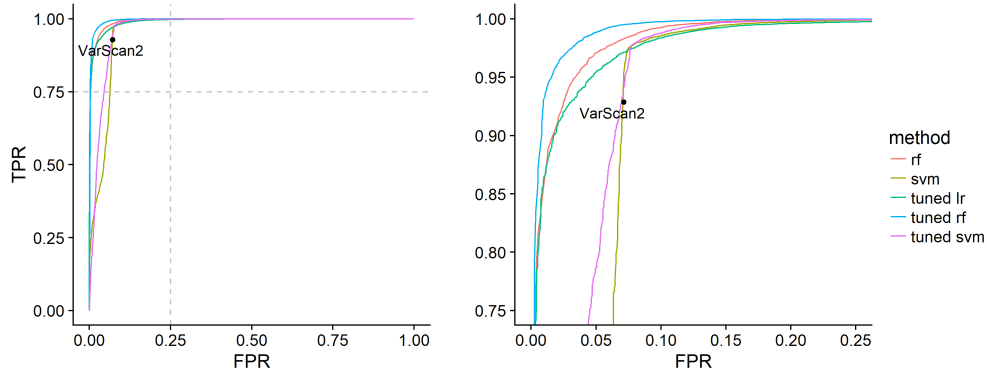


Figure 4.9: Performance of snv.predict in a cell line with genomic positive control. Receiver operating characteristic (ROC) plot showing the performance of VarScan2 SNV calling (labelled black point, taken as mean over all 192 K562 cell line samples) and ROC curves showing the performance of each machine learning approach in naive and tuned versions. The left hand plot shows full FPR/TPR scale and right plot shows a partial scale, as indicated by the grey dotted line bounded area in the left plot. Best performing model was tuned random forest (AUC=0.994). rf=random forest; svm=support vector machine learning; lr=elastic net logistic regression.

4.3 Single nucleotide variants in Barrett's oesophagus

Using machine learning to score SNVs gives a quantified prediction of a SNV being true. By altering the threshold at which a SNV is taken as true in upper gastrointestinal primary cells, the specificity of the call becomes higher (Figure 4.10). Interestingly, the corresponding decrease in true positive rate declines in a more linear fashion until very high prediction values are taken (>0.8), this is likely to be a consequence of the bias towards true positives used when model training.

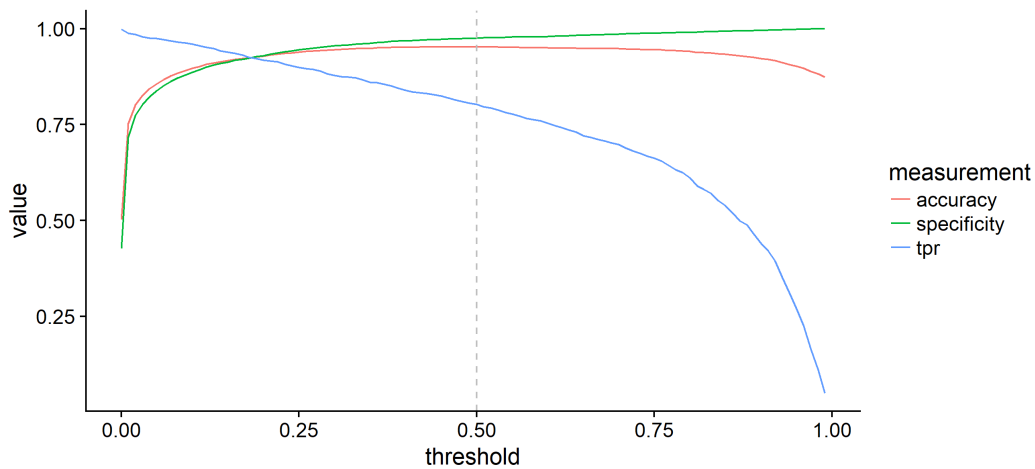


Figure 4.10: snv.predict score effects on statistical outcome measures. Line plot showing how specificity, accuracy and true positive rate change over different SNV prediction values taken from the snv.performance tool (applied to all upper gastrointestinal cells, using VarScan2 SNV calls). Vertical grey dotted line at 0.5.

This feature allows a threshold for a 'true' prediction to be altered. For instance, in an analysis of mutational burden a prediction value of >0.5 may be sufficient in order to maintain the number of SNVs, whereas analysis of specific SNVs may suit a more stringent threshold. Using snv.predict to assign a predictive value to variant calls made by VarScan2, we can begin to look at the mutational landscape in Barrett's oesophagus cells.

4.3.1 Mutational landscape of phenotyped single cells in Barrett's oesophagus

Taking all sequenced cells from four patients with Barrett's oesophagus (BO) together (n=371, patients A-D) and filtering SNV calls from VarScan2 with snv.predict leaves 46962 SNVS (after removing mutations in mitochondrial and unplaced scaffold regions). Further filtering was also performed by removing SNVs present in less than 10 cells and removing SNVs seen in bulk tissue sequencing, to suppress the effects of germline mutations in subsequent analysis and reduce the effect of single cell specific mutations. Hierarchical clustering of these BO cells shows that each patients' genotype remains important in how they cluster (Figure 4.11A), however there are some notable areas of cells that group together between patients (Figure 4.11B). Using the cell phenotypes defined in the previous chapter, it is possible to examine the cells

that group together despite coming from different patients. Interestingly, this suggests that certain cell phenotypes may share some elements of a mutational profile (Figure 4.11C).

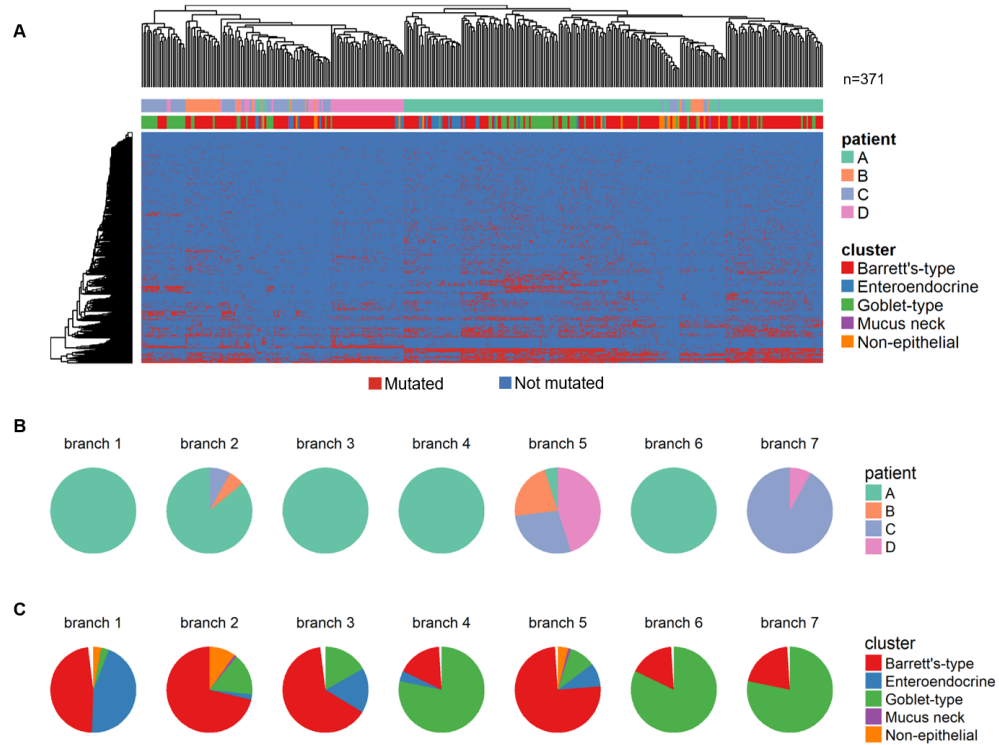


Figure 4.11: Hierarchical clustering of SNVs from four patients with Barrett's oesophagus. (A) Heatmap and dendrogram of SNVs called from single cell RNA-seq data of BO cells from four patients (A-D). SNVs were selected that were present in at least 10 cells and not present in matched bulk tissue RNA-seq. Annotation tracks above heatmap show which patient cells came from (top) and cell phenotype information (bottom, also see previous chapter). (B) Pie charts showing the proportions of cells contributed to each cluster by patient. Clusterings are derived from the dendrogram in A using a dynamic tree cutting method to group cells more robustly. (C) As in B, but showing proportions of cell phenotypes in each cell cluster.

4.3.2 Phenotype heterogeneity exists within the clonal structure of Barrett's oesophagus

Focussing on the patient with the most cells (n=214, patient A), to exclude the genotype effect observed in Figure 4.11A, distinct clusters develop (Figure 4.12A). As bulk tissue SNVs have been excluded, these are enriched for somatic mutations and thus these clusters are likely to represent the clonal structure previously described

in BO (Figure 4.12B, from Ross-Innes *et al.* (Ross-Innes et al., 2015a)). By adding phenotype details, we observe there is significant phenotypic heterogeneity among the clonal architecture but interestingly, clones tended towards a Barrett's-type (red) or goblet-type (green) phenotype. (Figure 4.12C).

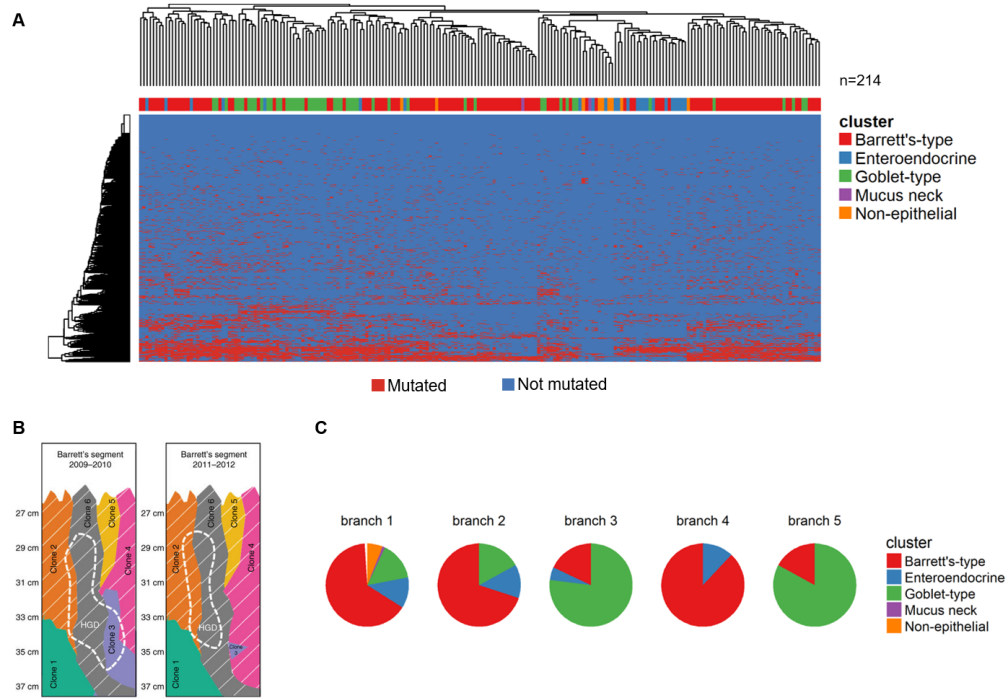


Figure 4.12: Hierarchical clustering of SNVs from a single patient with Barrett's oesophagus. (A) Heatmap and dendrogram of SNVs called from single cell RNA-seq data of BO cells from patient A. SNVs were selected that were present in at least 10 cells and not present in matched bulk tissue RNA-seq. Annotation tracks above heatmap show the phenotype of cells. (B) Example of the clonal structure seen in BO from whole genome and targeted sequencing, taken from Ross-Innes *et al.* (C) Pie charts showing the proportions of cells contributed to each cluster by phenotype. Clusterings are derived from the dendrogram in A using a dynamic tree cutting method to group cells more robustly.

4.3.3 Unique mutations in goblet cells

Branches 4 and 7 in Figure 4.11C and branches 4 and 5 in Figure 4.12C are mainly made of goblet cells. Since the mutability of goblet cells has previously been proposed as a factor in the neoplastic potential of BO, SNVs called from VarScan2 with high `snv.predict` values (>0.9) which were present in goblet cells were isolated to see if there is a direct link between cancer causing mutations and mutations observed in

these goblet cells' mutations. While some of these SNVs were in long non-coding genomic regions, identical mutations in mucin 6 (*MUC6*), trefoil factor 3 (*TFF3*) and anterior gradient 2 (*AGR2*) were detected in all four patients (Figure 4.13).

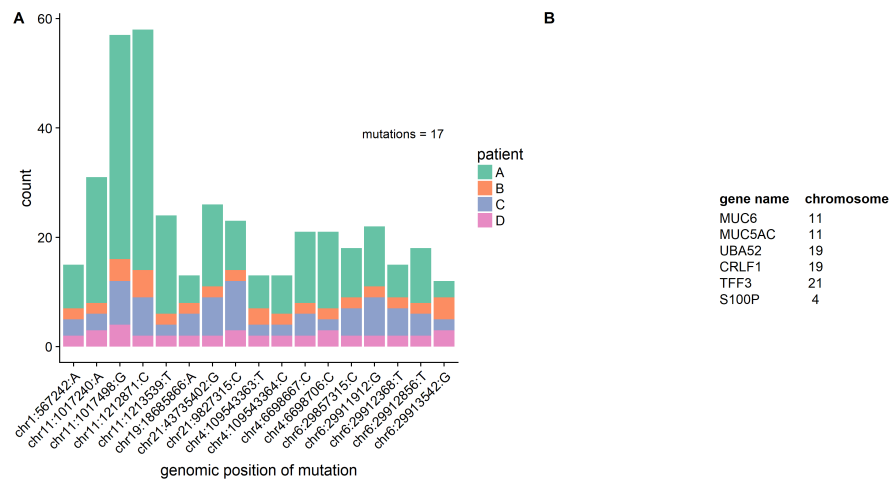


Figure 4.13: Mutations common to goblet cells from four patients with Barrett's oesophagus. (A) Stacked bar plot showing the frequency of SNVs at specific genomic positions in goblet cells from four patients with Barrett's oesophagus. SNVs shown were present in at least two cells from each patient included and had a snv.predict score of >0.9. X axis text show chromosomes, genomic positions and substituted bases, separated by colons. (B) Table showing the genes and chromosome location in which mutations were detected in goblet cells from all four patients. Pseudogenes, mitochondrial, long non-coding region and micro RNAs are excluded from this table, but are present in A hence the discrepancy in length.

4.3.4 Unique mutations in Barrett's oesophagus

Barrett's oesophagus has been described as having a high mutational burden, with single nucleotide polymorphism per megabase rates approaching that of some invasive cancers such as multiple myeloma and breast (Ross-Innes et al., 2015a). This has been shown to further increase in dysplastic BO (Stachler et al., 2015). Patients A-D did not exhibit any pathological features of dysplasia, and overall quantification of SNVs only showed a non significant increase in SNVs in BO when compared with duodenum, gastric or oesophagus cells (Figure 4.14A). Total SNV burdens did vary more significantly between patients for each tissue, for example, there was a higher SNV burden in BO cells from patient A; it will be interesting to see how this correlates in longer term clinical follow up (Figure 4.14B). Using the phenotype details established in the previous chapter (Figure 3.9), SNV burden also did not vary much between phenotypic cell type overall, or by tissue (Figure 4.14C, D).

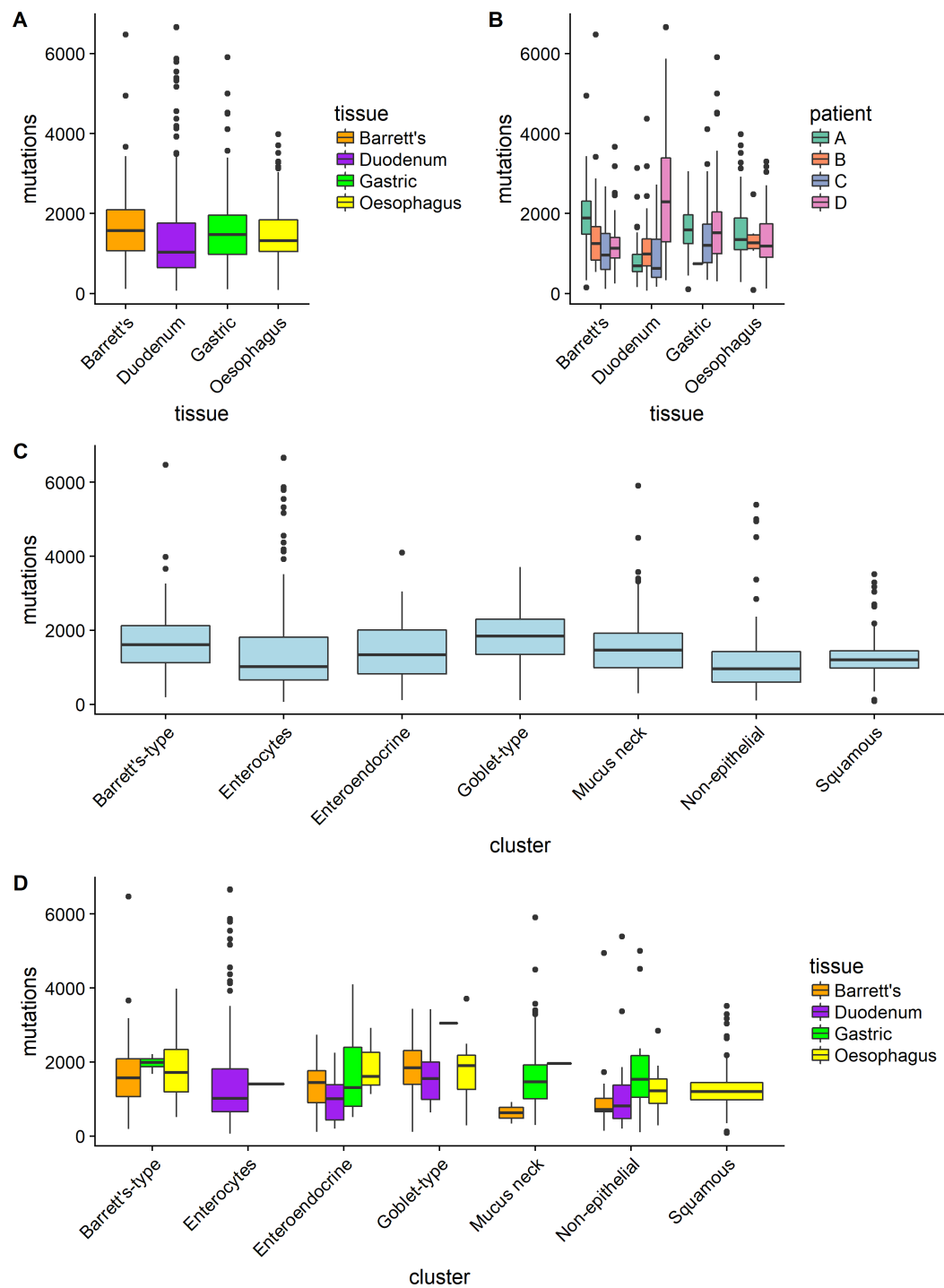


Figure 4.14: Mutational burdens in upper gastrointestinal cells. (A) Boxplots showing the total mutations in single cells from each sampled tissue in four patients (Barrett's oesophagus, duodenum, gastric and normal oesophagus). (B) As in A, but separate boxplots for each patient (A-D). (C) Boxplots showing the total mutation number in upper gastrointestinal cells by cell phenotype, as defined by gene expression in the previous chapter. (D) As in C, but separate boxplots for each tissue type sampled within each phenotype cluster.

Of the SNVs occurring in all patients with BO, 284 SNVs were identified in 84 genes in at least 9 cells from every patient (Figure 4.15A, B). *CDKN2A* mutations were also observed in multiple cells in two patients (A and B), consistent with expected mutations in BO. Interestingly, two cells from patient A also had *TP53* mutations, which were not detectable in bulk sequencing.

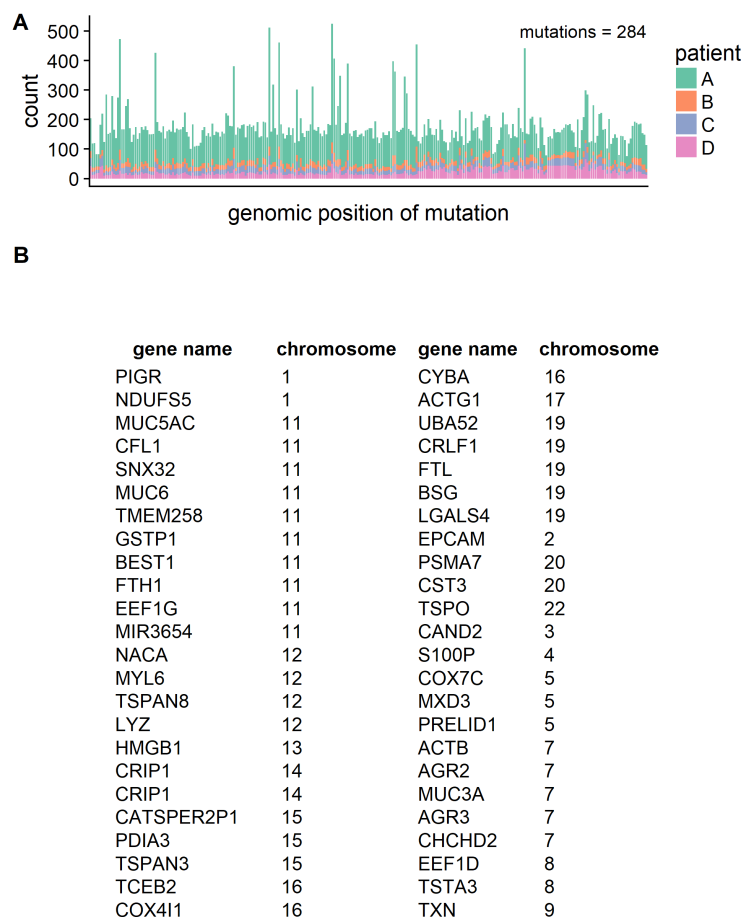


Figure 4.15: Mutations common to four patients with Barrett's oesophagus that were not present in other upper gastrointestinal cells. (A) Stacked bar plot showing the frequency of SNVs at specific genomic positions in Barrett's oesophagus cells from four patients. SNVs shown were present in at least nine cells from each patient included and had a snv.predict score of >0.9. Each X axis position marks the specific genomic positions of each SNV but these are not labelled in this plot due to the overall density of SNVs. (B) Table showing the genes and chromosome location in which mutations were detected in goblet cells from all four patients. Pseudogenes, mitochondrial, long non-coding region and micro RNAs are excluded from this table, but are present in A hence the discrepancy in length.

Next, SNVs specific to BO and not present in control tissues were next examined to understand if any BO driver mutations might exist. Mutations identified were present in all four patients with BO and not present in any other upper gastrointestinal single cell or bulk RNA-seq data, to suppress germline mutation detection and increase the likelihood of identifying mutations that were biologically important in BO. This identified 135 unique BO SNVs in 31 genes in at least one cell, including *S100P*, which was also mutated in goblet cells (Figures 4.13), suggesting that mutated *S100P* in BO could be specific to intestinal metaplasia, though gastric-type BO with no intestinal metaplasia would be needed as a further control to confirm this (Figure 4.16).

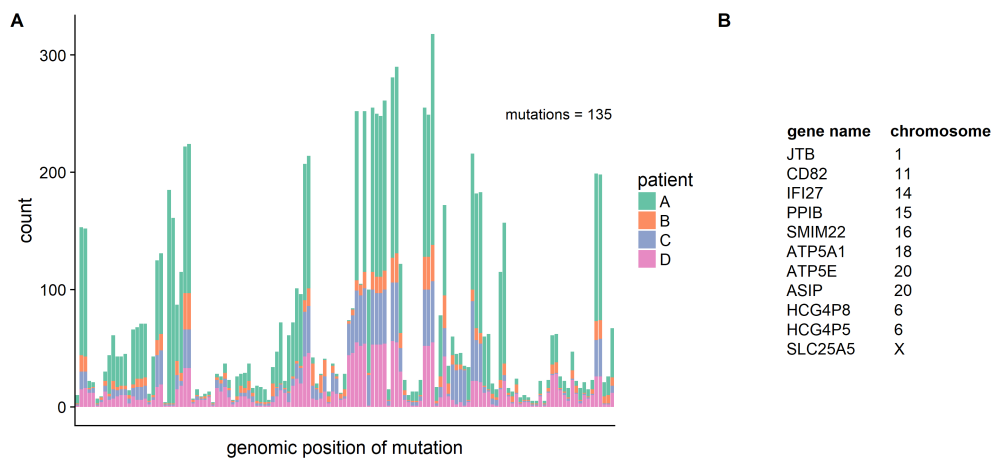


Figure 4.16: Mutations common to four patients with Barrett's oesophagus. (A) Stacked bar plot showing the frequency of SNVs at specific genomic positions in Barrett's oesophagus cells from four patients. SNVs shown were present in at least one cell from every patient included and had a snv.predict score of >0.9. SNVs were excluded that were present in duodenal, gastric or oesophagus cells, and in matched bulk RNA-seq, to suppress germline mutation detection. Each X axis position marks the specific genomic positions of each SNV but these are not labelled in this plot due to the overall density of SNVs. (B) Table showing the genes and chromosome location in which mutations were detected in goblet cells from all four patients. Pseudogenes, mitochondrial, long non-coding region and micro RNAs are excluded from this table, but are present in A hence the discrepancy in length.

4.3.5 Mutational signatures in Barrett's oesophagus

While specific mutations may prove interesting, this will require further validation. More general observations of the mutational types in BO may elucidate fundamental understanding of the aetiology of BO (Alexandrov et al., 2013). To do this, mutational signatures were generated from SNVs called in BO cells from all four patients

(Gehring et al., 2015). Grouping by cell phenotype clusters identified two mutational signatures that accounted for most of the variance over the data (Figure 4.17A). Interestingly, contributions to these signatures varied by phenotype (Figure 4.17B), with the mucus neck cluster contributing most to signature 1 (S1), the non-epithelial cluster contributing equally to S1 and S2, and the other phenotype clusters (goblet-type, enteroendocrine and Barrett’s-type) contributing to S2. Inspection of these mutational signatures shows dominant T>C mutations in a G:C context in both signatures, but S2 also shows C>T mutations in broader contexts (Figure 4.17C), as seen in OAC (Secrier et al., 2016). The effect of filtering out potential RNA editing sites when setting the control parameters must be considered here, as only A>G sites were excluded from positive control SNVs, and thus T>C mutations could represent RNA editing on the antisense strand. Despite this, visual comparison of these signatures to previously defined mutational signatures in cancer shows similarity with Alexandrov signatures 21 and 26, both seen in gastric cancer and the latter associating with DNA mismatch repair (Figure 4.17C) (Alexandrov et al., 2015, 2013). The goblet-type cell cluster matches Alexandrov signature 26 (gastric cancer), and cancer of the gastric cardia frequently exhibits intestinal histology and chromosomal instability (Cancer Genome Atlas Research Network, 2014, 2017). Interestingly, the Barrett’s-type cluster, mucus neck and non-epithelial clusters grouped separately from goblet-type and enteroendocrine clusters in hierarchical analysis, based on mutational profile (Figure 4.17D).

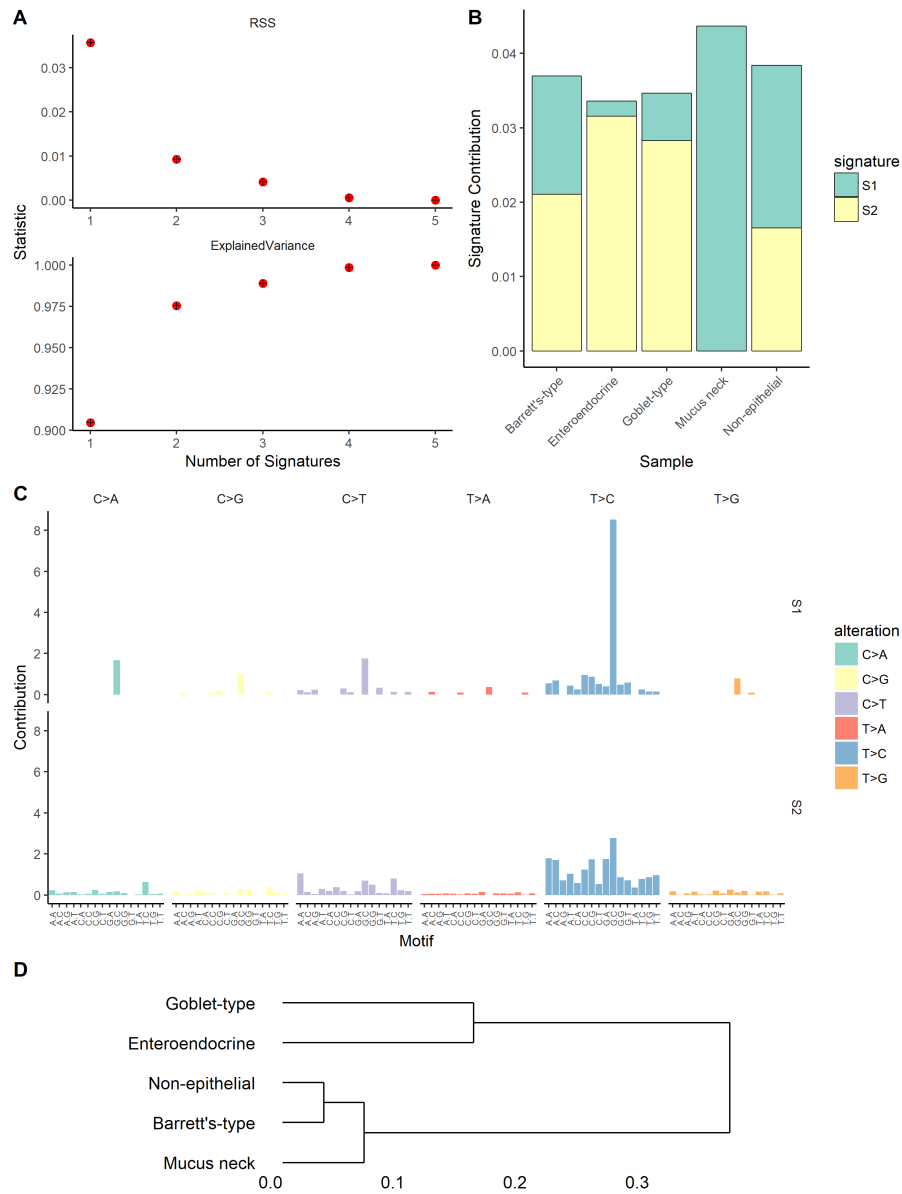


Figure 4.17: Mutational signatures in Barrett's oesophagus cells from four patients. (A) Dot plot showing how variance and root sum of squares (RSS) changes with increasing number of mutational signatures. At four signatures, no further increase in variance is seen therefore this is taken as n for further analysis. (B) Stacked bar plot showing how each phenotype (x axis) contributes to each identified mutational signature (colour legend). (C) Histograms showing the frequency of each SNV in any mutational context. Mutational signatures were grouped by matrix decomposition into 2 signatures (plot rows). Plot columns are arranged by base substitution and colour coded. (D) Dendrogram showing hierarchical analysis of cell phenotypes by mutational signature.

To check whether the heterogeneity in the Barrett's-type cluster BO showed different

4.3.6 Genomic links between Barrett’s oesophagus and oesophageal submucosal glands

In view of results in the previous chapter showing that OSGs may harbour the cellular niche from which BO generates, SNVs were isolated that were present only in cells from the Barrett’s-type phenotyped cluster identified in Figure 3.9. These cell groups comprise of BO and oesophageal submucosal gland (OSG) cells. To limit genotyping bias, only cells from patient A were included. SNVs were excluded that were present in any other tissue, and any other patient. Hierarchical clustering of SNVs shows that while most Barrett’s and oesophageal cells clustered independently (Figure 4.19A), despite their shared phenotype, some mutations were shared between the different tissue types supporting earlier observations of an OSG-BO relationship (Figure 4.19B). Overall, identical SNVs in 395 genes were present in both BO and OSG cells, with some occurring in the same genes multiple times (*CFB*, *UQCR11*). Unfortunately, there are insufficient OSG cells from other patients to perform a similar analysis, but it will certainly be intriguing to see if any of the same genes are mutated when this analysis can be replicated in more patients.

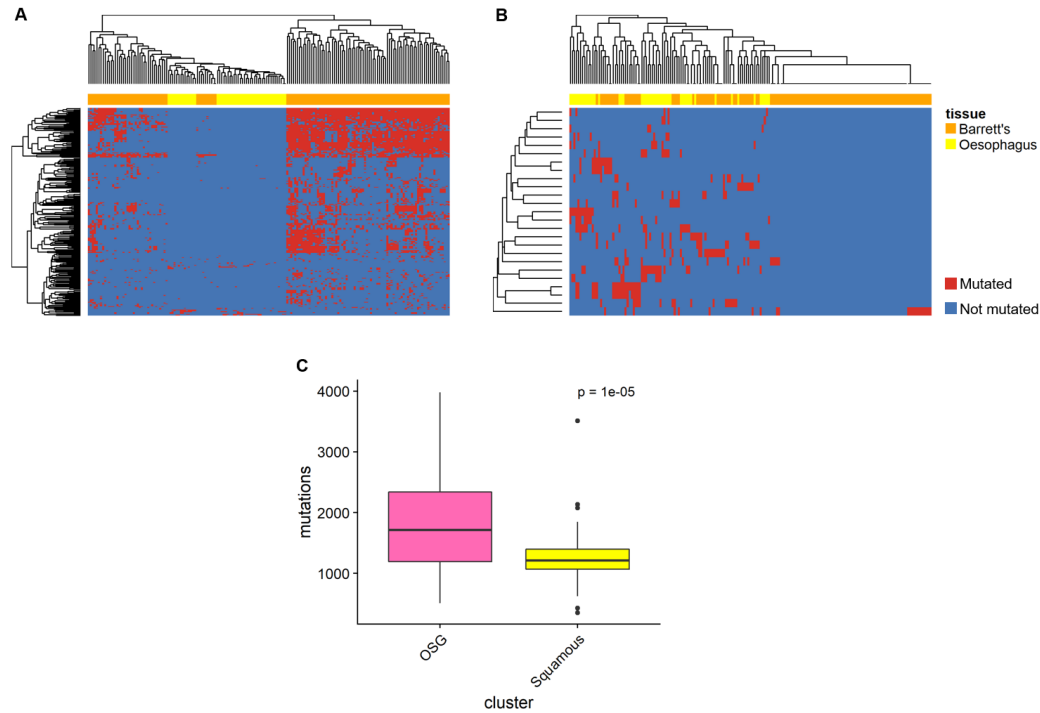


Figure 4.19: Genomic linking of Barrett's oesophagus and OSGs from a single patient. (A) Heatmap and dendrograms showing the hierarchical clustering of tissues (columns) by SNVs (rows) which were unique to Barrett's oesophagus and oesophageal cells present in the 'Barrett's-type' cluster previously defined. SNVs from matched bulk RNA-seq and from other tissues and patients were removed, to limit the effects of germ line mutations. (B) As in A but SNVs are restricted to only those seen in OSG cells. (C) Boxplot showing the number of SNVs in each oesophageal cell from patient A, labelled as squamous (yellow) or Barrett's-type (pink). P value shown is from t test.

Finally, in view of the previous observations of high mutational burdens in BO, total SNVs in squamous or Barrett's-type oesophageal cells from patient A were compared. This shows that Barrett's-type oesophageal cells have a significantly higher mutational burden than squamous cells in the same patient (Figure 4.19C).

4.4 Generalising SNV prediction models

In all data examined, tuned random forest machine learning appeared to outperform naive calling, logistic regression and support vector machine learning. Therefore tuned random forest machine learning has been used alone in further analysis. Since many publicly available datasets lack ERCC spike ins or reliable positive controls, 'cross.predict' was developed to investigate whether modelling on one data set could

be applied to another, and maintain the substantial improvements in SNV calling possible with `snv.predict`. Figure 4.20 shows a schematic of how `cross.predict` was applied in the upper gastrointestinal and K562 cells.

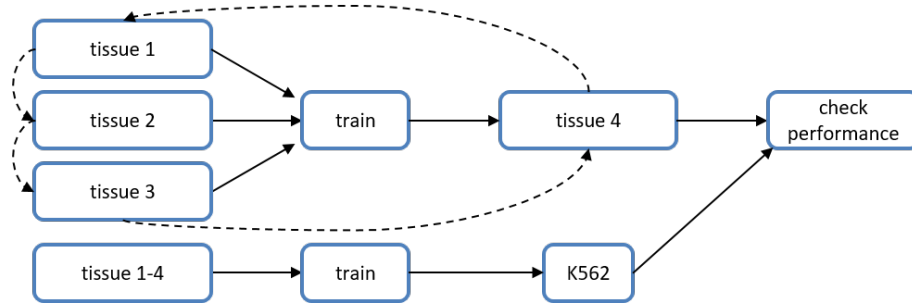


Figure 4.20: Outline of `cross.predict`. Schematic showing how `cross.predict`, a module within `snv.predict`, uses existing datasets to model SNV predictions before applying the model to a dataset with true positives and false positives already defined, to assess performance between tissues or cell types. Dashed lines show how cells from each tissue were rotated for each iteration of analysis.

Applying `cross.predict` in this way shows that the performance of SNV predictions was high (Figure 4.21). In particular, AUC and accuracy were high between columnar upper gastrointestinal cells (duodenum, gastric and BO). Performance in K562 cells and oesophageal cells was not as strong but still demonstrated high AUC and accuracy, presumably due to the more phenotypically different nature of these cells (squamous and myelogenous respectively). Notably, the specificity was very high with fairly modest thresholds (0.5) for calling predicted SNVs (Figure 4.21B).

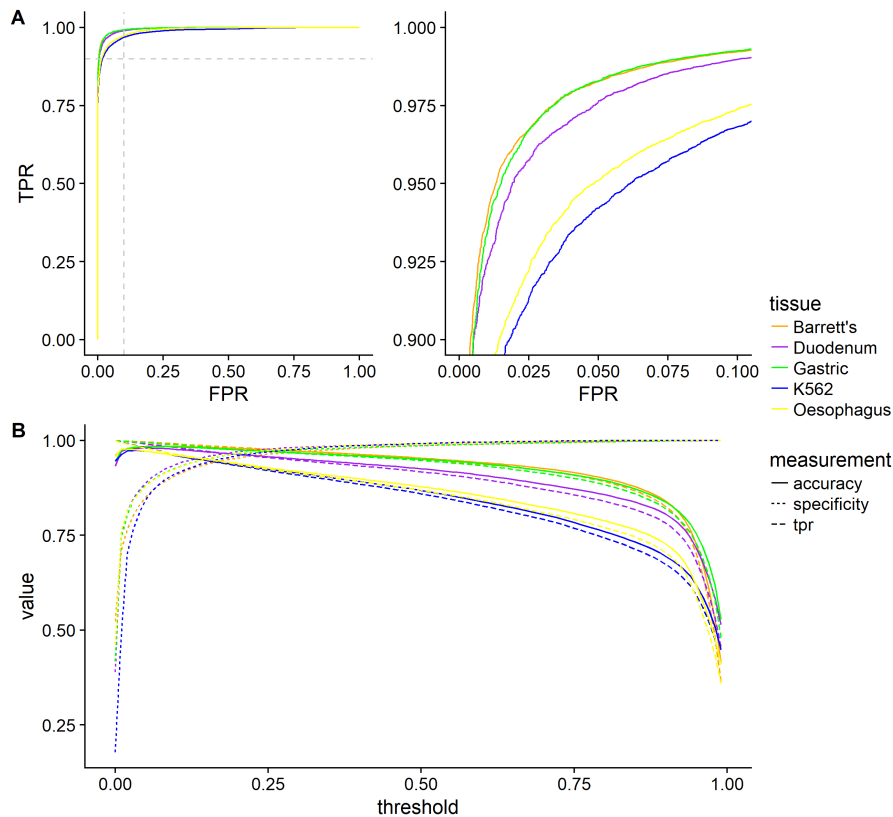


Figure 4.21: Performance of cross.predict in SNV calling between tissue and cell types. (A) ROC plot showing the performance of SNV calling in single cell RNA-seq data with machine learning enhancement. Each line corresponds to a tissue type in which predictions were made using a model trained by other cell types (see Figure 4.20). The right plot shows a partial scale, as indicated by the grey dotted line bounded area in the full scale left plot. VarScan2 was used for the baseline SNV calls and a tuned random forest model to enhance calling. (B) Line plot showing how true positive rate, accuracy and specificity change depending on prediction score output by cross.predict, a module in snv.predict.

4.5 Single cell mutations in breast carcinoma

Using cross.predict in a breast carcinoma data set, which lacked suitable spike-in controls for defining false positive SNVs for snv.predict, allowed quantification of mutations between tumour subtypes in 10 patients. This showed highly variable total SNVs between cells, even from the same patient and tumour (Figure 4.22A). Interestingly, I identified that the total predicted SNVs in cells was substantially reduced from primary breast carcinoma to lymph node carcinoma cells in two patients where these matched samples were available (Figure 4.22B, BC03 and BC07). This reduction in predicted SNVs was also observed in non carcinoma cells from the same

patients (Figure 4.22C). Whilst this reduction in predicted SNVs from primary to metastatic cancer was unexpected, many SNVs were identified in lymph node cells that were not present in any of the 10 patients' primary or recurrent cancer tissues, including a unique *BRCA2* mutation in one patient. Whilst no SNVs were identical in both patients lymph node cells (and not in any primary breast cells), SNVs in four genes were uniquely mutated in lymph node samples (*ERCC1*, *UTRN*, *SETX*, *CD3EAP*), which could suggest these genes are implicated in the disease spread.

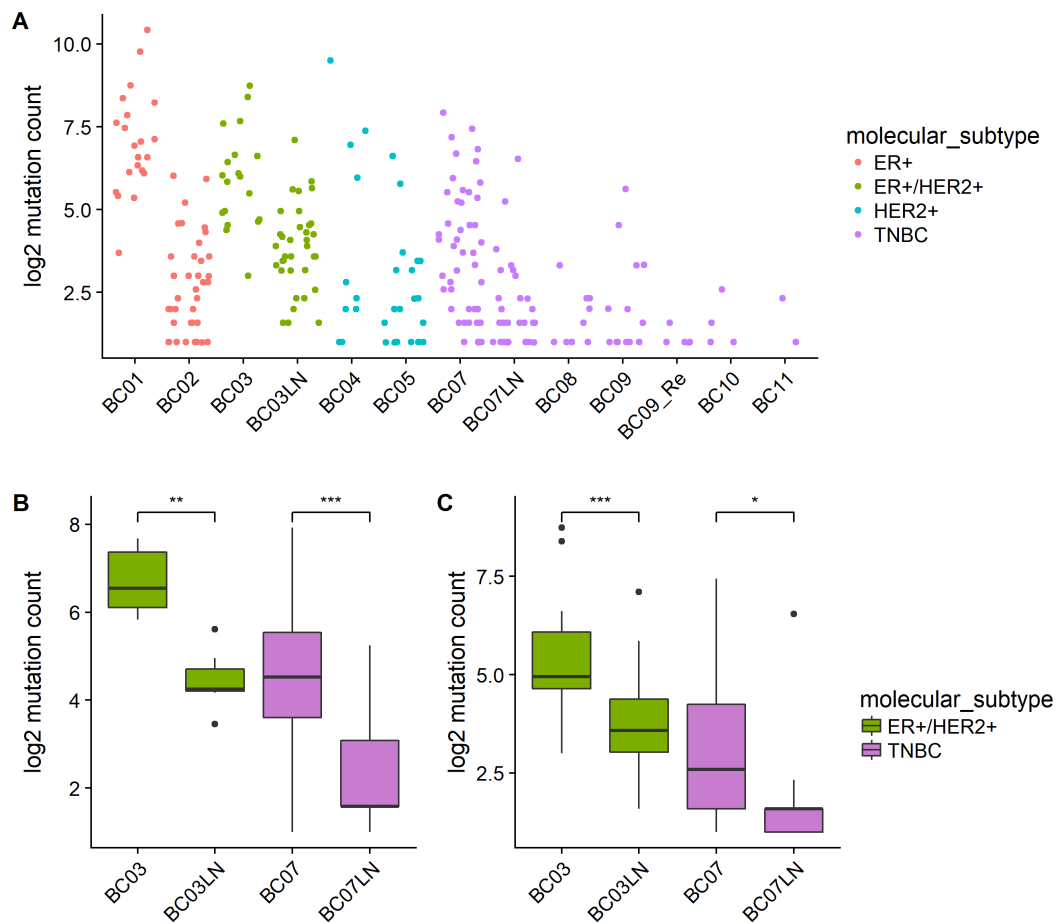


Figure 4.22: Mutational burdens of malignant and non-carcinoma cells from breast cancers with different molecular subtypes. (A) Jitter plot showing the total mutations in single cells from breast carcinoma in 10 patients. (B) Boxplot showing the total number of mutations in cancer cells sampled from primary tumour and from lymph nodes. Cells were labelled as cancer cells using and inferred copy number variation (CNV) algorithm. (C) As in B but applied to non carcinoma cells as labelled from inferred CNV. LN=lymph node; Re=recurrence; ER=oestrogen receptor status; HER2=human epidermal growth factor receptor 2 receptor status; TNBC=triple negative breast carcinoma.

Finally, hierarchical clustering of the breast carcinoma cells shows that these cells

shared very few SNVs, indicating the mutational heterogeneity and relatively scant number of SNVs in these data (Figure 4.23).

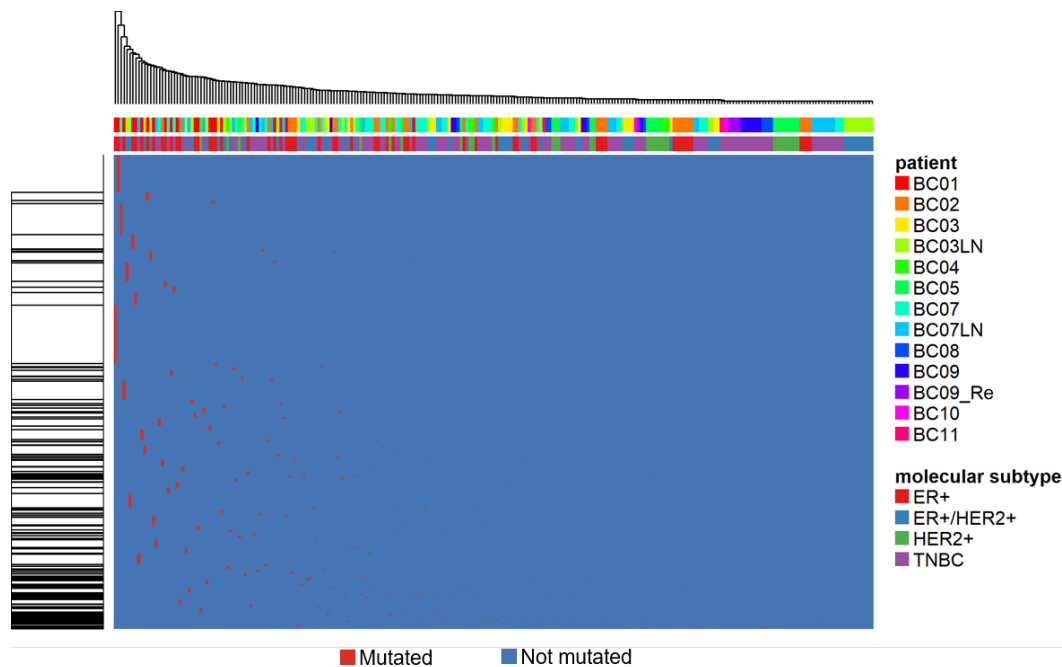


Figure 4.23: Mutational relationships between breast cancer cells. Heatmap and dendrogram of SNVs called from single cell RNA-seq data of breast carcinoma. Annotation tracks above heatmap show the patient origin of each cell and the molecular subtype of a given carcinoma.

4.6 Summary

In this chapter I have introduced an R package (`'snv.predict'`) designed to take single nucleotide variant (SNV) calls made in single cell RNA-seq data and exploit artificially spiked in sequences to train a model to predict whether a SNV is true. Using `snv.predict` I demonstrate the clonal nature of BO, show that BO clones have different phenotypic compositions and identify mutations unique to cell phenotypes. Finally, I show that `snv.predict` can be generalised between cell types and use publicly available single cell RNA-seq data to show how mutational burdens alter in breast cancer between primary and metastatic sites.

Chapter 5

Enhancements for the clinical application of single cell RNA-seq

Whilst single cell sequencing approaches have offered exciting insights and prospects in biology, many technical and practical challenges exist which have restricted the translation of laboratory based sequencing assays to *ex vitro* human applications (Saliba et al., 2014). In this chapter, I present the results of experiments designed to enhance the application of single cell RNA-seq to human tissues by addressing throughput, affordability and practicality. This is then put into practice by sequencing cells from an oesophageal cancer as part of a clinical trial investigating the safety of immunotherapy in the treatment of oesophageal cancer.

5.1 Tissue preservatives and single cell RNA sequencing

A key issue in processing single cells for plate based single cell RNA-seq is time. RNA degradation is accelerated in the conditions required for enzymatic tissue digestion thus a cell's transcription profile is at risk of changing due to temperature, chemical or nutritional changes once a cell is stripped from its normal environment (Kim et al., 2010). This effect is likely to be exacerbated when obtaining single cell suspensions from organised or fibrous tissues as a thorough digestion step is needed to reduce selection bias in the types of cells that are selected, whilst also trying to reduce the survivorship bias of less vital cell types.

In human tissue uses, the ideal solution would be a cheap and available reagent or method that quickly stabilises RNA in bulk samples and is then compatible with tissue disaggregation, cell isolation, RNA extraction, library preparation and sequencing. Several products exist to help stabilise RNA, though none have been applied to intact whole tissue samples before processing for single cell RNA-seq.

5.1.1 RNA stabilisation

5.1.1.1 Cryopreservation

Snap freezing tissue samples in liquid nitrogen or dry ice is the gold standard method of preserving tissue at the bedside prior to sample transfer to a longer term cold storage facility for later DNA, RNA or protein extraction. Freezing can sacrifice the cell structure, which is not important in obtaining DNA or RNA for bulk sequencing. This is less suitable for single cell RNA-seq where it remains important that RNA from each cell is kept separately. Slow-freezing methods commonly used in cell culture, with 10% Dimethyl sulfoxide (DMSO) supplemented freezing medium to prevent ice crystal formation and cell lysis, may be more suitable for single cell RNA-seq, but whether this can be applied to tissues for later cell sorting and RNA-seq remains to be seen (Guillaumet-Adkins et al., 2017).

5.1.1.2 *RNAlater*

RNAlater (QIAGEN) is a product designed for the preservation of RNA in bulk samples, which are then stored at ambient temperature or frozen for longer term

storage (Florell et al., 2001). This method of cell preservation has not been popular in single cell RNA-seq due to the high salt content of *RNAlater*, leading to cell shape changes and lysis, thus causing problems with any downstream application which requires intact cells of a consistent size, such as automated sequencing platforms like the Fluidigm C₁TM. To test this, 2mm biopsies were obtained from mouse stomach epithelium using endoscopic biopsy forceps (Radial Jaw 4 Standard Capacity, Boston Scientific, Natick, USA) and stored for 48 hours in *RNAlater* prior to disaggregation and cell sorting with fluorescence activated cell sorting (FACS). These FACS plots showed that recorded events (side scatter area and forward scatter area) consistent with epithelial cells, as seen in freshly processed tissue controls, are heavily diminished and cannot be reliably sorted (figure 5.1). These results did not justify ongoing use of *RNAlater* in subsequent experiments as a cell population suitable for sequencing was not identified.

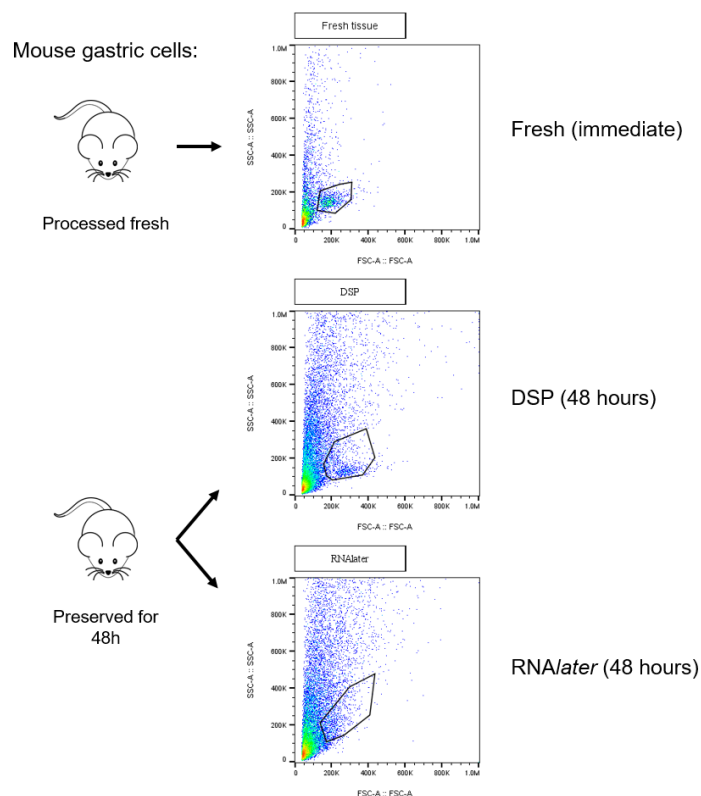


Figure 5.1: FACS plots of cells from mouse stomachs preserved in different ways. Gastric biopsies obtained from wild type mice were processed by immediate disaggregation then sorting (plot labelled fresh), storage in DSP for 48 hours prior to RNA crosslink reversal followed by disaggregation and sorting (plot labelled DSP) or storage in *RNAlater* for 48 hours followed by disaggregation and sorting (plot labelled *RNAlater*). The black polygons in the FACS plots outline the events consistent with epithelial cells which would be selected for sorting into plates in a single cell RNA-seq experiment.

5.1.1.3 Dithio-bis(succinimidyl propionate)

Dithio-bis(succinimidyl propionate) (DSP) or Lomant's reagent is a cell-permeable fixative which reversibly cross-links RNA. DSP has been previously used to preserve tissue prior to immunostaining, microdissection and RNA expression profiling (Xiang et al., 2004). More recently, DSP has been used successfully in preserving cell lines for single cell RNA-seq (Attar et al., 2018). To test whether this could be applied prior to tissue disaggregation, tissue biopsies obtained from mouse stomach were stored in DSP for 48 hours at 4°C followed by tissue disaggregation and sorting. As seen in figure 5.1, there was a large increase in debris produced from delayed processing of the tissue sample, however events are recorded which are consistent with the typical

epithelial cell population seen in fresh processed tissue. A potential limitation of DSP is that its use has only been proven when storing cells for up to 7 days at 4°C, which may be inadequate for clinical usage, especially in longitudinal studies where samples would ideally be processed at the same time but could be collected weeks or months apart. In view of this, a further condition was added in subsequent experiments which RNA was reversibly cross-linked with DSP and then the tissue sample was also slow-frozen with the aim of enabling longer term storage beyond 7 days.

5.1.2 Data quality with RNA preservatives

Tissue biopsies were endoscopically obtained from human stomach and processed fresh or preserved (slow-frozen, DSP treated or DSP treated and slow frozen). DSP treatments were carried out in accordance with published methodology (Attar et al., 2018). Preserved biopsies were stored for 48 hours. All biopsies were then disaggregated before sorting into wells of 96 well plates with removable well columns, to allow permutation of the cell positions to perform a more balanced experiment, thus avoiding the batch effects which would be encountered by processing cells preserved in one condition per plate (Hicks et al., 2017; Bacher and Kendzierski, 2016) (figure 5.2).

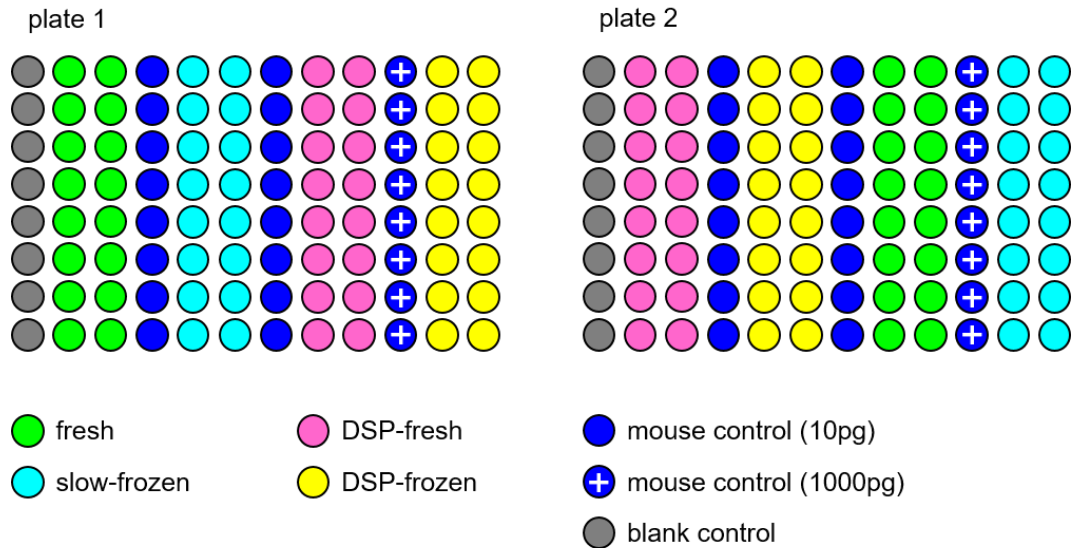


Figure 5.2: Sample permutations in multi-well plates to avoid batch effects. Final well positions after permuting the columns of 96 well plates with removable well columns. Wells contained sorted cells which were processed fresh or preserved by slow-freezing, DSP, or DSP and slow-freezing. Blank columns contained volume matched water and positive control wells were supplemented with mouse RNA from multiple tissues.

Plate based Smart-seq2 was then carried out for cDNA conversion, and Nextera XT (Illumina®) library preparation and sequencing (Illumina® Hiseq 4000). Data were mapped with STAR and gene counts obtained with htseq-count. Mean depth (total counts) per cell was 1.7×10^5 , 1.3×10^5 , 0.9×10^5 and 1.1×10^5 for fresh, slow-frozen, DSP-fresh and DSP-frozen cells respectively. Mean number of genes detected (at least one count per gene) per cell was 597, 1683, 527 and 553 genes for fresh, slow-frozen, DSP-fresh and DSP-frozen cells respectively (figure 5.3).

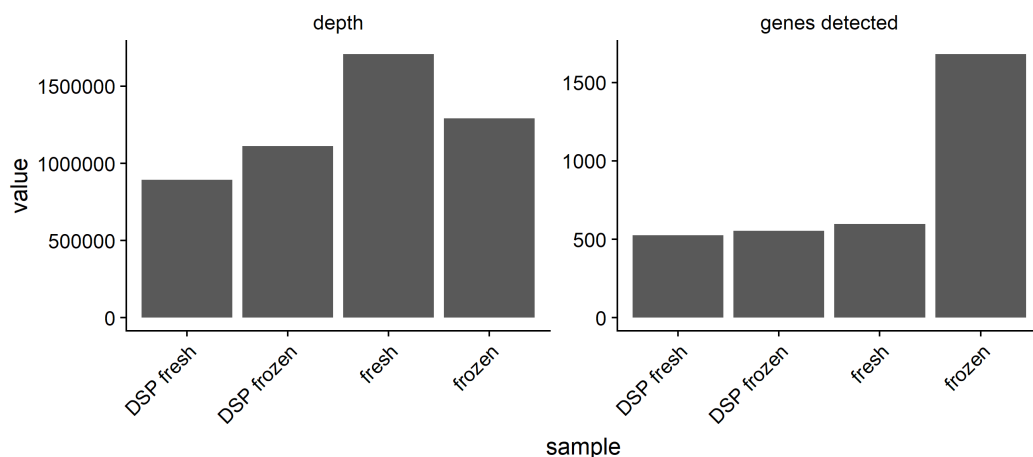


Figure 5.3: Overall performance of single cell RNA-seq with cells from tissues preserved in different ways. Bar plots showing mean depth (total number of gene counts, left side plot) and mean number of genes detected (with at least one count per gene, right side plot) in single cell RNA-seq data from human gastric biopsies either processed fresh, or preserved first by slow-freezing, treatment with DSP or treatment with DSP combined with slow-freezing.

Analysis of individual cells showed there was considerable spread of results, suggesting the potential for better quality data in each experimental condition (figure 5.4 A-B). The distribution of number of genes detected further demonstrates the better profile observed in the fresh and slow-frozen cell preservation conditions, with the DSP treated and slow frozen cells showing little improvement in gene detection than observed in water controls (note that all samples have ERCC spike-in controls, hence genes are detected in water controls) (figure 5.4 C).

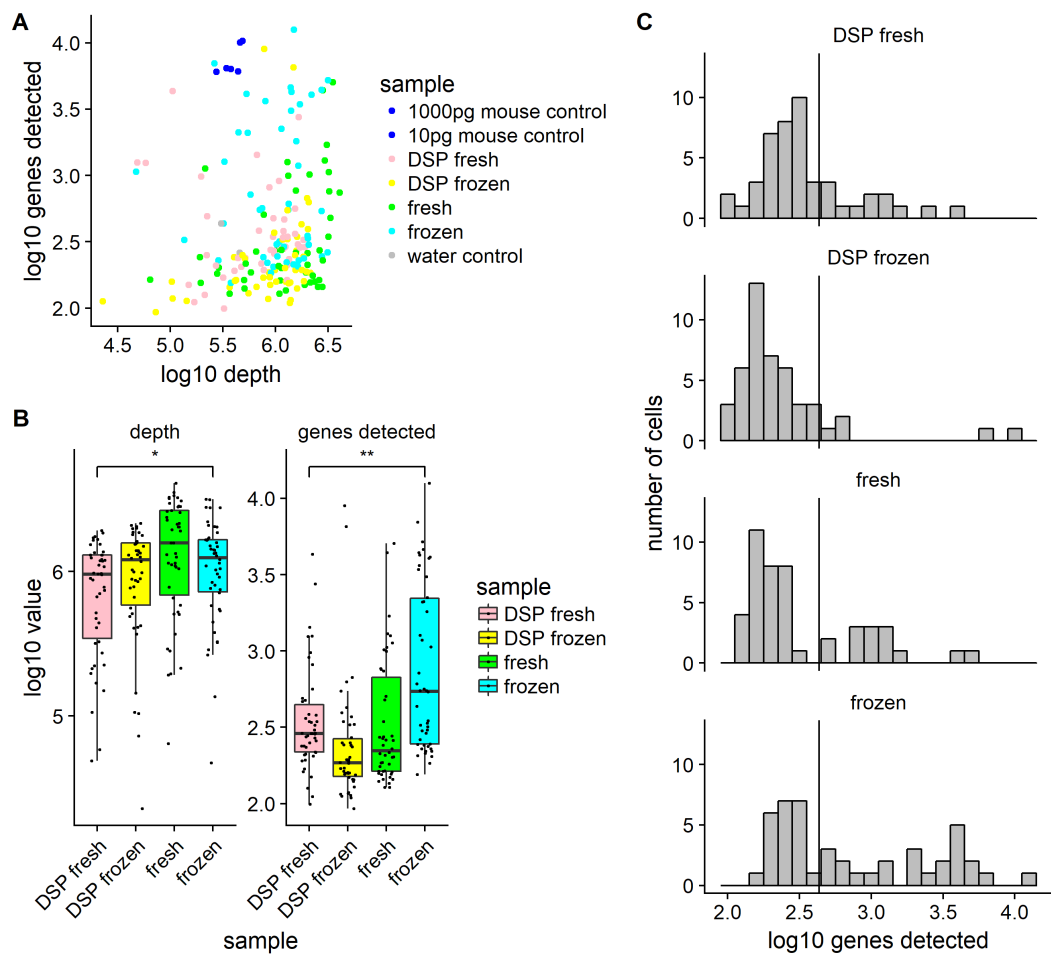


Figure 5.4: Individual cell performance of single cell RNA-seq with cells from tissues preserved in different ways. (A) Scatter plot showing log transformed depth and number of genes detected in individual samples including controls, coloured by sample type and processing conditions. (B) Boxplots showing depth and number of genes detected across all cells sequenced in each condition, jittered points show results for each individual cell. $p < 0.05 = *$; $p < 0.01 = **$; $p < 0.001 = ***$. (C) Histograms showing number of cells in which a given number of genes were detected (at least one count per gene). Vertical black lines indicate the maximum number of genes detected observed in a water control, and thus can be considered as a cut off of poor quality data.

As cells were processed with a consistent disaggregation protocol, batch bias was minimised, and sorted cells were processed together and pooled prior to sequencing, this would suggest that the preservative conditions were the largest influence on data quality. In the fresh and slow-frozen cells, the highest expressed genes were consistent with those expected in human gastric tissues, with the gastric genes *progastricsin* (*PGC*) and *trefoil factor 1* (*TFF1*) both highly expressed (figure 5.5).

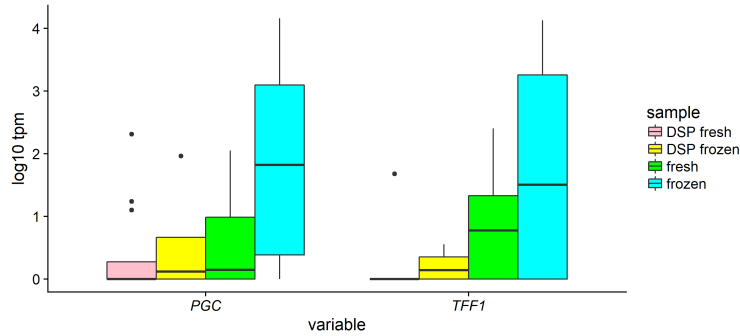


Figure 5.5: Expression of known marker genes in cells from tissues preserved in different ways. Boxplots showing the expression of the gastric genes *PGC* and *TFF1* in cells which were processed fresh or frozen for single cell RNA-seq with and without the addition of the reversible RNA crosslinking agent DSP. tpm=transcripts per million.

These results suggest that DSP treatment worsened the number of genes detected and depth obtained from sequenced human gastric cells in contrast to previous data showing improved results when applied to K562 cell suspensions (Attar et al., 2018). The poor performance of DSP in this setting may be due to extended incubations required to de-crosslink RNA, perhaps exposing RNA to an extended period of degradation in both fresh-DSP and frozen-DSP experimental conditions. Interestingly, data produced from the slow-frozen biopsies showed enhanced gene detection compared to all other conditions, including freshly processed, despite a reduced depth. This may be due to the ability to begin the freezing process at the bedside, rather than transporting samples to a facility for disaggregation and sorting, during which time RNA may degrade and thus impair the quality of the final data. A similar rationale has been made previously for the use of DSP.

5.2 Accelerating throughput with robotic sample preparation

Plate based sequencing methods have relied on manual preparation. This has been adequate in smaller experiments, especially where flexibility is needed such as in multiple condition cell line work, but most characterisation based single cell projects require consistency across plates to allow a larger number of samples to be tested without introducing preparation error. Additionally, greater speed is important to batch processing of larger sample numbers. After initial manually prepared experiments,

we developed a robotic process to enable up scaling of single cell sequencing experiments. Sample processing was divided into pre-polymerase chain reaction (PCR) and post-PCR using two robots (figure 5.6). Pre-PCR steps were carried out using a Biomek[®] FXP Laboratory Automation Workstation to automate cDNA synthesis (Smart-seq2) and sequencing library construction (Nextera XT) from cell lysate taken from human cells. Post-PCR clean-up, pooling and library QC were all automated using a Biomek[®] NXP Laboratory Automation Workstation.

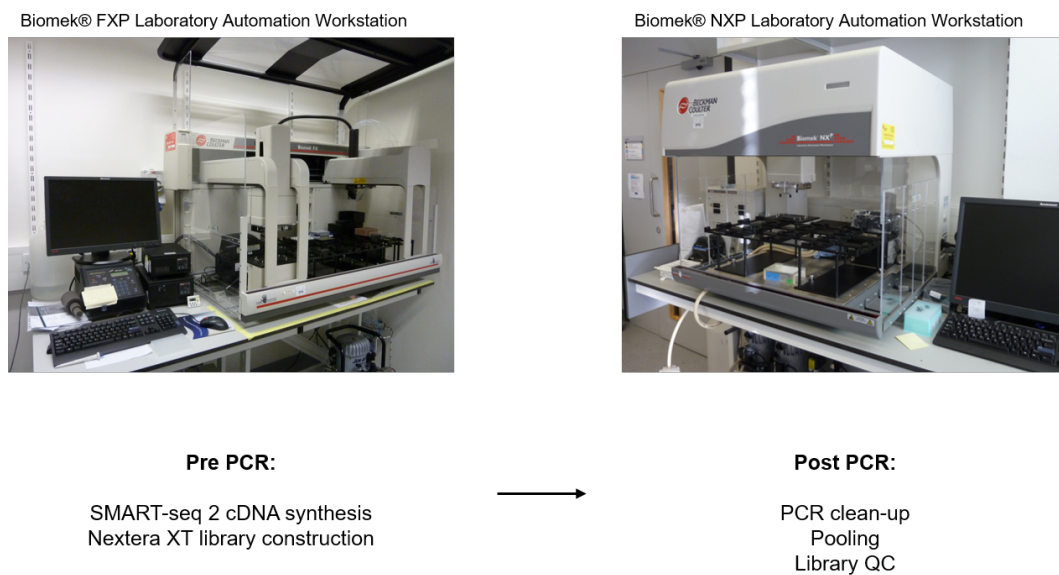


Figure 5.6: Robots used for single cell RNA-seq. Pictures of the two robots used for SMART-seq 2 cDNA synthesis, Nextera XT library construction, PCR clean-up, sample pooling and library QC.

A total of 171 Barrett’s oesophagus cells were prepared for sequencing (75 manually and 96 robotically) so a direct comparison could be made of data quality. Following sequencing using Hiseq4000 (Illumina[®]), data were mapped and gene expression data prepared using STAR and HTSeq as described in Materials and Methods (data were normalised post-processing to account for differing sample sizes). Comparing the mapping statistics manually and robotically prepared samples suggests that robotic preparation was significantly more efficient at producing uniquely mapping reads from the same number of total mapped reads (figure 5.7A, C). However, on examining gene count data there was little difference in total counts or the number of genes detected (at least one count per gene) between manual and robotic sample preparation (figure 5.7B, D-E).

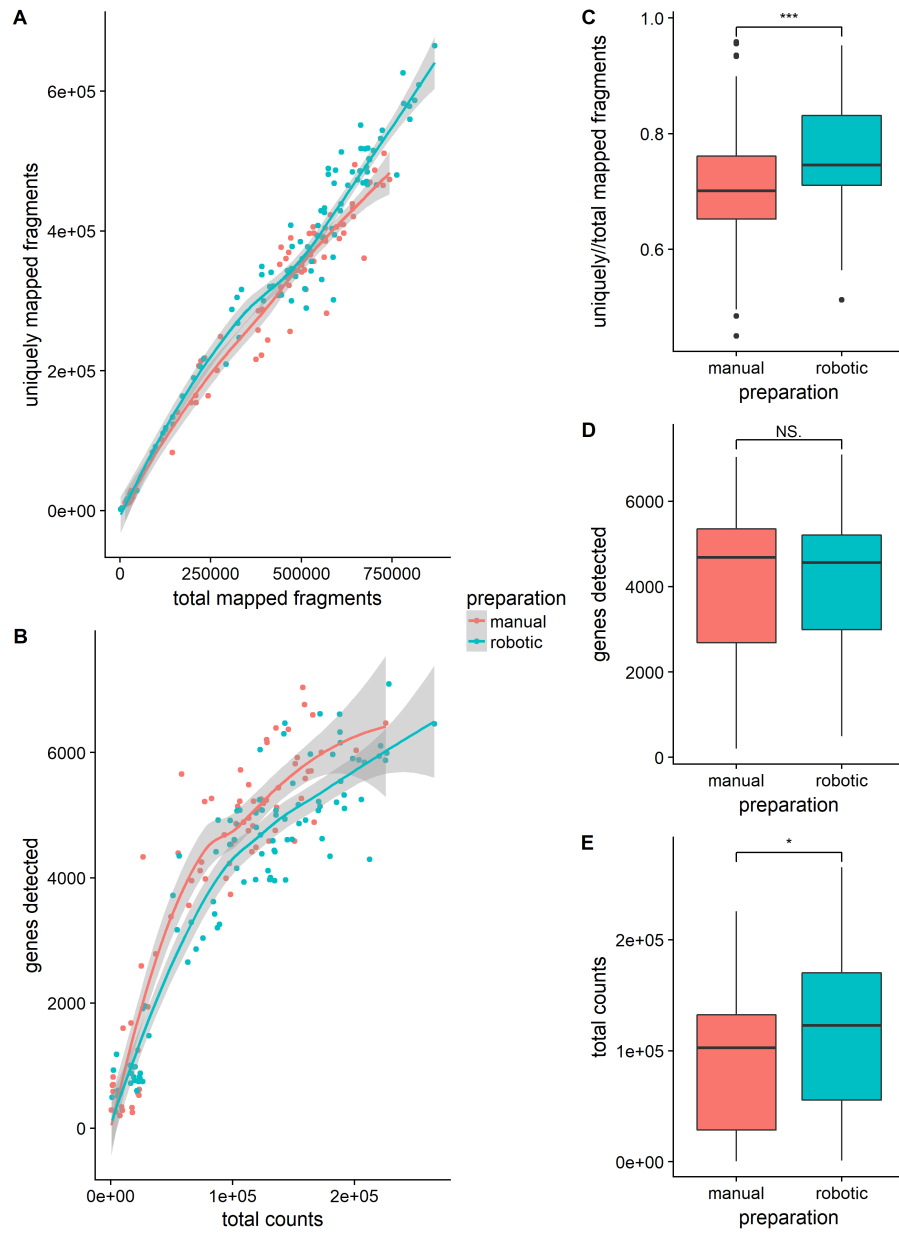


Figure 5.7: Comparison of manual and robotic preparation of single cells for RNA-seq. (A) Scatter plot with smooth curve comparing the total number of mapped fragments against uniquely mapped fragments. (B) Scatter plot with smooth curve comparing the number of genes detected (at least one count per gene) against total number of gene counts obtained. (C) Boxplot showing proportion of uniquely mapped fragments to total mapped fragments by sample preparation method (manual or robotic) (D) Boxplot showing the number of genes detected (at least one count per gene) by sample preparation method (manual or robotic, NS.=not significant) (E) Boxplot showing total number of gene counts by sample preparation method (manual or robotic). $p < 0.05 = *$; $p < 0.01 = **$; $p < 0.001 = ***$; not significant = NS.

Since little difference was observed between robotic and manual sample preparation

in processed data (total counts and number of genes detected), a more detailed analysis on a gene to gene basis was performed, also showing extremely high correlation between preparation methods (figure 5.8A). This correlation was maintained when comparing a reduced set of data made up of artificially spiked-in ERCC sequences (figure 5.8B). Comparing total counts across individual chromosomes also did not show any discrepancies (figure 5.8C).

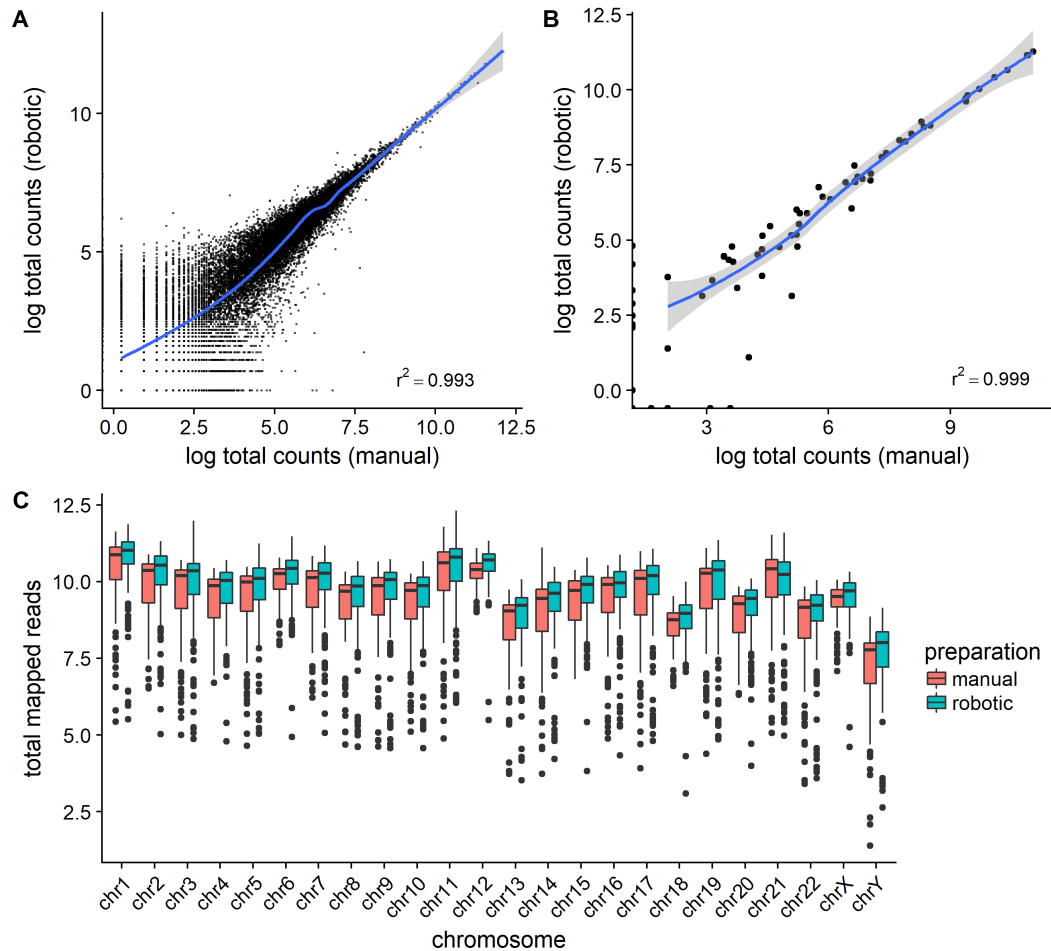


Figure 5.8: Comparison of single cell RNA-seq data between samples which were manually or robotically prepared. (A) Scatter plot with smooth curve showing the correlation between total gene counts of robotically and manually prepared samples. Value displayed is Pearson correlation coefficient. (B) Scatter plot with smooth curve showing the correlation between total gene counts from ERCC spike-in data only of robotically and manually prepared samples. Value displayed is Pearson correlation coefficient. (C) A box plot showing total gene counts by preparation method across each individual chromosome.

These data taken together show that the quality of data produced by robotic sample preparation does not outperform manual sample preparation. Despite this, in larger

plate based experiments the increased throughput of robotic sample preparation provide an important advantage, which may improve reproducibility.

5.3 Miniaturising sequencing reactions

Scaling plate based single cell RNA sequencing to larger experiments (>384 cells) remains an expensive proposition, but the high number of genes detected per cell and flexibility of the method make this an attractive option for some large single cell RNA-seq experiments (Baran-Gale et al., 2017). After sample acquisition the main costly steps in our experiments are divided into:

- RNA to cDNA conversion - Smart-seq2
- Preparation of sequencing library - Nextera XT (Illumina®)
- Sequencing by synthesis - Hiseq4000 (Illumina®)

The costs associated with cDNA conversion were reduced markedly when the non-proprietary Smart-seq2 method was published (Picelli et al., 2013, 2014), but remain significant in larger experiments due to large amount of reagent used per sample, certainly when compared to microfluidic methods. For similar reasons, library preparation costs also remain high with the added issues of requiring a proprietary product to remain compatible with the downstream sequencing technology used.

5.3.1 Reagent reduction

Advances in high accuracy micro-pipetting techniques, such as the Echo Acoustic Liquid Handling technology (LabcyteTM), enable precise reduction of reagent use on a per sample basis (Kanigowska et al., 2016), though the effects of this on data quality in plate based single cell RNA-seq are not known, though have been explored with Nextera XT in low input samples (Lamble et al., 2013), and quarter reaction volumes are used as standard in many single cell RNA-seq experiments. To test whether the amounts of Smart-seq2 and Nextera XT reagents could be further reduced without adversely harming data quality, human duodenal cells were sorted into a 96 well plate and processed with 1x and 0.25x volumes of the reagents used for Smart-seq2, and 0.25x (standard volume) and 0.02x volumes of the reagents used for Nextera XT library preparation (figure 5.9). Samples were then sequenced with Hiseq 4000 (Illumina®).

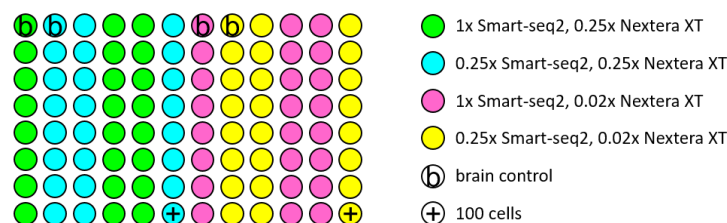


Figure 5.9: Permutation of samples to avoid batch effects when comparing different miniaturisation protocols for single cell RNA-seq. Layout of 96 well plate used for reaction miniaturisation experiments. Wells are colour coded by the amount of reagent used as indicated in the key, with 24 samples in each class. Brain controls were added after cells were 'washed out' from indicated wells.

As these samples were derived from a single duodenal biopsy and processed fresh, differences in sample quality are most likely to be due to either random experimental variation or the effects of reaction miniaturisation. Figure 5.10A-C gives an overview of the impact of changing both Smart-seq2 and Nextera XT reagent volumes by quarter and twentieth volumes respectively. Whilst reducing the Smart-seq2 volumes from 1x to 0.25x did not have a significant effect on the data, reducing Nextera XT from standard 0.25x volume further to 0.02x was detrimental to the data in both gene detection and depth (Figure 5.10D). This finding is further highlighted when looking at the gene detection profiles across each condition with even the best data in twentieth volume Nextera XT conditions showing very limited numbers of genes detected (Figure 5.10E).

Together, these data show that whilst reducing Smart-seq2 reagent volumes is feasible to reduce the cost of plate based single cell RNA-seq experiments, quarter volume Nextera XT is required to maintain data quality. Whether more modest reductions in Nextera XT reagents could preserve the data quality will require further study.

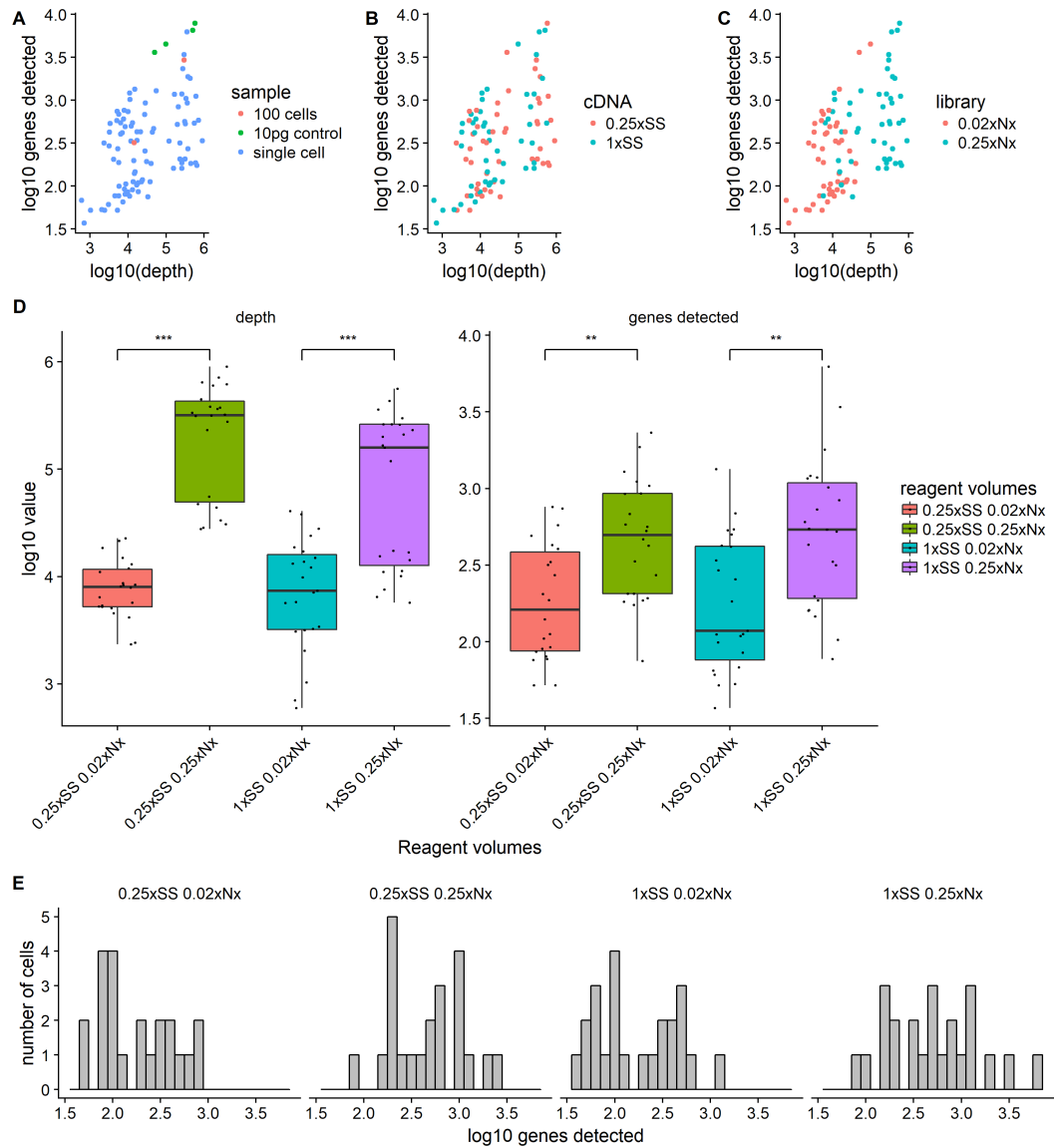


Figure 5.10: Comparison of single cell RNA-seq data between samples which were prepared with different miniaturisation protocols. (A) Scatter plot showing log transformed depth and number of genes detected in individual samples including controls, coloured by sample type. (B) Scatter plot as in A but points coloured by whether full volume Smart-seq2 (1xSS) or quarter volume Smart-seq2 (0.25xSS) reagents were used. (C) Scatter plot as in A but points coloured by whether quarter volume Nextera XT (0.25xNx) or twentieth volume Nextera XT (0.02xNx) reagents were used. (D) Boxplots showing depth and number of genes detected across all cells sequenced with each combination of reagent volumes used per cell, jittered points show results for each individual cell. (E) Histograms showing number of cells in which a given number of genes were detected (with at least one count per gene).

5.3.2 Sequencing depth requirements

Since the costs of actual sequencing are largely based on the depth of sequencing required per sample, this cost is also greatly inflated in larger experiments. Early plate based single cell RNA-seq experiments typically sequenced at a depth of over one million reads, though it is questionable at what point the sensitivity of the sample preparation chemistry prevents more gene detection, and thus at what point additional sequencing depth increases the duplication rates without increasing the number of new genes detected (Rizzetto et al., 2017).

To test the effects of reduced sequencing depth on the number of genes detected, 48 samples of diluted (to 10pg) human neural RNA were sequenced using full volume reagents and sequenced to a total depth per sample of 3×10^5 reads using Hiseq4000 (Illumina®). After mapping the data using STAR, downsampling was then undertaken using SAMtools and gene counts determined with HTseq to establish the final number of genes detected (at least one count per gene) at simulated reduced sequencing depths (0.05x, 0.1x then 0.1x increments to full depth). Figure 5.11 shows the expected effects of downsampling on the total read count with a corresponding reduction in total properly mapped reads (both mates of a given read pair map to the same chromosome correctly). Importantly, the properly:total mapped proportion is consistent, showing that the downsampling process does not unduly affect sequencing quality.

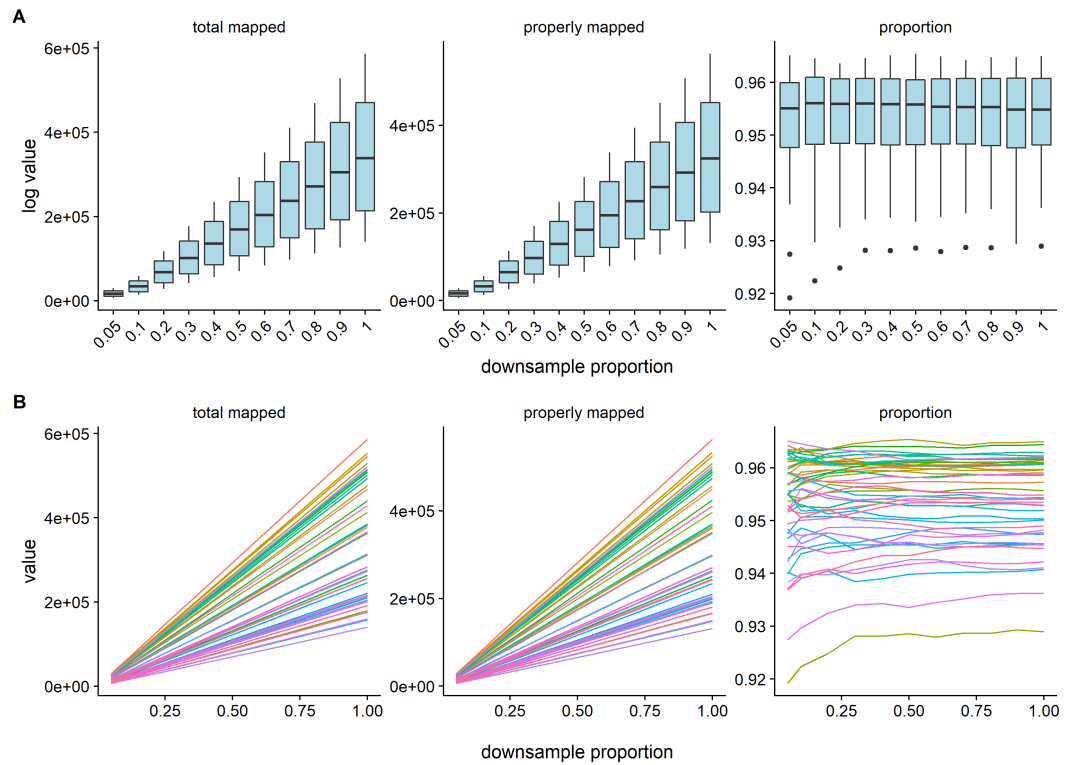


Figure 5.11: Comparisons of read mapping in diluted neural RNA sequenced with a single cell RNA-seq protocol with varying degrees of downsampling. (A) Box plots showing the effects of downsampling on total mapped reads (left plot), the total properly mapped reads, defined as reads in which both pairs of the paired end read map correctly (centre plot), and the proportion of total:properly mapped reads (right plot) in 48 samples of diluted human neural RNA, to mimic similar cDNA concentrations seen in single cell RNA-seq experiments (left plot), (B) Line plots showing the same data as in (A) with each line representing each sequenced sample.

Next, mean gene body coverage was plotted over a set of housekeeping genes for all 48 human neural samples using RSeQC (Wang et al., 2012) (Figure 5.12A). Sample 48 was selected as a typical example for downsampling and gene body coverage again plotted at each simulated reduced sequencing depth (Figure 5.12B). This shows that downsampling did not introduce 3' or 5' bias or skew gene body coverage, until the dataset was very substantially reduced.

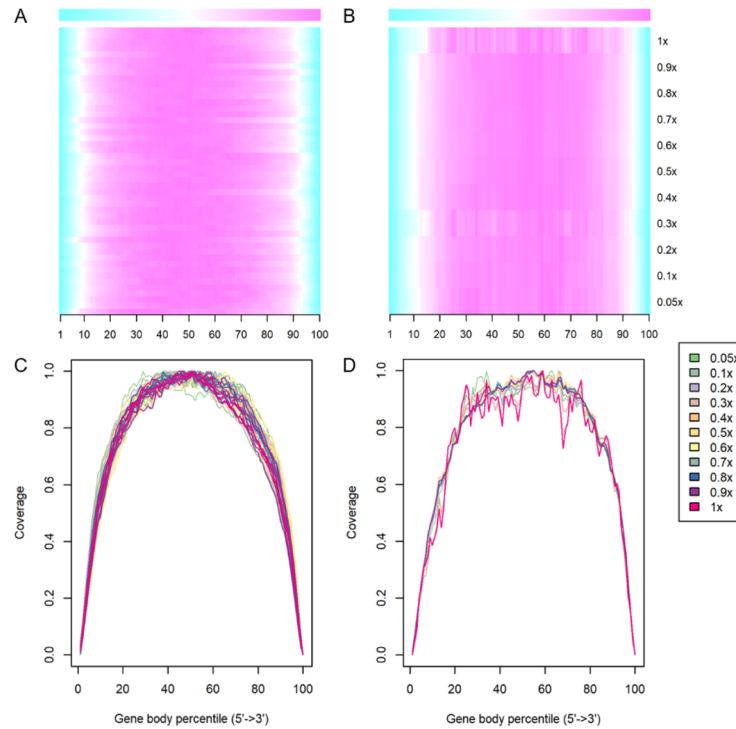


Figure 5.12: Comparisons of gene body coverage and prime bias in diluted neural RNA sequenced with a single cell RNA-seq protocol with varying degrees of downsampling. (A) Heatmap showing gene body coverage across a set of housekeeping genes in each sample (each row represents one sample, n=48). (B) As in A but applied to the first row sample with downsampling as indicated by the row labels. (C) Line plot representation of A. (D) Line plot representation of B.

Finally, to test the effects of downsampling on processed data, gene detection and depth (total gene counts) were generated from each downsampled BAM file. Reductions in gene detection and depth with downsampling were found to both be mild and linear, until data was diminished by greater than approximately 0.5x-0.6x (Figure 5.11). Assuming that epithelial cells from the upper gastrointestinal tract are likely to have fewer expressed genes than diluted total human neural RNA, it is reasonable to expect that the sequencing depth that has been applied to these diluted brain RNA samples can also be applied across human derived gut primary cells, and that 0.5x reductions in this sequencing depth will not adversely affect the processed data.

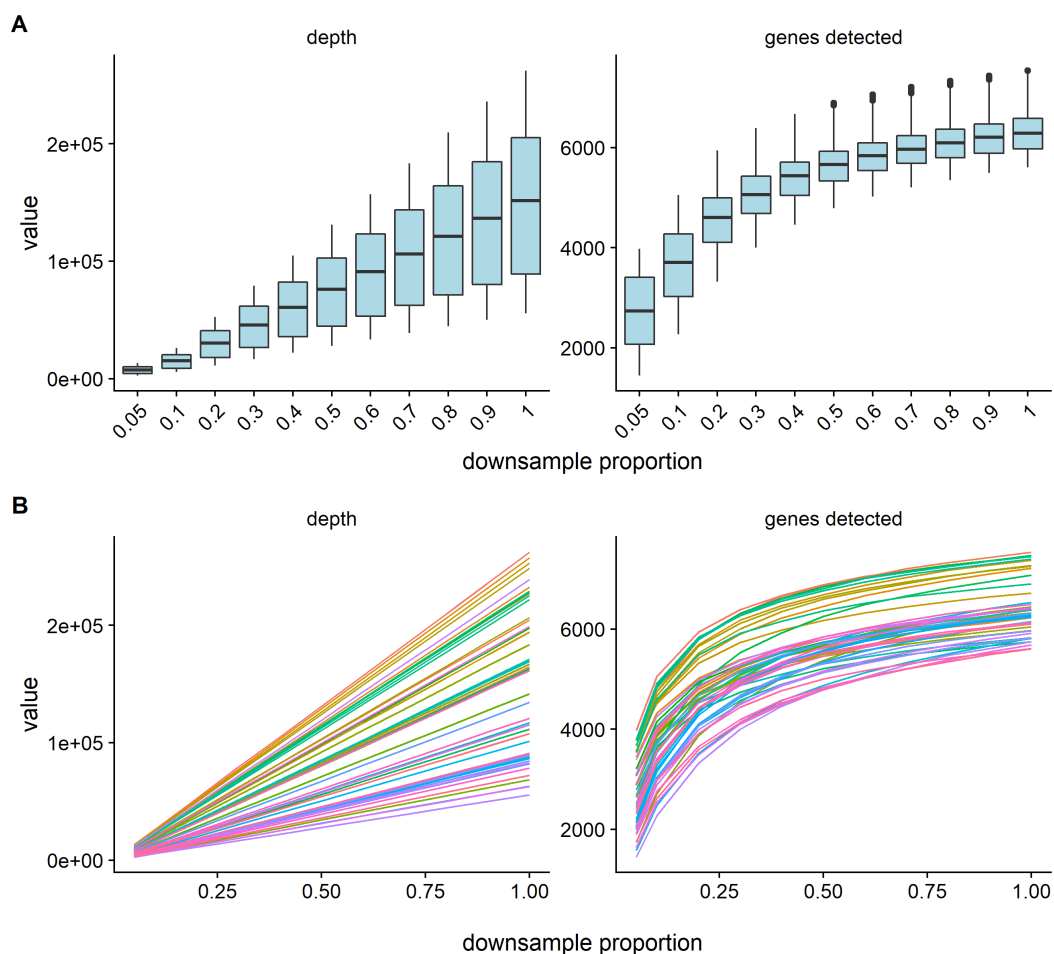


Figure 5.13: Comparisons of sequencing performance in diluted neural RNA sequenced with a single cell RNA-seq protocol with varying degrees of down-sampling. (A) Box plots showing the sequencing depth (total gene counts, left plot) and the number of genes detected (least one count per gene, right plot) in 48 samples of diluted human neural RNA, to mimic similar cDNA concentrations as seen in single cell RNA-seq experiments. (B) Line plots showing the same data as in (A) with each line representing each sequenced sample.

5.3.3 Cell selection

Whilst preparing cells for plate based single cell RNA-seq is laborious, the equipment and reagent costs prior to cDNA conversion are relatively small. Considering most plate based approaches use fluorescence activated cell sorting to select individual cells there is potential that the metrics obtained through this process could predict a given cell's likelihood of producing sequencing data of sufficient quality for analysis. To test this, we used all recorded FACS events (from the two main sorts carried out to generate data) and index sorted events relating to sequenced cells from four patients

(patients A-D, 620 BO, 352 duodenum, 678 gastric and 526 oesophagus). These events and cells were paired with forward scatter area, width and heights, side scatter area, width and heights, CD45 (a marker of immune cells) expression, EpCAM (an epithelial cell marker) expression, DAPI expression and sorting time metrics. Metrics were recorded for a total of 8248 BO, 23372 duodenum, 13610 gastric and 10847 oesophagus events (Figure 5.14).

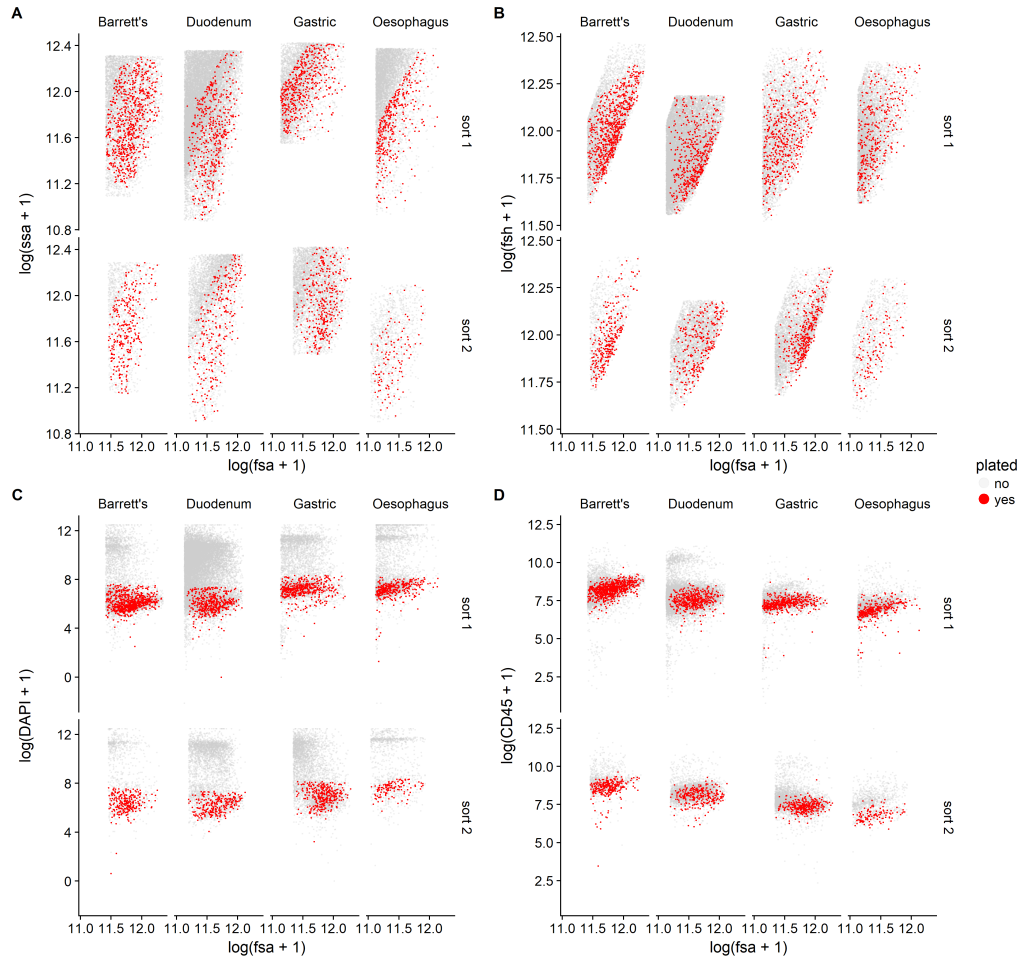


Figure 5.14: FACS metrics of cells sorted for single cell RNA-seq. (A) Forward scatter area (fsa) and side scatter area (ssa) of each cell type sorted, separated by sorting occasion. Points coloured red show those which were selected for sorting into a 96 well plate, although not all selected cells would necessarily be received into the plate due to sorting inefficiencies. (B) As in A, forward scatter area and forward scatter height (fsh), used to discriminate doublets or cell clusters. (C) As in A, forward scatter area and DAPI expression, low DAPI expressing events were selected to exclude dead cells. (D) As in A, forward scatter area and CD45 expression, where a CD45 high population of cells exists, these were excluded to avoid selecting immune cells.

Since sequenced controls were used to set a total gene count threshold for cells which

were useful for phenotype analysis, the plated cells which produced data of sufficient quality for analysis were identified from plated events (red points in Figure 5.14) and sequenced cells. Logistic regression was then used to identify whether any pre cDNA conversion metrics recorded from FACS could predict whether a cell was later used for analysis. Figure 5.14A shows that no single metric was highly significant in predicting whether a sequenced cell produced data suitable for phenotype analysis. Interestingly, sorting time was the most predictive single variable, which underlines the importance for speed in plate based single cell RNA-seq experiments, where cell preservation techniques are not employed. Taking all metrics together, the receiver operating characteristic curve in Figure 5.14B shows that we can not reliably predict whether a cell was later used for analysis (area under curve (AUC)=0.64).

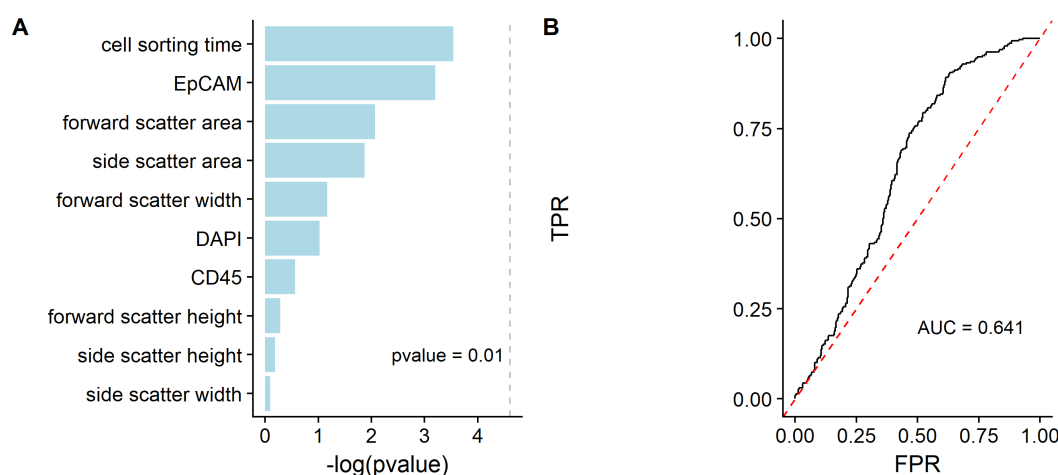


Figure 5.15: Predicting single cell RNA-seq data quality from metrics obtained before cell lysis and storage.(A) Results of a logistic regression model of a sequenced cells likelihood of producing data of sufficient quality for phenotype analysis. (B) Receiver operating characteristic curve showing model's overall performance in predicting whether a cell produced data of sufficient quality.

5.4 Single cell RNA-seq in a phase I/II trial of immunotherapy for oesophageal cancer

Cancer therapies designed to augment natural immune defenses against cancer have led to striking improvements in cancer care in many solid tumour types, perhaps most notably in melanoma (Wolchok et al., 2017). Interestingly, highly mutated cancers tend to respond better to immunotherapy (Rizvi et al., 2015), and oesophageal adenocarcinoma is one such cancer (Goodman et al., 2017; Lawrence et al., 2013). To

investigate the effects of immunotherapy in oesophageal cancer (OC) we designed a phase I/II open label trial (LUD2015-005) to test the safety of combining anti-PDL1 and anti-CTLA4 therapies with conventional chemotherapy, and chemoradiotherapy and surgery in up to 75 patients. Using the findings described above we were able to scale 384 well plate based RNA-seq to the relatively high sample numbers expected in this trial. We therefore aimed to isolate single cells from oesophageal cancers at different time points to look at how the transcriptional landscape of the tumour changed in response to therapy. Ultimately the aim is to look at changes that can separate responders and non responders, however this is beyond the scope of this thesis. The results of our first pilot experiment are presented in a single patient who was sampled at baseline then again a month later, after two cycles of anti PD-L1 therapy.

A total of 864 samples were sequenced with mean total gene counts of 51592 and mean of 922 genes detected (at least one count per gene). Figure 5.16 summarises the total counts and number of genes detected (at least one count per gene) by cell and control type. A further experimental condition was included in this pilot experiment to test whether Nextera XT library preparation with the Echo[®] acoustic liquid handling robot device produced similar results to the robotic methodology described above. All cells were collected, sorted and snap frozen, then thawed for miniaturised cDNA conversion and Nextera XT library preparation before sequencing with Hiseq4000 (Illumina[®]) to a reduced per sample depth.

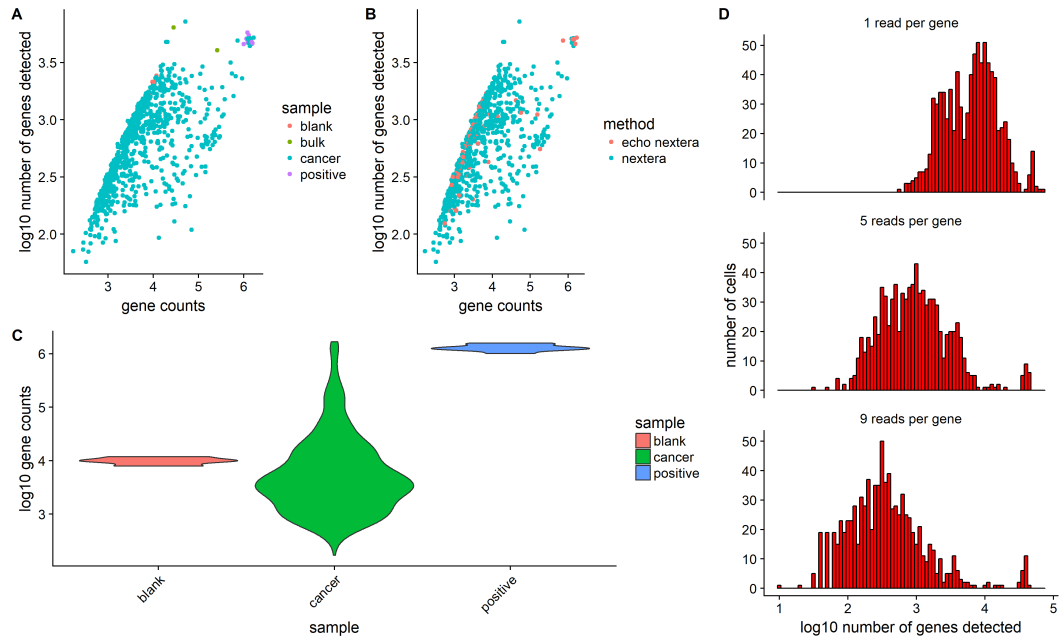


Figure 5.16: Single cell RNA-seq data quality from oesophageal adenocarcinoma samples from the LUD2015-005 trial. (A) Scatter plot showing the total number of gene counts and total number of genes detected per cell (one count per gene). Points are coloured according to cell or control type. (B) As in A but points are coloured to indicate which type of robotic library preparation (Nextera) was performed (Echo[®] Acoustic Liquid Handling or FXP Laboratory Automation Workstation). (C) Violin plot showing the total gene counts in cancer cells, positive controls and negative controls. (D) Histograms showing how the number of genes detected changes depending on how many counts per gene were taken as proof of gene detection.

After filtering out samples which did not achieve total counts of at least the maximum observed in blank controls, 237 cells (27%) were left for analysis. This high attrition rate is likely due to difficulties in processing samples contemporaneously for this patient. In cells used for analysis, mean gene counts was 1.8×10^5 per cell and number of genes detected (at least one count per gene) was 1610 genes per cell. Since no discernible difference was observed between sample library preparation using the different robot, these data were kept combined. Oddly, the mean gene counts were higher than in previous experiments, but presumably this was due to the relatively low gene detection observed and variability in true read count obtained from each sequencing run.

5.4.1 Single cell RNA-seq identifies a response signature in oesophageal cancer treated with immunotherapy

First, cancer cells were clustered using SC3 (Kiselev et al., 2017) (Figure 5.17A). Silhouette indices showed that $k=5$ gave the most stable clustering structure.

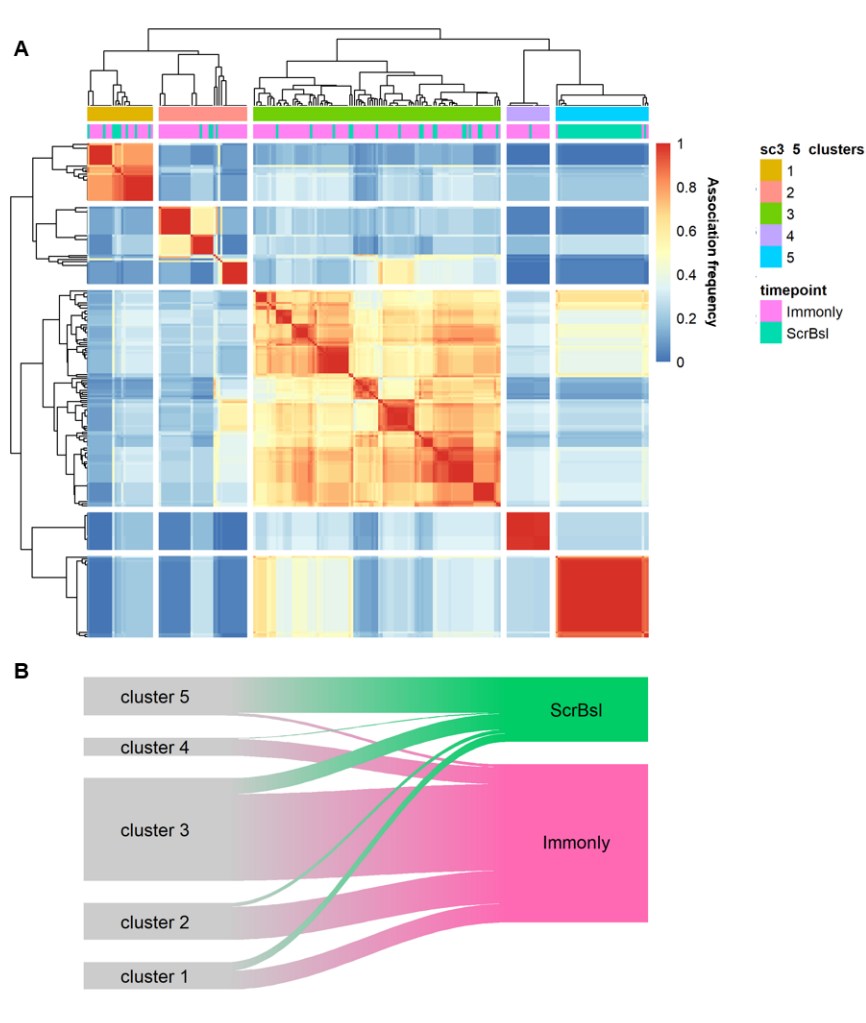


Figure 5.17: Clustering of single cells from a LUD2015-005 trial patient. (A) Consensus matrix heatmap showing how cells clustered across the five SC3 clusters. Annotation tracks above the heatmap show cluster number (top) and time at which the cell was sampled (bottom), before (ScrBsl) or after (Immonly) immunotherapy. (B) River plot summarising the results of A showing how screening baseline (ScrBsl) cells and post-immunotherapy (Immonly) cells clustered.

Of the five clusters, clusters 1-4 were mainly post immunotherapy cells, while cluster 5 was mainly baseline (pre therapy) cells (Figure 5.17B). By labelling clusters with less

than 20% crossover of cells from different time points, clusters can then be referred to as:

- cluster 1 - Mixed cluster (31% pre therapy cells)
- cluster 2 - Post immunotherapy cluster (10% pre therapy cells)
- cluster 3 - Post immunotherapy (16% pre therapy cells)
- cluster 4 - Post immunotherapy cluster (5% pre therapy cells)
- cluster 5 - Pre therapy cluster (93% pre therapy cells)

Differential gene expression in each cluster (comparing to all other clusters) shows that each cluster had a distinctive gene expression profile (Figure 5.18A). Profiling the biological processes using gene ontology shows that 'response to biotic stimulus' pathways were enriched in post immunotherapy cluster 3, while response to stimulus, programmed cell death and immune response pathways were enriched in pre therapy (ScrBsl) cluster 5 (Figure 5.18B).

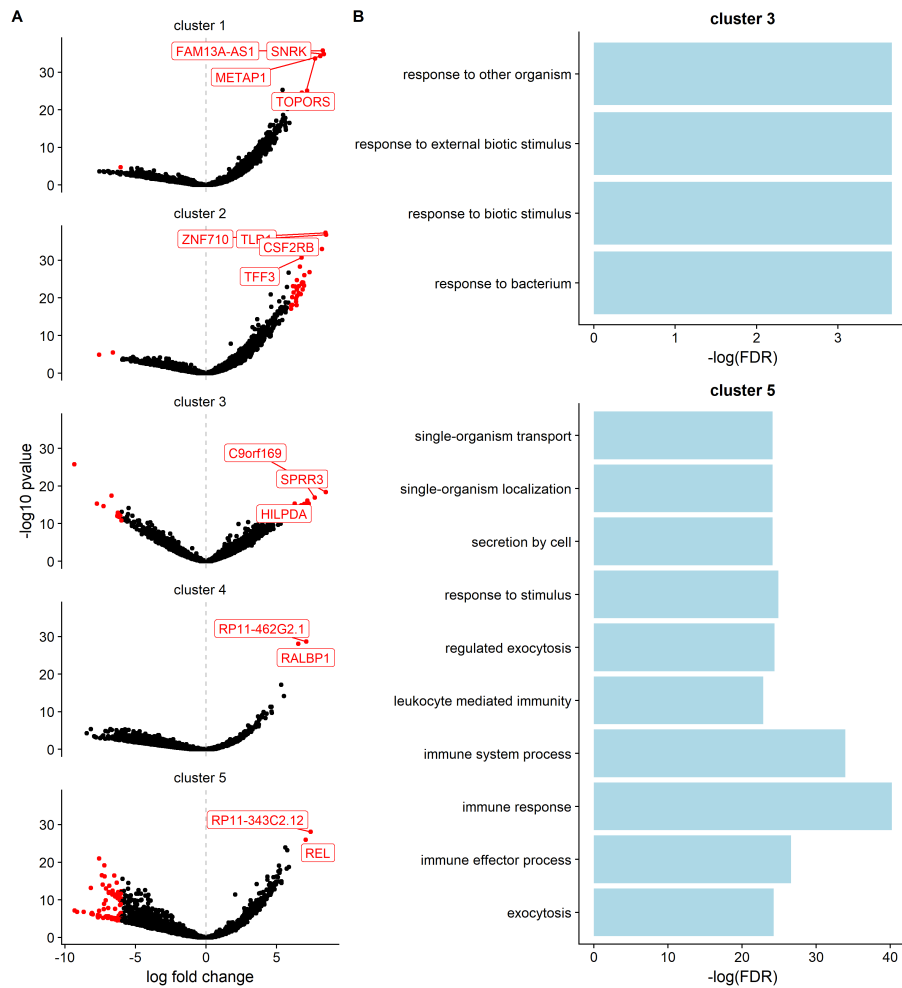


Figure 5.18: Differentially expressed genes and pathways between cell clusters from a LUD2015-005 trial patient. (A) Volcano plots showing the fold change and significance of differentially expressed genes across the five clusters defined in Figure 5.17. Points marked red have a q value of <0.001 and a log fold change between -6 and 6 . The most significant genes with a positive log fold change are labelled in red. (B) Top ten enriched biological processes (q value <0.01), based on the full results of A. Clusters not shown did not produce sufficient significant results.

The gene programmes specific to a given cluster are shown in more detail in Figure 5.19.

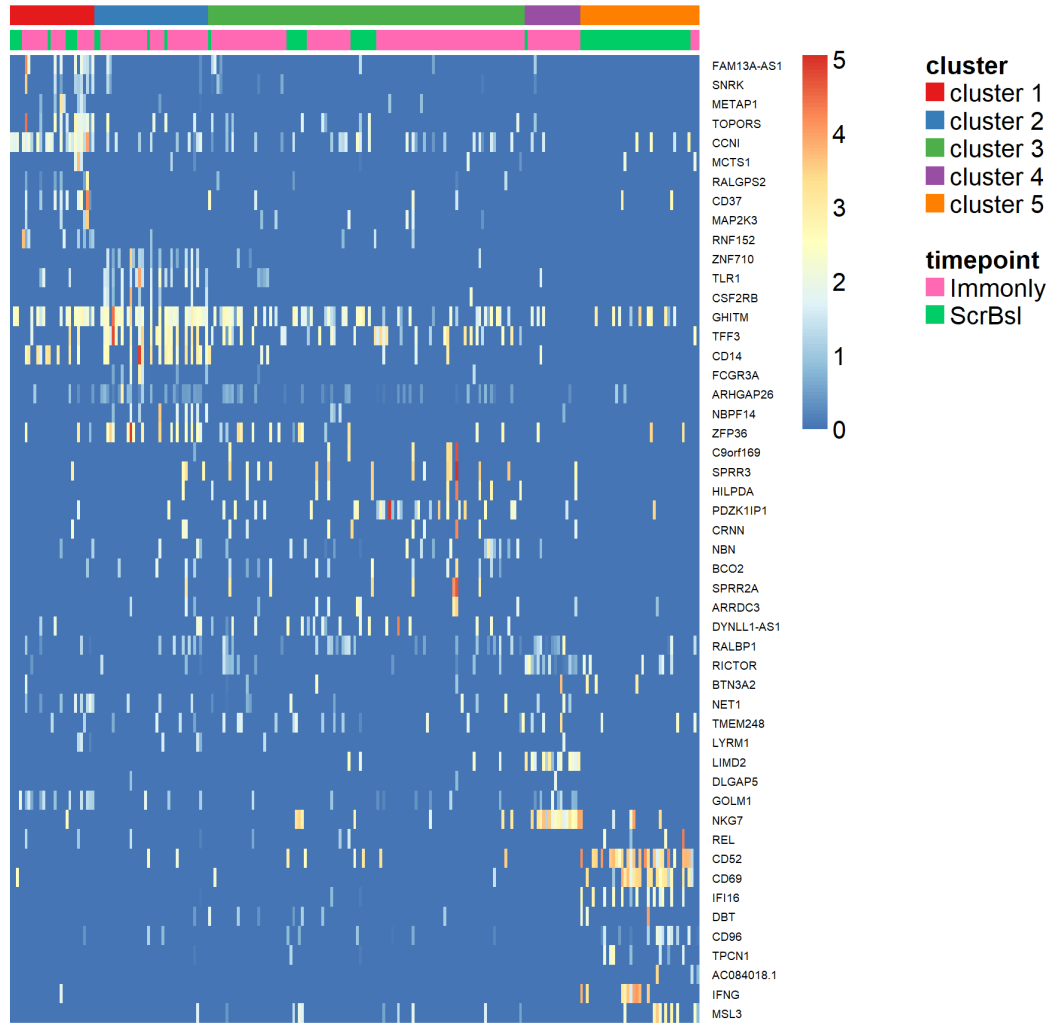


Figure 5.19: Gene programmes which defined each cell cluster from a patient in the LUD2015-005 trial. Heatmap showing the top ten differentially upregulated gene (log2 fold change >2) in each SC3 defined cluster (cluster and sampling time point indicated by the annotation tracks above the heatmap). ScrBsl timepoint is pre therapy, Immunonly timepoint is after immunotherapy.

Interestingly cluster 5, which is mainly a pre therapy cell cluster, differentially expressed *CD52* and *CD69*, both which are expressed by T-lymphocytes. Conversely, innate immune markers *TLR* and *CD14*, typical of dendritic cells or macrophages, were expressed in cluster 2, a mainly post immunotherapy cluster (Figure 5.20A). Direct comparison of immunity types, by separating adaptive and innate by *CD52* and *CD14* expression respectively, confirms that most of these cells are in the pre therapy cluster 5 and post immunotherapy cluster 2 (Figure 5.20B, Fishers test $p=0.007$). Some cells in cluster 3 also expressed *CD52*, but further analysis shows that most of

these CD52+ cells were rarely post immunotherapy (Figure 5.20A).

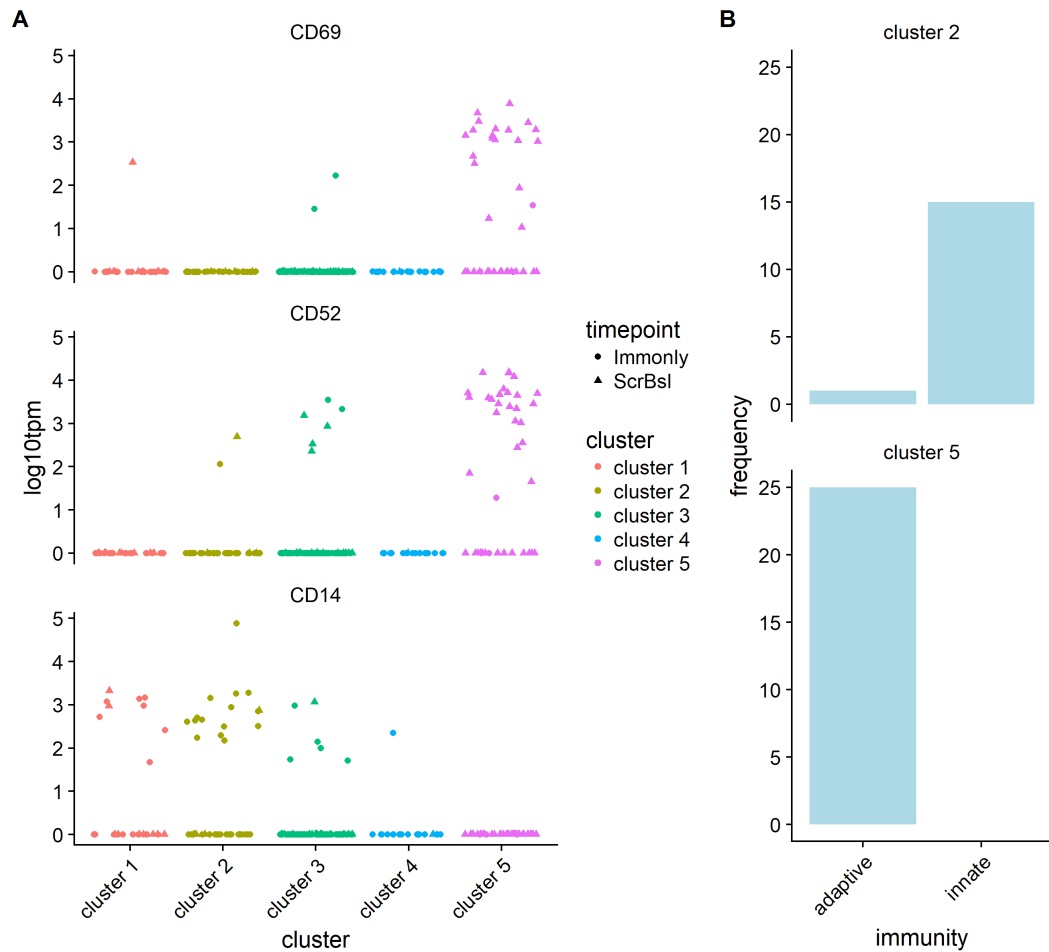


Figure 5.20: Characterisation of the immune switch seen in pre and post therapy cell clusters. (A) Jitter plots showing expression of *CD69*, *CD52* and *CD14* in cells by SC3 cluster (defined in Figure 5.17). Point colours indicate cluster and shapes indicate whether cells were pre therapy (ScrBsl) or post immunotherapy (Immonly). (B) Bar chart showing the frequency that innate immunity cells (*CD14*+) or adaptive immunity cells (*CD52*+) occurred in the mainly post immunotherapy cluster 2 and the mainly pre therapy cluster 5.

Since this patient showed a significant reduction in tumour volume between time points on clinical follow up, these results would seem to relate the shift from adaptive to innate immunity characterising an anti tumour immune response within the tumour microenvironment, as a consequence of anti PD-L1 treatment. To help interpret these results, an better characterisation of the T-cell receptor repertoire would be valuable, to help relate the effects of anti PD-L1 therapy with the adaptive immune switch I have observed. One way to achieve this would be to include T-cell receptor sequencing in future experiments, or at least to try to derive this from bulk RNA-seq using one

of several published bioinformatic methods. This work is now underway in a wider range of patients in the LUD2015-005 trial and will form a significant part of another student's work.

5.5 Summary

In this chapter I have demonstrated some modifications to a typical single cell RNA-seq work flow to increase the flexibility of this technique and make it more suitable for clinical science. These modifications include sample preservation to allow delayed samples processing, robotic techniques for increased throughput, and reaction miniaturisation to make plate based single cell RNA-seq more feasible. I have also shown that modest reductions in sequencing depth did not significantly affect processed data quality. The advent of cell encapsulation procedures using microfluidic systems or micro-well slides has increased the appeal of single cell RNA-seq and made much higher throughput experiments more achievable, however for some samples which are low yield and irreplaceable, the risk of technical failures such as a chip blockage causing a complete experimental failure may be too high, hence the modifications I propose could prove useful for sample storage or to enable plate based sequencing experiments, where this is a more acceptable methodological approach. As an example I used some modifications to show that a pre and post therapy signal is captured from single cell RNA-seq data from oesophageal cancer in a clinical trial investigating the safety of immunotherapy.

Chapter 6

Tissue homeostasis at the upper gastrointestinal squamocolumnar junction

Understanding how tissues are affected in gastro-oesophageal reflux has frequently relied on retrospective assessment of tissue samples, or *in vitro* assessment of two dimensional culture systems. In this chapter, I replicate some previous studies looking at the effects of acid and bile salts on squamous cells, demonstrate organoid gastric cell culture using mouse stomach for future study of these effects, and show some early results of squamo-columnar co-culture working towards the reconstruction of an *in vitro* squamo-columnar junction.

6.1 Squamocolumnar junction in humans and mice

As explained previously, a crucial difference between human and mouse gastro-oesophageal anatomy is the position of the squamo-columnar junction (SCJ). In humans the SCJ roughly corresponds to the anatomical division of the stomach and oesophagus, whilst in mice the SCJ is mid gastric (Figure 6.1A). In Barrett's oesophagus (BO), the SCJ no longer resides at the gastro-oesophageal boundary, but is found higher up in the oesophagus (Figure 6.1A, middle image and Figure 6.1B). This creates difficulties in assessing the true SCJ in mice, where a convenient and clear anatomical reference is not present. Squamous markers such as p63 and keratin 14 (KRT14), and columnar markers such as keratin 8 (KRT8) can demonstrate the position of the SCJ (Figure 6.1D).

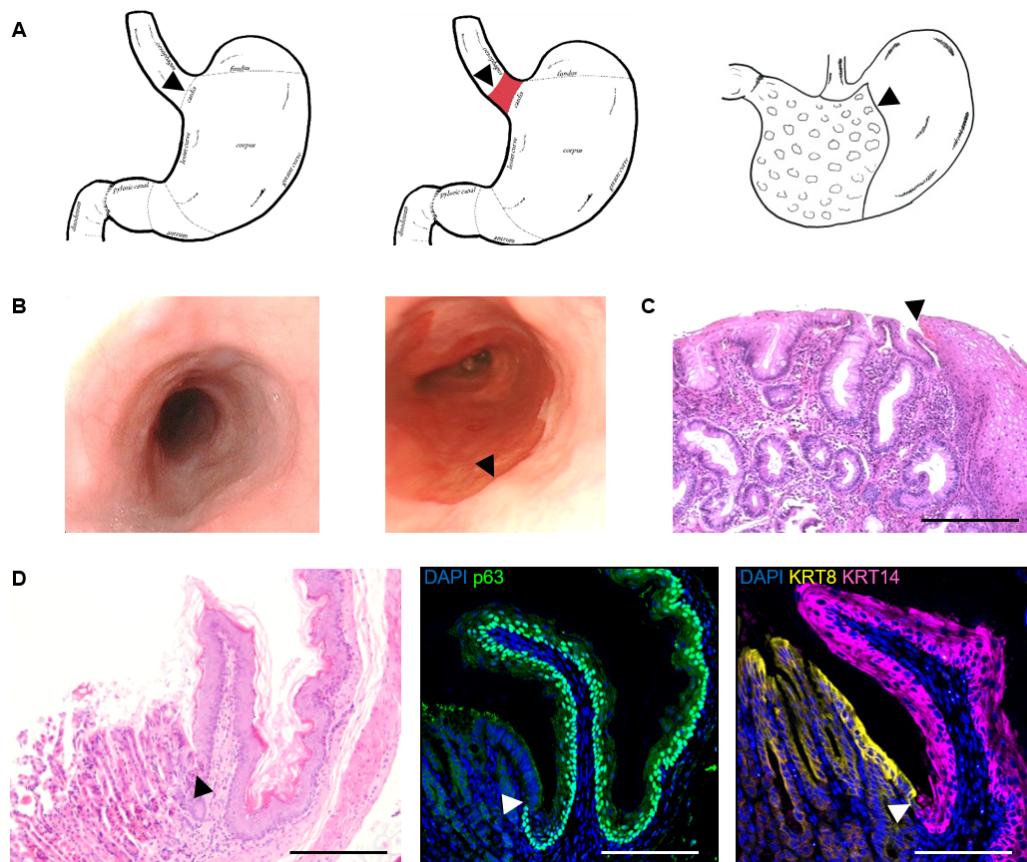


Figure 6.1: Key differences seen macroscopically, microscopically and in protein expression at the squamo-columnar junction in the upper gastrointestinal tract. (A) Cartoon images depicting the lower oesophagus, stomach and first part of the duodenum in normal human (left side), human with BO (middle) and mouse (right side). (B) Endoscopic images of the lower part of the oesophagus in normal human (left side) and in BO (right side). (C) H&E stained section of the SCJ in a human tissue sample. (D) Sections of mouse stomach including the SCJ. The left side image is H&E stained, the middle image is immunofluorescence staining with p63 (green) and DAPI (blue), and the right side image is stained with keratin 8 (KRT8, yellow), keratin 14 (KRT14, pink) and DAPI (blue). Scale bars are 250 μ m. Black and white arrow heads indicate SCJ positions.

This lack of an obvious sampling point at the SCJ is a key reason why murine studies of the SCJ are difficult and require destructive sampling to complete. Alternative methods to investigate changes in the oesophagus use squamous derived cell lines.

6.2 Effects of gastro-oesophageal reflux on a squamous and columnar derived cell line

Since BO occurs in a chronic reflux context, simulating acid and bile reflux to investigate for squamous-to-columnar changes has been investigated previously, with upregulation of columnar markers and downregulation of squamous markers shown on exposure to acidified bile salt solutions. To replicate these previous experiments, OE21 (squamous derived, oesophageal squamous cell carcinoma) and OE33 (columnar derived, oesophageal adenocarcinoma) cell lines were cultured to 70% confluence then exposed to low pH conditions in phosphate buffered solution for 30 minutes, to simulate gastro-oesophageal reflux. Since the role of bile salts has also been clearly implicated in causing changes in squamous cells, 200 μ M of deoxycholic acid (DCA) was included as an additional condition in experiments. Despite previous studies using up to 1000 μ M DCA and exposure times of up to 24 hours, I found that few cells remained viable after more than a one hour exposure to pH 5.2 PBS with 200 μ M DCA, and that creating a DCA solution was difficult below a pH of 5.2, in PBS or culture medium. Following exposure, cell cultures were then split and culturing continued for three days in normal conditions prior to harvest and protein extraction with urea lysis. Using cancer derived cell lines for this work is not ideal and prompts caution in how findings are interpreted, but unfortunately time limitations and difficulties in culturing primary cell cultures from relevant tissues prevented more rigorous experiments. Considering the data presented in previous chapters, a cell type of particular interest would be from oesophageal submucosal glands (OSGs) however no OSG cell lines exist and I have not had success in extracting these from primary tissues yet, though further analysis of sequencing data may help identify a suitable target on which OSG cells can be isolated with fluorescence activated cell sorting. Table 6.1 shows more details of the cell lines used.

name	p53 mutation	oesophageal position	type	differentiation	stage
OE21	truncation	mid	SCC	moderate	2A
OE33	404G>A, protein missense C135Y	distal	AC	poor	2A

Table 6.1: Cell lines used. Table summarising the two cell lines used for low pH, bile exposure and co-culturing experiments. SCC=squamous cell carcinoma; AC=adenocarcinoma.

6.2.1 p63

Without DCA, some subtle decrease in p63 was observed after three days of culture following low pH exposure, however this was not apparent in shorter or longer culturing periods (24 hours and five days) (Figure 6.2A). In view of this, subsequent experiments were carried out following three days of normal condition culture. As expected, immunoblotting for p63 in the columnar derived OE33 cells showed little p63 activity before or after low pH exposure (Figure 6.2B). With the addition of 200 μ M DCA a subtle reduction in p63 was observed in low pH conditions (Figure 6.2C, upper blots), however this experiment failed to replicate the changes with pH alone observed in Figure 6.2A. In view of the potential for protein insolubility with chaotropic lysis, a radioimmunoprecipitation assay (RIPA) buffer was added to separate the soluble and insoluble protein fractions. This showed a more convincing change in p63 expression, presumably due to increased monomeric protein presence, although again this was limited to the combined low pH and DCA condition (Figure 6.2C, lower blots). For clarification, and to check whether a heterogeneous response could account for the conflicting results, cells were cultured on cover slips to 50% confluence and exposed to the low pH and 200 μ M DCA for 30 minutes, then cultured for three days. This showed a consistent reduction in p63 expression, more in keeping with previous studies (Figure 6.2D). This different reduction in p63 expression seen between immunofluorescence staining and immunoblotting may be due to on how proteins are denatured differently for each experiment, hence antibodies may be less able to detect the protein in immunoblots due to the harsher processing conditions. Further experiments with different antibodies may help address this difference. In addition to using different p63 antibodies, testing for upregulation of a generic columnar marker, such as KRT7, might be useful as I would expect to see a far higher proportionate increase (relative to a decrease in squamous markers).

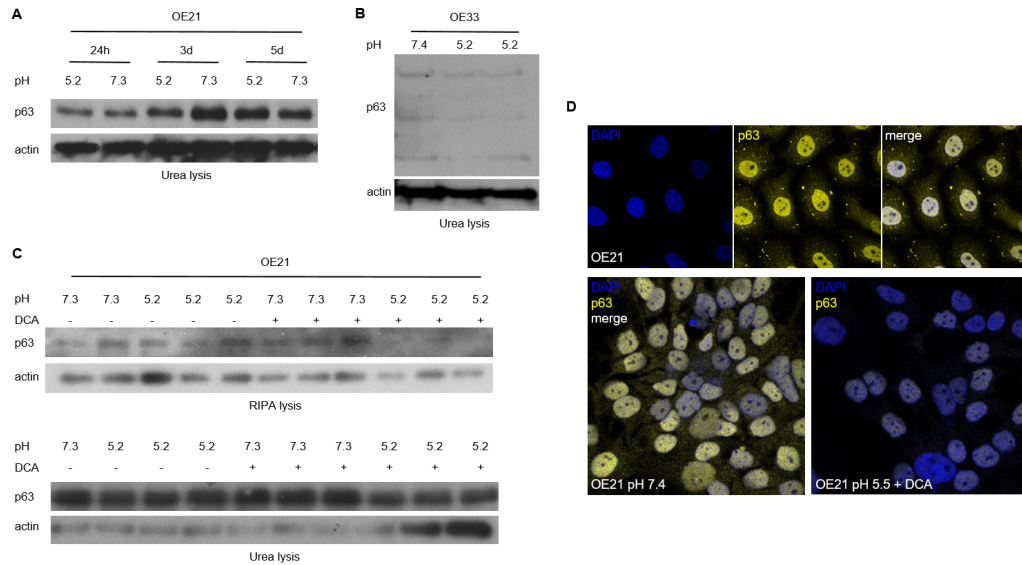


Figure 6.2: Changes in expression of p63 in response to acid and bile exposure to oesophageal cell lines. (A) Immunoblotting for p63 from urea lysates of OE21 cells (oesophageal squamous cell carcinoma derived). Cells were exposed to normal or low pH PBS for 30 minutes, then cultured in normal condition for 24 hours, three days and five days as indicated. (B) Immunoblotting for p63 from urea lysates of OE33 cells (oesophageal adenocarcinoma derived). Cells were exposed to normal or low pH PBS for 30 minutes, then cultured in normal conditions for three days. (C) Immunoblotting for p63 from urea and RIPA lysates of OE21 cells. Cells were exposed to normal or low pH PBS with and without 200 μ M DCA for 30 minutes, then cultured in normal condition for three days. (D) Immunofluorescence staining for p63 in OE21 cells in normal conditions (upper images), and after exposure to normal pH PBS and low pH PBS with 200 μ M DCA, followed by 3 days of culture in normal conditions (lower images). β -actin controls shown for all blots. DCA=deoxycholic acid.

6.2.2 Keratin 14

As shown in Figure 6.1D, KRT14 gives sharp demarcation of the SCJ, and thus its expression would be expected to diminish as squamous tissues take on columnar characteristics. Following the same experimental design as above, immunoblotting and immunofluorescent staining for KRT14 were carried out in OE21 cells, following low pH and DCA exposure. Initial results with urea lysis were very poor and thus the experiment was repeated with a RIPA buffer step, following the same reasoning as above. With RIPA lysis, a modest reduction in KRT14 expression was seen in immunoblotting (Figure 6.3A), whereas this was not seen in the insoluble (urea lysis) protein fraction (Figure 6.3B). To confirm this finding, immunofluorescence staining

of OE21 cells exposed to similar pH and DCA concentrations was undertaken, showing a reduction in KRT14 expression (Figure 6.3C).

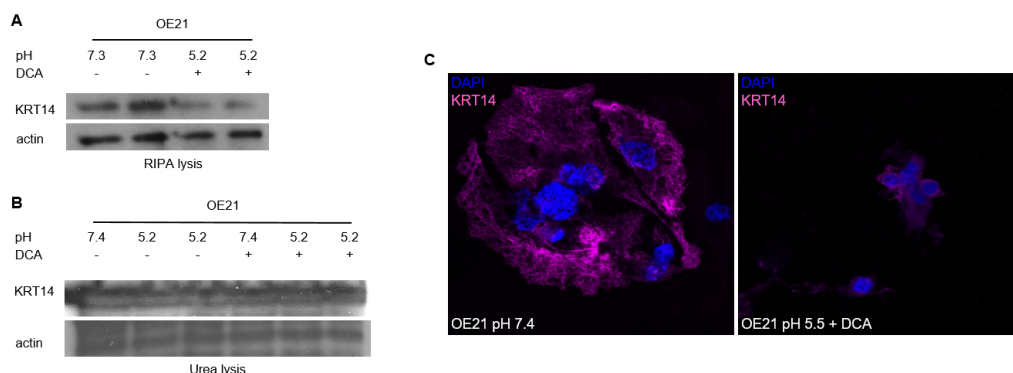


Figure 6.3: Reduced expression of KRT14 in response to acid and bile exposure to a squamous derived cell line. (A) Immunoblotting for KRT14 from RIPA lysates of OE21 cells. Cells were exposed to normal or low pH PBS with and without 200 μ M DCA for 30 minutes, then cultured in normal condition for three days. (B) As in A but applied to urea lysate (insoluble protein fraction). (C) Immunofluorescence staining for KRT14 in OE21 cells after exposure to normal pH PBS (left image) and low pH PBS with 200 μ M DCA (right image), followed by 3 days of culture in normal conditions. β -actin controls shown for all blots. DCA=deoxycholic acid.

6.2.3 Tumor Protein P53 Binding Protein 2 and Caudal Type Homeobox 2

Previous studies investigating the induction of a BO-type change in squamous cells in acidified bile exposure have used Caudal Type Homeobox 2 (CDX2) as a marker of intestinal phenotype. Since the squamous genes p63 and KRT14 were reduced on low pH-DCA exposure in OE21 cells, a similar condition might be expected to produce increased CDX2 expression, however this was not observed (Figure 6.4A). Immunofluorescence staining for CDX2 in OE21 cells was negative, even after low pH-DCA exposure, which might be expected in a squamous-derived cell line, at least in normal conditions (Figure 6.4B). This finding conflicts previous study but may be due to a lack of optimisation of the experiment or lack of conditions explored. It is also possible that the culture conditions may damage the CDX2 protein preventing antibody binding to its epitope. To confirm that this antibody was working, a tissue section of BO with intestinal metaplasia (IM) was stained, showing strong nuclear CDX2 expression (Figure 6.4D, upper image), thus a more relevant control in cell line experiments may have been a columnar cell line such as OE33.

The ASPP (Apoptosis Stimulating Protein of p53) family of proteins consists of three members: ASPP1, ASPP2 and iASPP. ASPP2 (or Tumor Protein P53 Binding Protein 2) stimulates apoptosis through its ability to alter the promoter selectivity of p53 family members (Samuels-Lev et al., 2001). ASPP2^{+/-}BALB/c mice represent the first haploinsufficient spontaneous squamous cell carcinoma (SCC) mouse model, and ASPP2 was shown to repress SCC by inducing nuclear NF κ B/p65RelA mediated repression of p63 expression (Tordella et al., 2013). ASPP2 could therefore connect p63 and NF κ B inflammatory signalling pathways to carcinogenesis in the oesophagus. Also, ASPP2 and Notch co-localise in stratified squamous epithelium (SSE) and gene array data has shown that ASPP2 is required for Notch to activate Hey2 and HeyL (Notch targets that induce SSE differentiation), and to repress Hes1 and Hes5 (Notch targets that maintain progenitor status gastric epithelial cells) through an unknown mechanism (Tordella et al., 2013). Therefore ASPP2 may influence epithelial homeostasis through its ability to regulate Notch directly, and via NF κ B/p63 indirectly. As p63 expression in squamous-derived cells is reduced by low pH-DCA exposure, it might be expected that ASPP2 should increase, if an NF κ B/p65RelA-dependent mechanism is at play. In fact, RIPA lysis immunoblotting for ASPP2 in OE21 cells exposed to low pH or DCA conditions showed reduced ASPP2 expression (Figure 6.4C). However, immunofluorescence for ASPP2 in OE21 cells did show cytoplasmic staining in some cells (Figure 6.4C), though it is surprising that this antibody did not stain more floridly in normal pH-no DCA exposed cells since ASPP2 expression is expected in squamous tissues, as shown in normal human oesophagus (Figure 6.4D, lower image). These results offer conflicting evidence for a p63/ASPP2 relationship in OE21 cells exposed to low pH and DCA, and could suggest that the p63 expression changes observed in this cell line may not only follow the p63-NF κ B/p65RelA-ASPP2 mechanism previously described, though further conditions and cell lines are needed to test this further and explore alternative interactions.

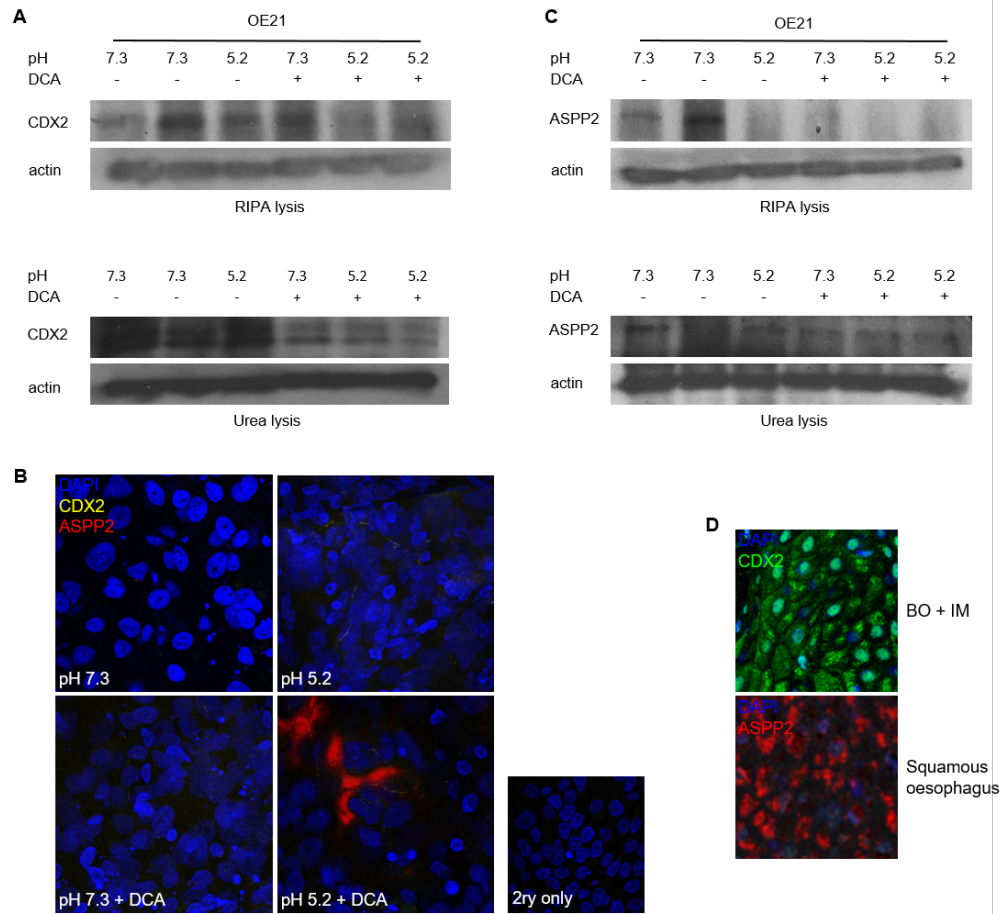


Figure 6.4: Expression of CDX2 and ASPP2 in response to acid and bile exposure to a squamous derived cell line. (A) Immunoblotting for CDX2 from RIPA lysates (upper blots) and urea lysates (lower blots) of OE21 cells. Cells were exposed to normal or low pH PBS with and without 200 μ M DCA for 30 minutes, then cultured in normal condition for three days. (B) As in A but blotted for ASPP2. (C) Immunofluorescence co-staining for CDX2 (yellow) and ASPP2 (red) in OE21 cells after exposure to low pH PBS with and without 200 μ M DCA for 30 minutes, then cultured in normal condition for three days. Secondary antibody only control shown in small image on right side (no PBS exposure). (D) Immunofluorescence staining for CDX2 and ASPP2 in BO with IM and normal squamous epithelium tissue sections respectively, as positive controls for the antibodies used in C. β -actin controls shown for all blots. DCA=deoxycholic acid.

To explore the relationship between ASPP2 and CDX2 in the context of BO, gastric-BO and intestinal BO tissue sections from two patients were co-stained with CDX2 and ASPP2. Usually, ASPP2 expression is seen in the squamous oesophagus (Figure 6.5A) and some glandular tissues, and CDX2 expression is restricted to intestinal tissues (as seen in BO with IM in Figure 6.4D, upper image). In gastric-BO, ASPP2 staining was positive in some cell nuclei, while CDX2 staining was mainly cytoplasmic

with no co-staining seen (Figure 6.5B, upper images and upper expanded image). In intestinal-BO, nuclear CDX2 expression was seen in tissues, however ASPP2 was absent (Figure 6.5B, lower images). Neither of these proteins were seen in goblet cells (Figure 6.5B, lower expanded image). In view of these findings, a direct relationship between CDX2 and ASPP2 in BO seems unlikely.

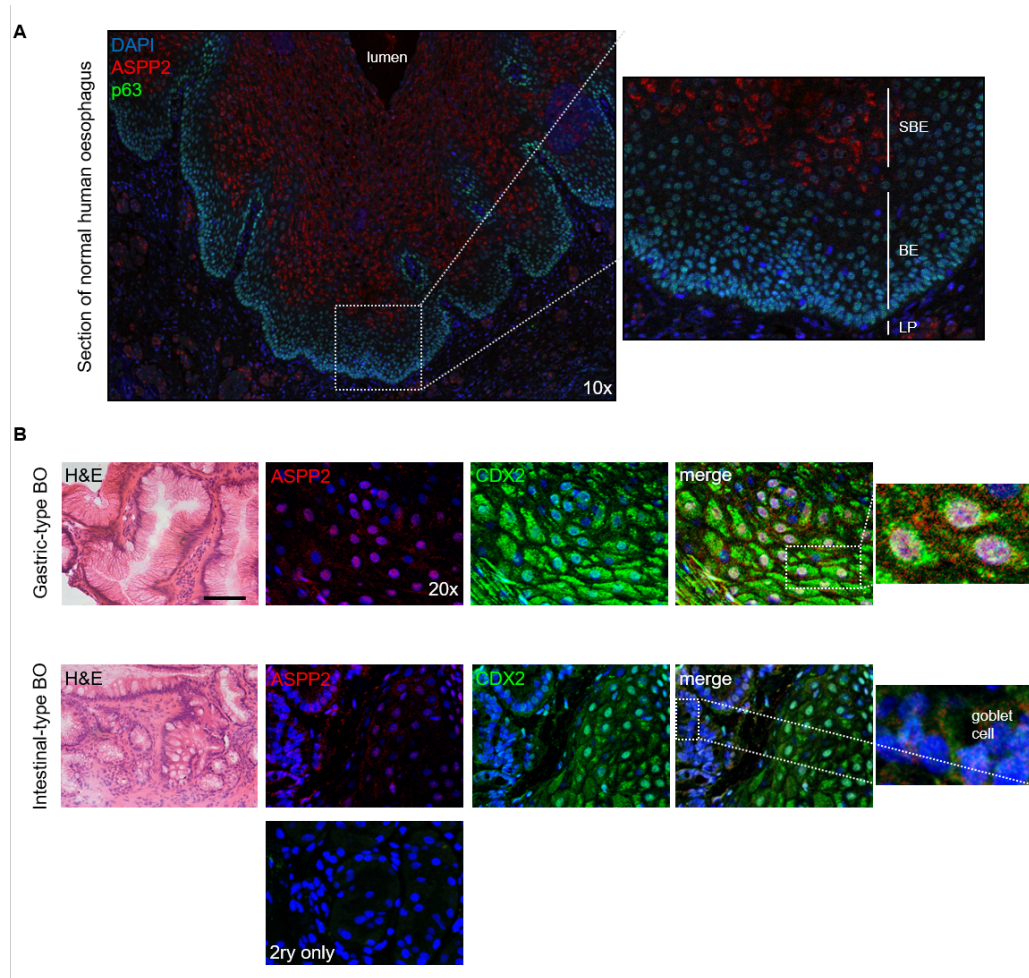


Figure 6.5: Expression of CDX2 and ASPP2 in normal human oesophagus and Barrett's oesophagus. (A) Immunofluorescence co-staining for ASPP2 (red) and p63 (green) in a large section of normal human oesophagus. Expanded image includes annotation for squamous epithelial layers (SBE=supra-basal epithelial layer, BE=basal epithelial layer, LP=lamina propria). (B) H&E images on left show sections of BO with mainly gastric phenotype (top) and intestinal phenotype (bottom). Immunofluorescence staining shows the corresponding expression of ASPP2 (left side), CDX2 (middle) and merged (right side) in the BO sections, with expanded images showing areas of co-staining in gastric-phenotype BO (top) and a goblet cell in the intestinal-phenotype BO (bottom). Secondary antibody only control shown in small image at bottom.

6.2.4 Effects of ASPP2 on organoid development in murine gastric tissue

Organoid culture is a method which has been used to model the development of upper gastrointestinal tissues. Together with Dr Michael White (DPhil student, Ludwig Institute for Cancer Research, Oxford) and Dr Lai Mun Wang (Consultant Pathologist, John Radcliffe Hospital), I looked at the SCJs of transgenic mice used for other experiments by other investigators in the laboratory of Professor Xin Lu (Ludwig Institute for Cancer Research, Oxford). Interestingly, the SCJs of $p53^{+/-}ASPP2^{-/-}$ mice showed disturbed epithelial homeostasis, here demonstrated by underdeveloped columnar and squamous epithelium with fibroproliferation (Figure 6.6A). In brief, these mice were developed from $p53$ -heterozygous mice (on a 129SvJxC57BL/6 background) from the Jackson Laboratories crossed once with C57BL/6J to generate the $p53$ -heterozygous (C57BL/6Jx129SvJ) colony. $ASPP2^{+/-}$ mice (also on a C57BL/6Jx129SvJ background and obtained from our laboratory) were crossed with $p53^{+/-}$ mice to generate double heterozygotes. The latter were intercrossed to generate compound genotypes - full details are available in the paper by Vives et al. (2006). In view of the changes seen in the SCJs of $p53^{+/-}ASPP2^{-/-}$ mice, and with reference to the conflicting observations of ASPP2 expression with low pH-DCA exposure in OE21 cells, I grew gastric organoids from $ASPP2^{\Delta 3loxP-CreER}$ mice, also from the Lu laboratory, culled at 8 weeks of age. Gastric organoid cultures were then split and re-plated after seven days of culture, with 100nM 4-hydroxy tamoxifen (suspended in corn oil) added to the culture media in 50% of wells to knock out ASPP2 expression and corn oil only in remaining wells (Figure 6.6B). After 24 hours, culture medium was changed (without tamoxifen) and organoids were then imaged every 40 minutes with multiple Z stack time lapse microscopy over seven days, with 48 hourly culture medium changes. This showed some morphogenic differences with restricted branching observed in some ASPP2 deficient organoids (Figure 6.6C). Attempts to stain organoids to confirm ASPP2 knock down were not particularly successful (Figure 6.6D).

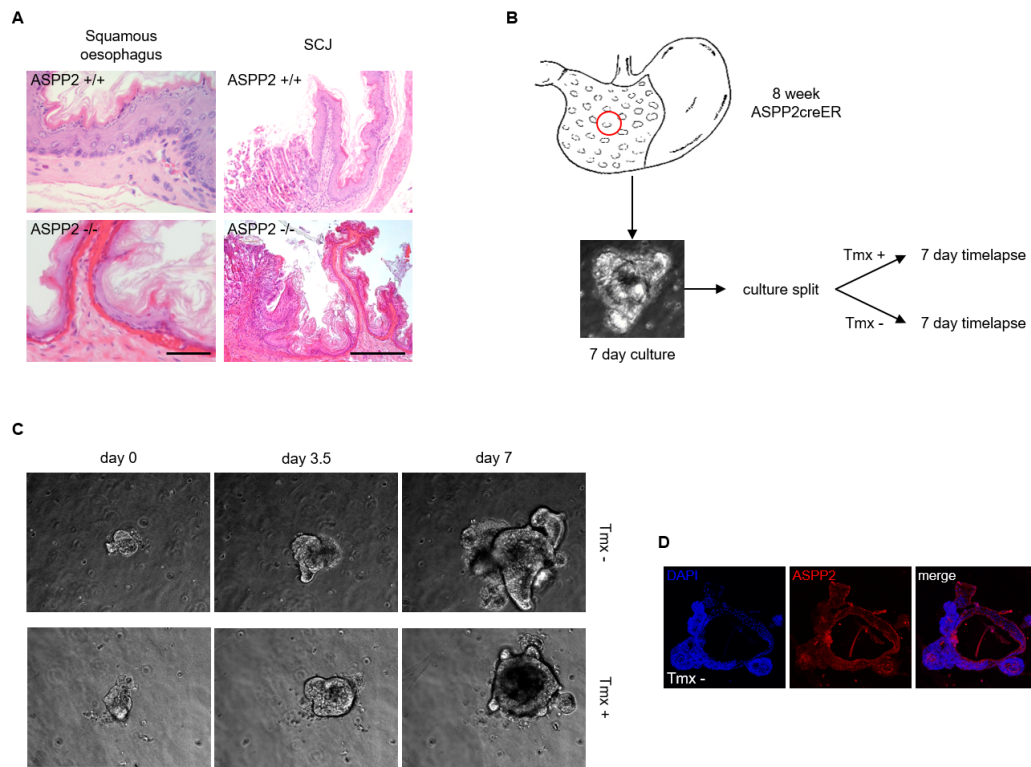


Figure 6.6: Tissue and organoid characterisation of the effects of ASPP2 knock-down in murine SCJ and stomach. (A) H&E stained sections of squamous oesophagus (left side, scale bar 50 μ m) and squamo-columnar junction (right side, scale bar 250 μ m) in ASPP2^{+/+} and ASPP2^{-/-} mice (both also p53^{+/-}). (B) Representation of where in the stomach samples were taken from ASPP2 ^{Δ 3loxP-CreER} mice prior to crypt extraction and organoid culture. After seven days, cultures were split and samples cultured in normal conditions or exposed to 100nM 4-hydroxy tamoxifen in corn oil or corn oil only (control), before continuing culture for seven days with time lapse microscopy. (C) Bright-field time lapse microscopy images of murine gastric organoids, with and without tamoxifen exposure, at initial (day 0), mid point (day 3.5) and final (day 7) imaging times. (D) Attempt at immunofluorescence staining for ASPP2 (red) in murine gastric organoids (tamoxifen naive). SCJ=squamo-columnar junction, Tmx=4-hydroxy tamoxifen.

Before testing the effects of ASPP2 knockdown in a higher throughput screen with similar methodology, a simple macro was developed with Dr Mark Shipman to obtain organoid metrics of size (area), roundness and solidity from maximal intensity images of organoids in each video frame. These measurements are generated from inbuilt imageJ shape descriptor measurement functions. In this limited experiment, this failed to capture any clear differences between ASPP2 ^{Δ 3loxP-CreER} organoids with and without 4-hydroxy tamoxifen treatment (Figure 6.7).

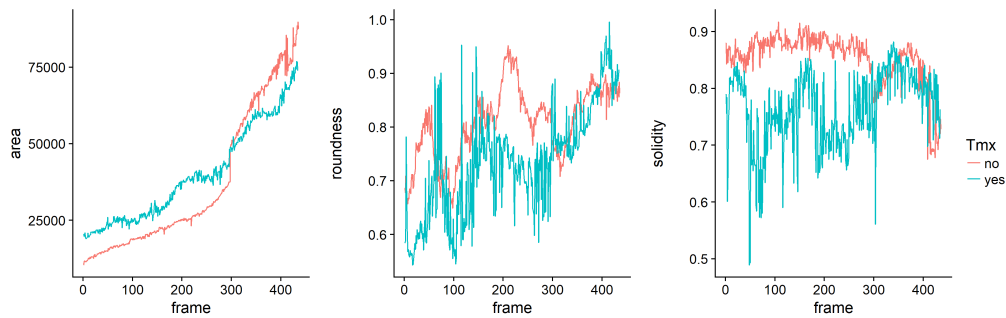


Figure 6.7: Growth dynamics of murine gastric organoids with conditional ASPP2 knockdown. Line plots showing how tamoxifen treated and tamoxifen naive murine ASPP2^{creER} gastric organoids changed over seven days. Metrics were obtained from time lapse microscopy for area, roundness and solidity of organoid, in two dimensions. Tmx=tamoxifen.

6.2.5 Co-culture of squamous and columnar oesophageal cells

Following from earlier discussion regarding the unsuitability of mouse models, due to gastro-oesophageal anatomical differences from human, and the difficulty in controlling experiments *in vivo*, recreating a squamo-columnar junction for testing effects of reflux, or simply observing cell-cell interactions is of some interest. Towards this, columnar derived (OE33) and squamous derived (OE21) cells were plated together, to make initial assessments of how cells interact, and whether this changed the 'squamous-ness' of OE21 cells by reducing p63 expression.

Placing cells in geographically distinct locations with a single point of cell-cell interaction on culture proved very difficult to reliably achieve, so cell membrane dyes (VybrantTM Multicolor Cell-Labeling Kit, ThermoFisher) were used to label OE21 and OE33 cells independently before co-culturing them together with environment controlled fluorescence time lapse microscopy. Results over seven days of microscopy with 30 minute interval imaging showed that both OE21 and OE33 cells moved freely, seemingly independent of each other, although cell labelling diminished over time and dye uptake in OE33 cells was poor (Figure 6.8). When the label was reversed, a similar result was obtained with poor uptake in OE33 cells suggesting this is a cell specific effect, rather than a feature of the dye or imaging technique (data not shown).

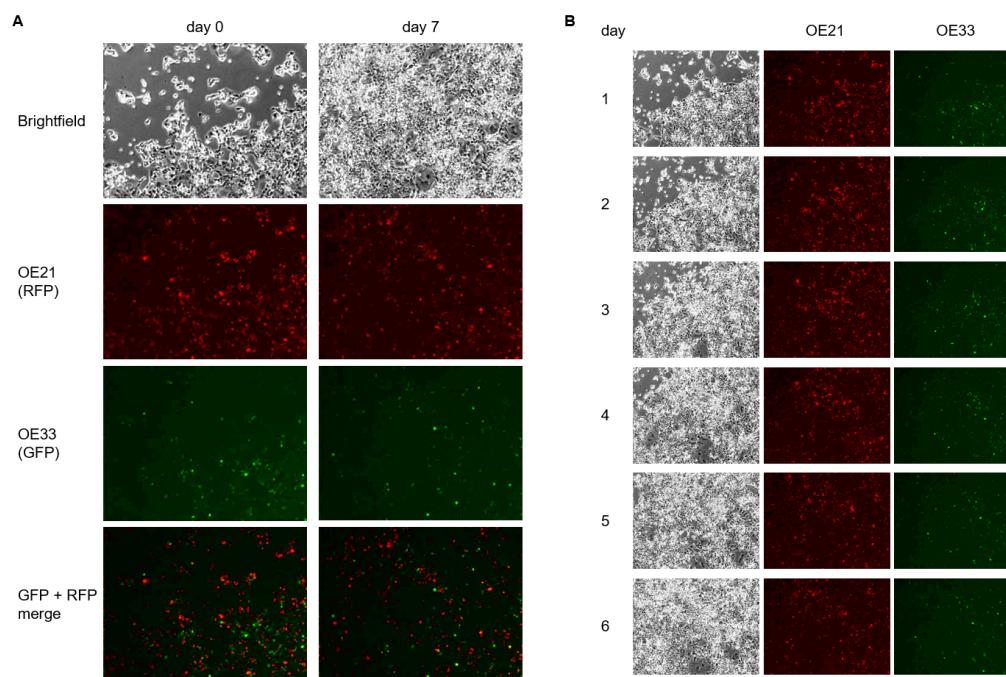


Figure 6.8: Labelled co-culturing of squamous and columnar derived oesophageal cell lines. (A) From top to bottom, bright-field, RFP channel, GFP channel and merged RFP/GFP channel images from first (left side, day 0) and last (right side, day 7) after plating of lipophilic dye-labelled OE21 and OE33 cells together (50:50 split). (B) Bright-field, GFP and RFP images from each 24 hour period of cell co-culture, as indicated by day annotation on left side.

After seven days of OE21-OE33 co-culture, cells were fixed and stained for p63. Confocal microscopy was able to detect remaining membrane dyes and p63 staining, where present (Figure 6.9, main image on left). Notably, where cell types could be determined by membrane colour, OE21 cells did not appear to express p63 when in direct contact with OE33 (Figure 6.9, lower expanded image), however the OE21 cells seen independently from OE33 expressed nuclear p63 (Figure 6.9, upper expanded image). However, this was a rare observation in these data and transfer of membrane labelling dye or 'squamous-isation' of OE33 cells are also possible explanations.

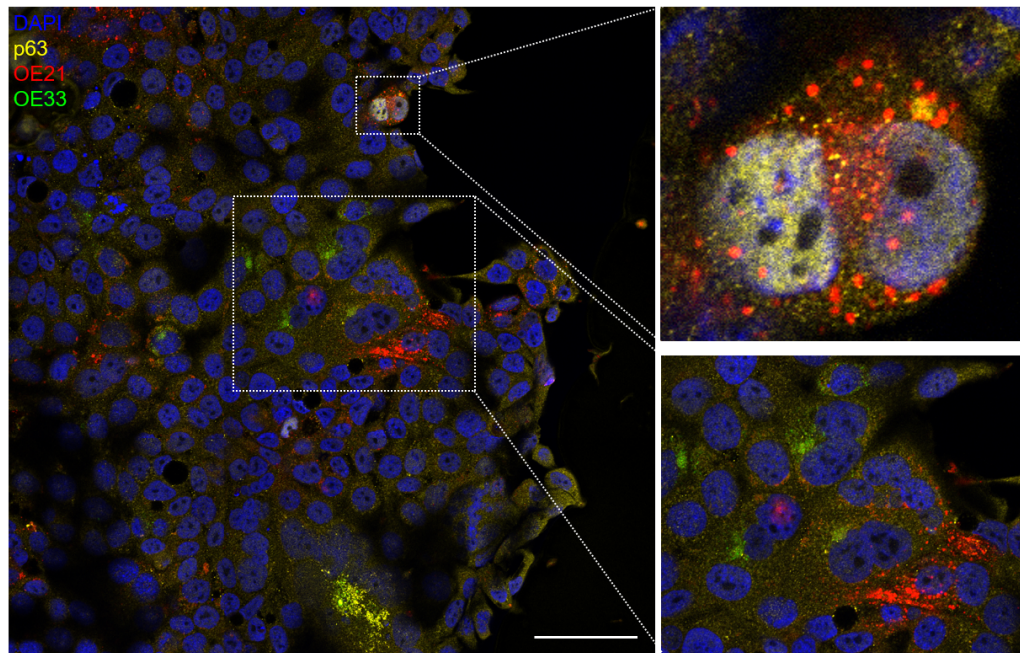


Figure 6.9: Effects of co-culturing squamous and columnar derived oesophageal cell lines on p63 expression. (A) Immunofluorescence staining for p63 (yellow) in fixed cells taken from day 7 of figure 6.8. Fluorescence channels were also used for RFP (OE21, red) and GFP (OE33, green). Expanded images show two OE21 cells in close contact (top), and an area of OE21 and OE33 cell interaction, where the lipophilic dyes have been retained by cells (bottom). Scale bar is 100/*mum*.

6.3 Summary

In this chapter I show p63 suppression in OE21 (oesophageal squamous cell carcinoma) cells with acid-bile exposure and suggest that ASPP2, a regulator of p63, could act through an NF κ B-independent mechanism in OE21 cells, though results were conflicting. I then demonstrate disturbed epithelial homeostasis in p53^{+/-}ASPP2^{-/-} mice, and try to recapitulate this with conditional knock out of ASPP2 in gastric organoids. Finally, using cell co-culture, I briefly investigate the effects of cell-cell interactions between columnar and squamous-origin cell lines on p63 expression. Subsequent experiments with organoid culture, quantitative polymerase chain reaction experiments to confirm ASPP2 knock down, and computational assessment of organoid morphology were undertaken by colleagues as part of their DPhil studies.

Chapter 7

Conclusions

7.1 Single cell RNA-seq in upper gastrointestinal tissues and Barrett's oesophagus

Using single cell RNA-seq data of 4237 cells from eight patients, six with Barrett's oesophagus (BO), I resolved cell sub-populations in gastrointestinal epithelia and show a profound similarity in the RNA composition profile between oesophageal submucosal gland (OSG) cells and BO cells. This is supported by observations that this sub-population of BO and OSG cells express the stem cell-associated gene *OLFM4*, in line with the notion that these populations might contain similar progenitor cells. Glandular epithelial cells are replaced by squamous epithelium during development of the oesophagus and OSGs are functionally important structures formed from remaining glandular epithelium or squamous ingrowth (Long and Orlando, 1999; Rishniw et al., 2011). It is thus not surprising that I identified expression of the developmental gene *LEFTY1* in OSGs, and that as these structures expand during the development of BO, increased levels of *LEFTY1* and *OLFM4* are observed in these tissues. Notably, *LEFTY1* is regulated by $TGF\beta$ signalling and bone morphogenic proteins (BMPs) (Smith et al., 2011; Miyata et al., 2012). Since $TGF\beta$ is often perturbed in BO, and BMPs have been shown to play a major role in the development of BO, it will be interesting to explore these relationships further (Hyland et al., 2014; Mari et al., 2014).

My observations support a previously proposed hypothesis that BO originates from OSGs. This model suggests that acid and bile reflux-induced damage to the oesophagus is repaired by the expansion or selection of OSGs, which have alkaline secretions,

and are thus able to play a role in protecting the oesophagus from gastro-oesophageal reflux damage. Observations I have replicated showing downregulation of p63 in squamous cells on bile acid exposure supports a mechanistic process whereby the enrichment and outgrowth of OSGs into BO could be supported by squamous differentiation cell fate-dependent transcription factor suppression (Figure 7.1).

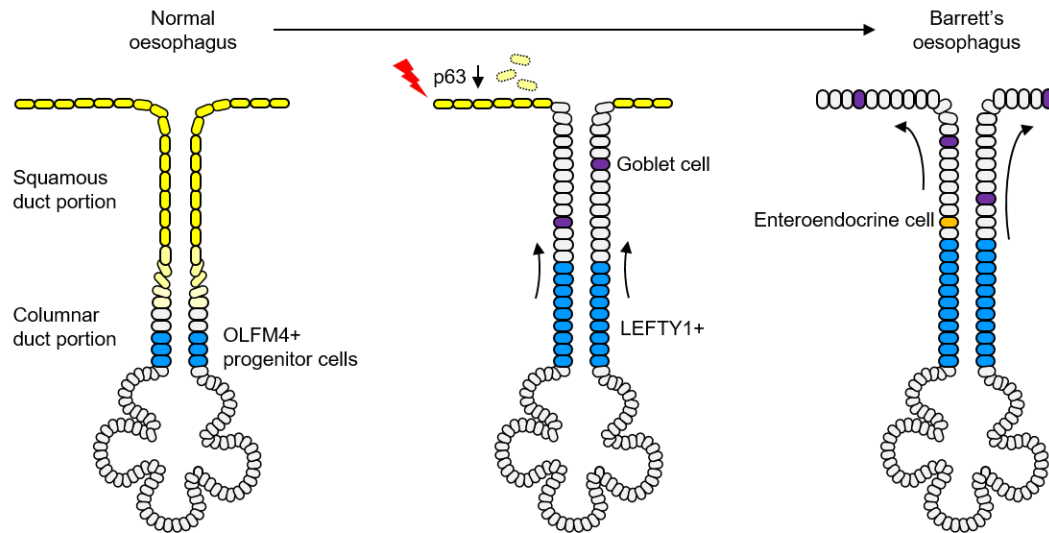


Figure 7.1: Proposed model for the development of Barrett's oesophagus from oesophageal submucosal glands. Cartoons summarising the proposed method by which BO develops from OSGs via expansion of the OLFM4+ progenitor cell pool in the columnar portion of the OSG ducts, with corresponding upregulation of LEFTY1, and bile induced suppression of p63 leading to unrepaired injury to surface squamous epithelium.

Together with the observations of normal stem cell dynamics in BO from StemID and in other work (Sato et al., 2011; Lavery et al., 2014), these data would support that BO is unlikely to develop from specific initiating mutational events, as has been proposed (Orloff et al., 2011), but rather it arises opportunistically when certain environmental conditions are met, such as those encountered in GORD. Previous observations of mutations associated with BO risk are perhaps consequent on these conditions and further study might identify mutational associations with GORD, and perhaps help to identify the as-yet-undefined mutational signatures that relate to acid exposure (Palles et al., 2015; Dong et al., 2018; Alexandrov et al., 2015). Further consideration of the functional overlap of other secretory structures with BO and OSGs, such as salivary and mammary glands may help guide our understanding of an adaptive response to injury that drives metaplasia in a broader context. Studies are also needed to experimentally demonstrate the potential of isolated OSG cells, p63+ or p63- OSGs in particular, to develop into a range of specialised columnar cells, BO

cells and OAC. To do this would require adequate starting material and optimisation of primary cell culture conditions, ideally in an organoid or organotypic set up. In our experience, finding OSGs in endoscopic samples is extremely difficult (we found three out of over 120 biopsies analysed from 80 patients), and imaging the complete OSG with duct is not possible with simple microscopy due to their tortuous nature. Tackling these issues would be valuable as OSGs are also relatively uncharacterised, and in light of our data, careful staining of fixed specimens for OLFM4 and LEFTY1 would help identify the niche within the gland that the progenitor cell occupies. Further staining with p63, KRT5, KRT7 and carbonic anhydrase 4 (CA4) would also help establish whether our observations are consistent with previously described BO origin cells (Wang et al., 2011; Jiang et al., 2017), or whether they represent a new BO source. Histopathological spatial mapping of BO and OAC has been achieved (Lavery et al., 2016), so applying this in OSGs would be invaluable.

Given that rodents lack OSGs, and the current lack of an *in vitro* model of human OSGs, analysis of human biopsies currently provides the most reliable approach to dissect the cell relationships of BO. Future improvements in single cell sequencing techniques may enable more systematic genetic confirmation of the cellular origin of BO through DNA analysis. Now, it would also be possible to perform a much larger analysis using higher throughput scRNA-seq, though sensitivity differences between methods may reduce the clustering analysis power. Importantly, higher throughput scRNA-seq might reduce selection bias inherent in the methodology in this study, especially with respect to gastric cells which were likely to have been detrimentally affected by acid exposure explaining the poorer cell yields seen after quality control filtering. Higher throughput droplet based single cell RNA-seq approaches are fairly recent and were not used in this work as they pre-date my project. Despite the obvious benefits of a higher throughput approach, it would preclude a combined mutational analysis and recent meta-analysis suggests that plate based smart-seq2, as used in my work, provides a highly sensitive approach, but cost and logistic implications clearly constrain the experiment.

It is important to note that this study cannot identify the origins of OAC. Future studies are needed to address the relationship between BO and OAC on a cellular level, and how this relates to recent work suggesting that OAC is highly similar to a sub-set of gastric cancers (Cancer Genome Atlas Research Network, 2017). Lending support to an OSG theory of BO development is valuable, as this would suggest that

BO-associated OAC could derive from a different primary cell source than gastric-related malignancies, where many current OAC therapies are borrowed. Currently, it seems that genetic studies of OAC have been unable to identify whether there are molecular differences in OACs that arise at or below the gastro-oesophageal junction, and are therefore probably gastric in origin, and OACs that develop in previously squamous oesophagus, and probably derive from OSGs or BO. As larger genomic studies are published it will be interesting to see if a pattern develops with regard to OAC and gastro-oesophageal junction location, and whether this impacts on therapeutic efficacy. The majority of OACs present without prior knowledge of BO, and metaplasia around OAC is not universally seen. This could mean that any BO has been completely overtaken by malignancy, or that some cancers may develop from OSGs or tiny areas of BO alone. Previous observations of adenomas and adenocarcinomas developing in OSGs might suggest that the role of BO in OAC development is simply a consequence of there being a larger surface area of susceptible exposed columnar mucosa (Figure 7.2). If true, current surveillance strategies, which rely on the 'size' of BO by Prague classification, should correlate with cancer risk, but may lack the precision of identifying mutational drivers occurring in native oesophageal structures, and observing for this in OSGs could also yield a small but important high risk patient group. The more recent addition to p53 staining to help stratify surveillance timing has begun to address exactly this issue, though this is not yet in widespread use.

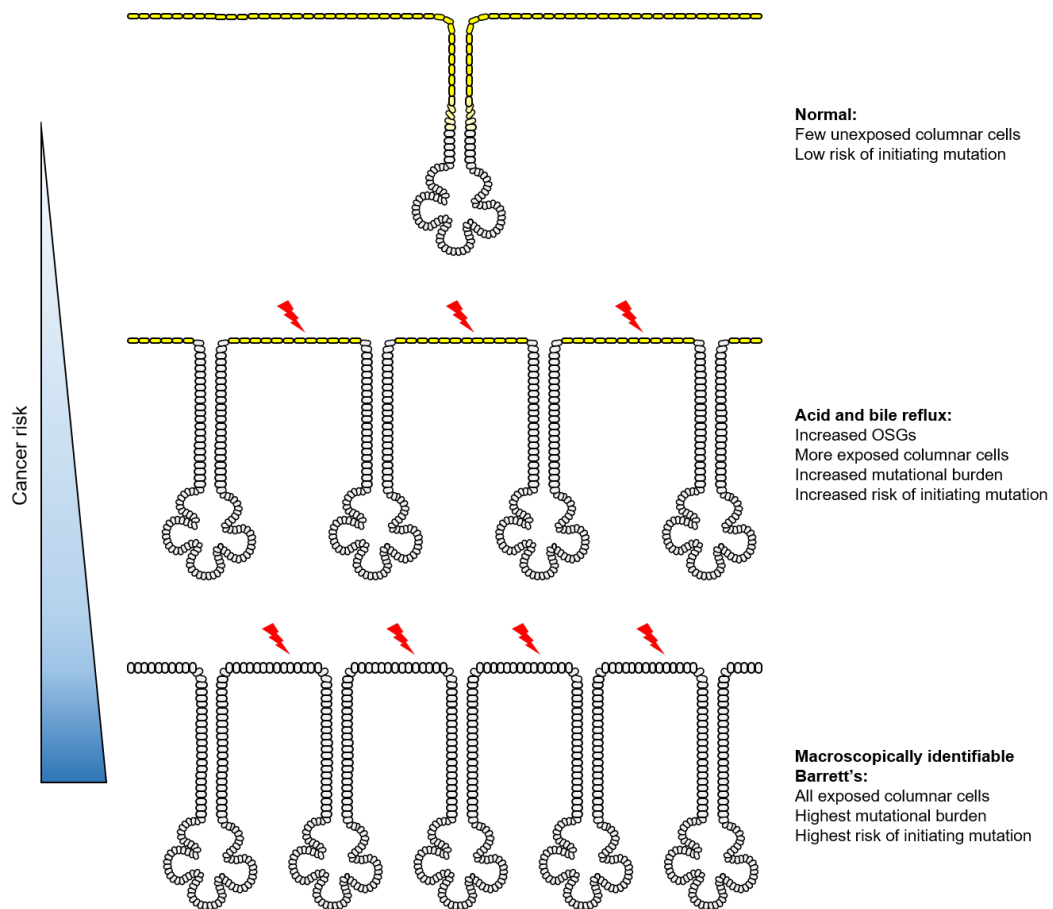


Figure 7.2: Proposed model for oesophageal adenocarcinoma risk. Cartoon demonstrating how the risk of dysplasia or adenocarcinoma developing in the oesophagus could be influenced by increasing numbers of columnar cells and increasing exposure to inflammatory signals through acid and bile exposure, seen from top (normal oesophagus) to bottom (established Barrett's oesophagus).

Beyond BO, this work also shows that goblet cells from normal and pathological cells in the upper gastrointestinal tract share an RNA composition profile. A power of single cell RNA-seq is that it can provide a 'pseudo-biopsy' by clustering similar cells and exploring their gene expression profile beyond known and selected markers. As a demonstration of this, I observed that SPINK4 seems to identify an earlier stage of intestinal metaplasia than marked by MUC2, given that SPINK4 is expressed lower in intestinalised crypts in BO, and SPINK4+ cells are phenotypically atypical for MUC2 expressing goblet cells in BO. Of clinical importance, these results suggest that intestinal goblet cell characteristics exist even in the absence of morphologically identifiable goblet cells, supporting the view that diagnosis of BO should not require the detection of goblet cells, as is still the case in the US.

7.2 Identifying single nucleotide variations in single cell RNA-seq data

While analysis of scRNA-seq data can identify phenotypic similarities between cell populations, showing shared somatic mutations would provide additional evidence of cell-cell relationships. Attempting this naively in single cell RNA-seq data produced vast quantities of single nucleotide variants (SNVs) over the 1046 cells examined, of questionable validity. Taking inspiration from our work using synthetic ERCC spike-ins to model technical variation between samples (Severson et al., 2018), I leveraged the false positive calls made in ERCCs to produce SNV accuracy predictions by classifying each SNV with metadata from a previously designed tool using machine learning approaches to learn 'true' SNVs. I was able to validate these results in human primary cells, using bulk RNA-seq as a positive control, and also in a cell line (K562) with a WGS control data set. Surprisingly, these models could predict SNVs with high accuracy between cell types, and even from primary cell to cell line. Thus, I developed `snv.predict`, an R package to curate SNV classifiers (`snv.bench`), test performance of SNV prediction using multiple machine learning approaches (`snv.performance`), run a tuned random forest model on data (`snv.predict`) and make SNV predictions between data sets where controls might not be available (`cross.predict`).

Using `snv.predict` I demonstrate the previously shown clonal nature of BO, and by overlaying phenotypic details from transcriptional data, show that cell phenotype may have a role in defining some features of clonal heterogeneity, as clones tended towards a Barrett's type or goblet-type phenotype. Since clonal diversity is an important predictor of BO progression, further comparisons of single cell derived clones in longitudinal analyses will be important (Martinez et al., 2016). Additionally, analysis of different pathological stages of BO-OAC (with IM and non IM BO) may develop our understanding of the clonal nature of BO more clearly, specifically whether non IM BO recapitulates the same clonal arrangements seen with goblet cells, providing a further basis for harmonising BO diagnostic criteria.

A weakness in using single cell RNA-seq for mutational calling not addressed here is the high variability of gene expression seen in different cell phenotypes. Considering most single cell RNA-seq experiments are designed to explore exactly this issue, different normalisation processes will need to be applied to prevent hierarchical analysis consistently recapitulating phenotype, though this problem is true of any RNA based SNV calling. This issue is not so important when looking at specific mutations or

mutational signatures in BO alone. My observation of *CDKN2A* mutations in BO is expected, however I also saw rare *TP53* mutations in a single patient, which is much more controversial. Considering the power of single cell approaches to identify rare mutations it is an intriguing idea that single cell approaches could identify *TP53* mutations in benign BO where bulk sequencing methods have not, and perhaps start debate on the role of p53 in initiating and maintaining BO. An orthogonal single cell sequencing approach, or multi region ultra deep sequencing of *TP53* will help clarify my findings. The abundance of mutations of *S100P* in all BO patient is an exciting observation, as this gene has been shown to decrease expression of cytokeratin 8, a key columnar epithelial gene, and therefore further functional studies of *S100P* mutation are warranted (Whiteman et al., 2007).

Anticipating the scepticism likely from focussing on individual mutations called in single cell RNA-seq data without separate validation, I also use mutational signatures to generalise the patterns of mutations to see if a known signature is present in BO. Interestingly, BO cells tended towards two known mutational signatures, both likened to gastric cancer, and one proposed to be a result of DNA damage response (DDR). DDR has previously been proposed as a molecular mechanism of BO development and progression through inactivation of *TP63* and *TP53* respectively (Roman et al., 2007; Das et al., 2011), particularly intriguing in light of p63 suppression in response to GORD conditions. I also observe that the age of one patient with BO, significantly higher than the other three examined, appears to result in a unique mutational signature, although further data are needed to confirm this observation.

Finally, I show that identical SNVs were present BO and OSG cells in one patient, after removing SNVs from matched bulk sequencing, supporting my earlier observations of a direct link between these cell types. Interestingly, the total mutational burden in phenotype-labelled oesophageal cells showed that cells from OSGs were significantly more mutated than normal squamous cells. This suggests that genomic instability may play an earlier role in metaplasia than has been proposed (Watari et al., 2009), at least in this single patient. As a future project, it would be extremely useful to define whether this observation is a product of a reflux-driven selection bias, in which squamous cells are shed while OSG cells survive and accumulate mutations. Understanding this would help to dissect the roles of random and specific mutations and environmental conditions in the development of BO and subsequent mutagenesis and carcinogenesis (Figure 7.2).

Taken together, these findings help characterize BO in humans and demonstrate the power of single cell analysis of clinical samples to uncover biological relationships among cell types and cellular heterogeneity in healthy and diseased tissues. Considering my results supporting a OSG origin of BO, and that OAC is not commonly surrounded by BO, snv.predict or a combined DNA/RNA sequencing approach may provide a better way to investigate when driver mutations occur by isolating columnar type cells from oesophageal biopsies using gene expression profile and then exploring which mutations exist. An easier, cheaper and more targeted approach to this may also be possible by identifying and validating suitable FACS markers from my data and isolating OSG cells for bulk targeted sequencing to map which mutations occur and when in at-risk patients.

7.3 Enhancing the application of single cell RNA-seq in clinical trials

Studies using Smart-seq2 plate based scRNA-seq with large numbers of cells are fairly rare, due to labour-intensiveness and cost. The advent of 3' sequencing and microfluidic platforms has reduced the per cell cost of single cell RNA-seq, while also increasing throughput. Since proponents of plate based single cell RNA-seq argue for its reliability, gene detection sensitivity and high success rate in clinical science where samples are rare and irreplaceable, I endeavoured to modify the workflow we use in our BO data to address some practical issues I encountered. I demonstrate that data quality is equivalent between fresh prepared samples and frozen preserved samples, robotic sample preparation with miniaturised sample reaction volumes and reduced depth sequencing. Using these modifications together substantially increased the feasibility of single cell RNA-seq in an early phase clinical trial where short notice of sample arrival and need to reduce batch effect from separate time processing were features. To prove this, I demonstrate the feasibility of plate based RNA-seq in the LUD2015-005 trial in a single patient. The first cohort of this trial investigates the safety of anti PD-L1 immunotherapy before conventional chemotherapy.

Despite an overall poor data quality, due to prolonged processing time from endoscopy, clustering cells by gene expression showed clear separation between pre and post immunotherapy. Pathway analysis showed that an immune response signature was captured in the post therapy cell cluster, and specific differentially expressed genes confirm a shift from adaptive to innate immunity between time points, with increased

immune infiltration previously correlated with improved survival (Ji et al., 2012). Indeed, this patient showed a reduction in tumour volume in clinical follow up between the two time points analysed. Macrophage recruitment has been associated with restriction of the adaptive immune system's ability to target cancer cells in PD-1 therapy (Peranzoni et al., 2018), however macrophages also express high levels of PD-L1, suggesting anti PD-L1 monotherapy in this situation may have unleashed an effective innate immune attack on OAC cells. *In vitro* confirmation using tumour associated immune cells would help us understand how a more nuanced macrophage response influences OAC, such as how macrophages have polarised in response to anti PD-L1 (Dhupkar et al., 2018).

Whilst these observations can most likely be recapitulated in bulk RNA-seq data, or simple immune cell profiling, cell phenotype heterogeneity remains an exciting avenue of exploration in immunotherapy which may shed light on treatment response.

Appendix A

Differentially expressed genes in cell clusters from each tissue

Clustering analysis of single cells from each tissue analysed by single cell RNA-seq from four patients with Barrett's oesophagus identified four duodenal clusters, three gastric clusters, four oesophageal clusters and four Barrett's clusters. Highly significant upregulated genes in each cluster are shown in this appendix with a manual annotation of cell type. See Materials and Methods for explanation of how PValue thresholds were determined.

A.1 Duodenum cluster D1 - enterocytes

gene_id	external_gene_name	logFC	PValue
ENSG00000187288	CIDEA	6.772233	7.02E-47
ENSG00000214782	MS4A18	6.439932	6.94E-26
ENSG00000173237	C11orf86	6.266525	1.36E-37
ENSG00000182156	ENPP7	6.035019	1.56E-32
ENSG00000110245	APOC3	5.975416	3.47E-34
ENSG00000143595	AQP10	5.954492	8.77E-22
ENSG00000110244	APOA4	5.784948	5.36E-34
ENSG00000160868	CYP3A4	5.645533	4.82E-42
ENSG00000084674	APOB	5.444359	2.76E-35
ENSG00000169906	S100G	5.385164	2.03E-21
ENSG00000104808	DHHD	5.332841	2.01E-22
ENSG00000160870	CYP3A7	5.26889	4.44E-26
ENSG00000172689	MS4A10	5.142668	5.76E-29
ENSG00000244503	RP11-278L15.6	5.016014	5.57E-16
ENSG00000184163	FAM132A	4.969959	5.87E-19
ENSG00000118137	APOA1	4.965345	1.84E-28
ENSG00000111701	APOBEC1	4.944591	1.12E-22
ENSG00000023839	ABCC2	4.903113	1.30E-17
ENSG00000148942	SLC5A12	4.748637	8.40E-15
ENSG00000163295	ALPI	4.658962	8.51E-21
ENSG00000223865	HLA-DPB1	4.649173	6.58E-15
ENSG00000076351	SLC46A1	4.628649	8.03E-20
ENSG00000115850	LCT	4.547508	5.55E-27
ENSG00000088386	SLC15A1	4.394417	1.03E-24
ENSG00000114455	HHLA2	4.381431	3.05E-18

gene_id	external_gene_name	logFC	PValue
ENSG00000118094	TREH	4.280211	3.51E-18
ENSG00000169903	TM4SF4	4.175157	3.81E-19
ENSG00000091138	SLC26A3	4.02518	1.04E-21
ENSG00000074276	CDHR2	3.97604	5.11E-19
ENSG00000118777	ABCG2	3.936404	5.00E-16
ENSG00000154646	TMPRSS15	3.899449	6.50E-18
ENSG00000182938	OTOP3	3.853617	6.32E-16
ENSG00000141434	MEP1B	3.792204	3.84E-15
ENSG00000142583	SLC2A5	3.787927	8.63E-21
ENSG00000107798	LIPA	3.765233	1.24E-17
ENSG00000130234	ACE2	3.727181	3.17E-16
ENSG00000104267	CA2	3.547486	9.36E-19
ENSG00000008517	IL32	3.502063	5.53E-16
ENSG00000136872	ALDOB	3.43864	6.44E-23
ENSG00000124253	PCK1	3.333954	6.28E-16
ENSG00000100170	SLC5A1	3.313218	1.02E-21
ENSG00000173702	MUC13	2.788719	4.70E-16
ENSG00000162482	AKR7A3	2.622521	2.79E-15

A.2 Duodenum cluster D2 - enteroendocrine cells

gene_id	external_gene_name	logFC	PValue
ENSG00000100604	CHGA	9.813004	4.33E-72
ENSG00000016490	CLCA1	9.488249	2.31E-90
ENSG00000171951	SCG2	9.467887	2.29E-68
ENSG000000089199	CHGB	9.213406	1.59E-62
ENSG00000198788	MUC2	9.072622	3.23E-101
ENSG00000162998	FRZB	8.887019	2.71E-64
ENSG00000145321	GC	8.818897	9.95E-55
ENSG00000179914	ITLN1	8.51476	2.52E-72
ENSG00000129167	TPH1	8.17198	2.29E-37
ENSG00000166922	SCG5	8.030384	1.13E-36
ENSG00000163497	FEV	7.728097	8.90E-28
ENSG00000163515	RETNLB	7.639118	1.43E-28
ENSG00000079689	SCGN	7.582934	1.19E-30
ENSG00000162992	NEUROD1	7.506706	3.02E-28
ENSG00000118271	TTR	7.446156	8.96E-28
ENSG00000163499	CRYBA2	7.423655	5.66E-27
ENSG00000064787	BCAS1	7.35693	2.56E-37
ENSG00000198691	ABCA4	7.178766	1.61E-25
ENSG00000175426	PCSK1	7.01268	1.89E-22
ENSG00000146674	IGFBP3	6.934361	1.90E-20
ENSG00000109472	CPE	6.757564	1.56E-24
ENSG00000122711	SPINK4	6.753938	3.79E-36
ENSG00000132639	SNAP25	6.699171	6.77E-25
ENSG00000108551	RASD1	6.543521	6.51E-24
ENSG00000134193	REG4	6.521098	3.30E-54
ENSG00000125851	PCSK2	6.519045	3.27E-18
ENSG00000172238	ATOH1	6.444416	3.48E-24
ENSG00000167964	RAB26	6.368088	6.86E-24
ENSG00000178538	CA8	6.243446	7.94E-19
ENSG00000204983	PRSS1	6.216465	2.10E-18
ENSG00000178695	KCTD12	5.865969	4.50E-17
ENSG00000160180	TFF3	5.864314	4.45E-57
ENSG00000184709	LRRC26	5.759057	4.89E-24
ENSG00000102109	PCSK1N	5.628301	6.78E-14
ENSG00000060709	RIMBP2	5.482717	5.81E-14
ENSG00000074527	NTN4	5.340119	4.26E-14
ENSG00000128917	DLL4	5.283551	6.05E-15
ENSG00000198719	DLL1	5.278852	7.29E-14
ENSG00000146205	ANO7	5.222765	1.75E-12
ENSG00000143797	MBOAT2	5.073472	4.04E-13
ENSG00000107281	NPDC1	4.96918	1.75E-14
ENSG00000122176	FMOD	4.959846	3.85E-12

gene_id	external_gene_name	logFC	PValue
ENSG00000129538	RNASE1	4.959785	2.35E-23
ENSG00000076864	RAP1GAP	4.921999	8.60E-14
ENSG00000144485	HES6	4.882611	8.81E-17
ENSG00000069974	RAB27A	4.684446	1.84E-11
ENSG00000126709	IFI6	4.673454	1.37E-11
ENSG00000167748	KLK1	4.592894	1.47E-12
ENSG00000111319	SCNN1A	4.573488	2.32E-12
ENSG00000157613	CREB3L1	4.18114	6.96E-12
ENSG00000174992	ZG16	4.175327	1.58E-14
ENSG00000070526	ST6GALNAC1	4.061795	3.84E-16
ENSG00000115756	HPCAL1	4.029924	3.01E-17
ENSG00000148180	GSN	4.026383	3.62E-12

A.3 Duodenum cluster D3 - crypt cells

gene_id	external_gene_name	logFC	PValue
ENSG00000102837	OLFM4	6.009896	1.27E-48
ENSG00000144354	CDCA7	5.807554	5.13E-30
ENSG00000168748	CA7	5.125482	1.33E-09
ENSG00000131153	GINS2	4.479937	1.94E-12
ENSG00000184178	SCFD2	4.346936	5.46E-10
ENSG00000152785	BMP3	4.144628	6.39E-12
ENSG00000136997	MYC	3.997686	5.41E-14
ENSG00000049541	RFC2	3.848504	5.69E-08
ENSG00000167513	CDT1	3.835981	6.42E-09
ENSG00000073536	NLE1	3.804046	3.21E-08
ENSG00000136875	PRPF4	3.684516	3.75E-08
ENSG00000163517	HDAC11	3.623571	1.58E-07
ENSG00000111641	NOP2	3.593746	1.75E-08
ENSG00000167747	C19orf48	3.592617	1.16E-12
ENSG00000118513	MYB	3.55406	4.01E-10
ENSG00000187908	DMBT1	3.529512	1.56E-25
ENSG00000166483	WEE1	3.440569	1.05E-07
ENSG00000149573	MPZL2	3.36007	1.88E-07
ENSG00000073111	MCM2	3.258082	2.49E-07
ENSG00000177426	TGIF1	3.208098	2.12E-08
ENSG00000130147	SH3BP4	3.138604	1.60E-07
ENSG00000134757	DSG3	3.110402	1.28E-07
ENSG00000166803	KIAA0101	3.08008	1.13E-07
ENSG00000128050	PAICS	3.027147	6.24E-11
ENSG00000185885	IFITM1	2.972194	2.35E-10
ENSG00000048162	NOP16	2.965151	3.12E-10
ENSG00000174136	RGMB	2.942336	6.01E-09
ENSG00000156471	PTDSS1	2.912882	2.79E-08
ENSG00000240583	AQP1	2.87119	2.97E-10
ENSG00000136717	BIN1	2.834095	3.17E-07
ENSG00000172016	REG3A	2.828941	1.57E-09
ENSG00000122779	TRIM24	2.755135	8.94E-08
ENSG00000101361	NOP56	2.720506	9.22E-09
ENSG00000115386	REG1A	2.69834	3.70E-18
ENSG00000166508	MCM7	2.61749	1.37E-08
ENSG00000205358	MT1H	2.572654	2.10E-08
ENSG00000126602	TRAP1	2.571435	3.29E-08
ENSG00000128951	DUT	2.525586	2.09E-07
ENSG00000130826	DKC1	2.51728	3.91E-07
ENSG00000187134	AKR1C1	2.363523	4.25E-09
ENSG00000165732	DDX21	2.352263	2.95E-07
ENSG00000135316	SYNCRIP	2.291286	1.42E-08
ENSG00000162599	NFIA	2.251402	2.06E-08
ENSG00000131142	CCL25	2.13433	1.08E-13

A.4 Duodenum cluster D4 - non epithelial cells

gene_id	external_gene_name	logFC	PValue
ENSG00000132465	JCHAIN	9.710531	3.97E-75
ENSG00000170476	MZB1	9.39588	2.75E-60
ENSG00000211899	IGHM	8.785804	5.27E-49
ENSG00000211890	IGHA2	8.560104	2.13E-51
ENSG00000211949	IGHV3-23	8.24168	1.99E-41
ENSG00000026751	SLAMF7	8.063692	1.62E-36
ENSG00000211895	IGHA1	7.746403	1.21E-45
ENSG00000239855	IGKV1-6	7.4463	7.66E-28
ENSG00000167077	MEI1	7.234189	1.64E-30
ENSG00000048462	TNFRSF17	6.972045	5.63E-30
ENSG00000211592	IGKC	6.970546	6.02E-29
ENSG00000137265	IRF4	6.967371	1.43E-23
ENSG00000143297	FCRL5	6.097617	2.31E-19
ENSG00000102096	PIM2	5.995711	1.27E-19
ENSG00000230006	ANKRD36BP2	5.989733	4.62E-15
ENSG00000152689	RASGRP3	5.964261	2.72E-17
ENSG00000099958	DERL3	5.554701	5.40E-16
ENSG00000183508	FAM46C	4.681497	1.75E-10
ENSG00000051108	HERPUD1	3.247964	3.57E-10

A.5 Gastric cluster G1 - enteroendocrine cells

gene_id	external_gene_name	logFC	PValue
ENSG00000079689	SCGN	8.477593	1.32E-66
ENSG00000145198	VWA5B2	8.254949	3.60E-54
ENSG00000118271	TTR	8.198001	5.90E-50
ENSG00000109472	CPE	8.194048	1.62E-66
ENSG00000102109	PCSK1N	8.163755	6.33E-63
ENSG00000054356	PTPRN	8.088903	8.45E-54
ENSG00000089199	CHGB	7.830879	2.85E-39
ENSG00000114115	RBP1	7.66518	2.01E-40
ENSG00000177669	MBOAT4	7.639084	4.54E-37
ENSG00000176956	LY6H	7.622422	3.22E-50
ENSG00000185002	RFX6	7.537414	9.52E-41
ENSG00000132639	SNAP25	7.370717	6.22E-39
ENSG00000133169	BEX1	7.29813	5.36E-39
ENSG00000165507	C10orf10	7.216046	1.48E-38
ENSG00000136698	CFC1	7.102495	1.54E-28
ENSG00000157017	GHRL	6.940161	1.65E-44
ENSG00000148798	INA	6.85591	2.57E-31
ENSG00000139219	COL2A1	6.777976	2.45E-23
ENSG00000169213	RAB3B	6.553235	1.89E-28
ENSG00000058404	CAMK2B	6.537027	9.26E-26
ENSG00000007516	BAIAP3	6.474654	3.31E-22
ENSG00000176463	SLCO3A1	6.397368	4.47E-26
ENSG00000109072	VTN	6.34053	3.22E-19
ENSG00000104112	SCG3	6.274672	9.51E-24
ENSG00000126218	F10	6.265717	8.50E-25
ENSG00000133958	UNC79	6.217056	2.60E-23
ENSG00000139767	SRRM4	6.207809	5.58E-18
ENSG00000166922	SCG5	6.19645	4.91E-19
ENSG00000091428	RAPGEF4	6.18352	7.85E-24
ENSG00000188175	HEPACAM2	6.140299	1.97E-21
ENSG00000162992	NEUROD1	6.130387	1.83E-19
ENSG00000162552	WNT4	5.983102	1.18E-18
ENSG00000151726	ACSL1	5.946345	1.79E-25
ENSG00000175426	PCSK1	5.937333	2.48E-17
ENSG00000060709	RIMBP2	5.927499	2.25E-20
ENSG00000116774	OLFML3	5.918355	4.34E-15
ENSG00000197594	ENPP1	5.744487	1.19E-18
ENSG00000181234	TMEM132C	5.705999	2.13E-15

gene_id	external_gene_name	logFC	PValue
ENSG00000117707	PROX1	5.670002	3.11E-16
ENSG00000188130	MAPK12	5.649954	1.34E-17
ENSG00000133878	DUSP26	5.543098	1.43E-14
ENSG00000130294	KIF1A	5.518187	1.25E-16
ENSG00000139970	RTN1	5.484362	4.37E-16
ENSG00000120457	KCNJ5	5.477779	1.76E-13
ENSG00000069535	MAOB	5.471301	1.68E-13
ENSG00000153558	FBXL2	5.450384	2.02E-13
ENSG00000140323	DISP2	5.427908	2.22E-14
ENSG00000104723	TUSC3	5.38455	1.86E-14
ENSG00000111674	ENO2	5.367859	1.16E-14
ENSG00000077782	FGFR1	5.328956	2.34E-14
ENSG00000169903	TM4SF4	5.26362	8.58E-13
ENSG00000118137	APOA1	5.206163	2.62E-13
ENSG00000145321	GC	5.191757	5.27E-17
ENSG00000178695	KCTD12	5.187902	1.20E-14
ENSG00000100604	CHGA	4.98595	2.69E-11
ENSG00000167964	RAB26	4.9221	1.83E-11
ENSG00000102181	CD99L2	4.904141	1.23E-11
ENSG00000136854	STXBP1	4.770088	5.87E-11
ENSG00000211445	GPX3	4.625438	2.22E-11
ENSG00000016402	IL20RA	4.514668	2.56E-12
ENSG00000186318	BACE1	4.457159	3.42E-11
ENSG00000197249	SERPINA1	4.375002	5.50E-11
ENSG00000235162	C12orf75	2.96163	3.54E-11

A.6 Gastric cluster G2 - surface mucus cells

gene_id	external_gene_name	logFC	PValue
ENSG00000171431	KRT20	5.081951	2.21E-14
ENSG00000145908	ZNF300	4.521963	8.75E-12
ENSG00000170312	CDK1	4.467884	8.03E-12
ENSG00000105641	SLC5A5	4.349407	2.83E-13
ENSG00000100162	CENPM	3.824912	7.97E-09
ENSG00000169605	GKN1	3.644098	2.30E-16
ENSG00000198901	PRC1	3.609446	7.04E-08
ENSG00000233041	PHGR1	3.550346	2.62E-12
ENSG00000183607	GKN2	3.41494	2.24E-14
ENSG00000164611	PTTG1	3.369752	9.24E-08
ENSG00000187134	AKR1C1	3.288943	2.07E-17
ENSG00000111863	ADTRP	3.216271	2.10E-09
ENSG00000166920	C15orf48	3.117646	1.14E-15
ENSG00000151632	AKR1C2	2.935674	9.13E-09
ENSG00000107731	UNC5B	2.845145	3.68E-08
ENSG00000172831	CES2	2.815641	4.92E-10
ENSG00000137628	DDX60	2.809172	4.40E-10
ENSG00000173467	AGR3	2.738813	4.10E-09
ENSG00000131620	ANO1	2.737553	7.74E-08
ENSG00000125775	SDCBP2	2.727392	1.30E-09
ENSG00000090530	P3H2	2.69901	7.84E-08
ENSG00000105835	NAMPT	2.600729	1.19E-07
ENSG00000240065	PSMB9	2.588357	1.04E-09
ENSG00000127528	KLF2	2.412352	1.38E-07
ENSG00000160182	TFF1	2.235229	5.68E-12
ENSG00000198074	AKR1B10	2.229954	2.28E-11
ENSG00000074370	ATP2A3	2.209345	5.84E-08
ENSG00000224259	LINC01133	2.120357	8.86E-08
ENSG00000174501	ANKRD36C	2.087612	3.15E-09
ENSG00000173597	SULT1B1	2.086574	1.12E-07
ENSG00000243955	GSTA1	2.069225	1.10E-07
ENSG00000155380	SLC16A1	2.010588	1.62E-07

A.7 Gastric cluster G3 - mucus neck cells

gene_id	external_gene_name	logFC	PValue
ENSG00000118017	A4GNT	5.649412	1.77E-15
ENSG00000184226	PCDH9	5.332498	9.65E-14
ENSG0000012223	LTf	5.29974	2.58E-11
ENSG00000184956	MUC6	5.159393	2.75E-25
ENSG00000138115	CYP2C8	4.722615	3.87E-09
ENSG00000244405	ETV5	4.197573	9.80E-09
ENSG00000158296	SLC13A3	4.156613	3.57E-08
ENSG00000122176	FMOD	3.544766	2.14E-12
ENSG00000153446	C16orf89	3.50261	1.01E-07
ENSG00000229314	ORM1	3.405963	1.44E-08
ENSG00000182333	LIPF	3.326237	1.55E-14
ENSG00000229859	PGA3	3.114288	1.40E-18
ENSG00000229183	PGA4	3.107755	5.40E-19
ENSG00000225630	MTND2P28	2.628557	7.48E-10

A.8 Oesophagus cluster O1 - squamous cells

gene_id	external_gene_name	logFC	PValue
ENSG00000169245	CXCL10	8.387438	1.21E-67
ENSG00000138755	CXCL9	7.948493	9.20E-77
ENSG00000173432	SAA1	7.88148	7.16E-130
ENSG00000131203	IDO1	7.686289	1.87E-65
ENSG00000204252	HLA-DOA	6.976924	3.65E-44
ENSG00000196735	HLA-DQA1	6.805577	6.08E-48
ENSG00000182326	C1S	6.801503	2.21E-71
ENSG00000164683	HEY1	6.427516	8.75E-45
ENSG00000149131	SERPING1	6.270712	6.71E-62
ENSG00000114115	RBP1	6.231378	9.98E-92
ENSG00000143341	HMCN1	6.194237	8.26E-31
ENSG00000169248	CXCL11	5.959261	4.54E-31
ENSG00000154451	GBP5	5.903161	8.60E-30
ENSG00000162493	PDPN	5.736755	4.22E-27
ENSG00000064886	CHI3L2	5.710386	1.24E-17
ENSG00000104368	PLAT	5.668957	1.83E-38
ENSG00000130303	BST2	5.66123	4.53E-62
ENSG00000159403	C1R	5.557266	2.05E-48
ENSG00000182492	BGN	5.41212	9.62E-45
ENSG00000225630	MTND2P28	5.361742	1.19E-127
ENSG00000140459	CYP11A1	5.346061	6.28E-23
ENSG0000013016	EHD3	5.114874	1.90E-28
ENSG00000206073	SERPINB4	5.105944	1.73E-34
ENSG00000162692	VCAM1	5.037664	2.71E-15
ENSG00000109472	CPE	4.913475	1.09E-21
ENSG00000041982	TNC	4.878504	8.21E-22
ENSG00000050730	TNIP3	4.845309	4.87E-21
ENSG00000087494	PTHLH	4.83791	2.59E-29
ENSG00000198502	HLA-DRB5	4.720672	7.18E-45
ENSG00000223865	HLA-DPB1	4.703518	1.36E-35
ENSG00000125730	C3	4.455689	1.03E-14
ENSG00000162654	GBP4	4.451316	1.98E-26
ENSG00000185860	C1orf110	4.386571	8.86E-24
ENSG00000135919	SERPINE2	4.365099	1.33E-26
ENSG00000179344	HLA-DQB1	4.352668	1.39E-35
ENSG00000078081	LAMP3	4.319174	1.79E-19
ENSG00000137573	SULF1	4.316083	1.06E-14
ENSG00000231389	HLA-DPA1	4.30905	2.66E-26
ENSG00000104332	SFRP1	4.251612	5.93E-20
ENSG00000186891	TNFRSF18	4.190832	2.84E-19
ENSG00000119917	IFIT3	4.163434	6.32E-19
ENSG00000205076	LGALS7	4.15722	5.15E-60
ENSG00000117228	GBP1	4.15505	8.09E-26

gene_id	external_gene_name	logFC	PValue
ENSG00000239713	APOBEC3G	4.124108	6.35E-15
ENSG00000167779	IGFBP6	4.092977	7.24E-24
ENSG00000105928	DFNA5	4.065871	1.16E-12
ENSG00000000971	CFH	3.992825	9.05E-21
ENSG00000025708	TYMP	3.909291	3.70E-25
ENSG00000137959	IFI44L	3.882926	3.03E-18
ENSG00000204287	HLA-DRA	3.88051	1.51E-29
ENSG00000057149	SERPINB3	3.874022	3.25E-20
ENSG00000050165	DKK3	3.801593	5.69E-17
ENSG00000164308	ERAP2	3.786442	5.13E-25
ENSG00000186867	QRFPR	3.764881	2.19E-14
ENSG00000196126	HLA-DRB1	3.726079	1.00E-27
ENSG00000140105	WARS	3.69273	7.17E-35
ENSG00000168306	ACOX2	3.634628	6.16E-14
ENSG00000115415	STAT1	3.600106	1.49E-33
ENSG00000133321	RARRES3	3.586454	2.21E-24
ENSG00000140511	HAPLN3	3.557573	1.10E-12
ENSG00000079215	SLC1A3	3.477487	2.20E-16
ENSG00000116661	FBXO2	3.427935	1.11E-12
ENSG00000157601	MX1	3.427845	2.16E-22
ENSG00000187054	TMPRSS11A	3.357445	9.15E-13
ENSG00000106278	PTPRZ1	3.342742	3.12E-22
ENSG00000168394	TAP1	3.338009	2.95E-31
ENSG00000242574	HLA-DMB	3.29939	5.99E-14
ENSG00000137965	IFI44	3.276539	3.33E-19
ENSG00000156587	UBE2L6	3.253608	6.74E-25
ENSG00000206337	HCP5	3.224097	5.64E-14
ENSG00000100097	LGALS1	3.153512	7.13E-15
ENSG00000162645	GBP2	3.147092	7.67E-17
ENSG00000186377	CYP4X1	3.122252	1.40E-16
ENSG00000186847	KRT14	3.115166	6.01E-14
ENSG00000111335	OAS2	3.041996	3.52E-14
ENSG00000002549	LAP3	3.024514	3.99E-22
ENSG00000163874	ZC3H12A	2.999242	2.15E-13
ENSG00000101347	SAMHD1	2.976792	5.07E-18
ENSG00000132274	TRIM22	2.959712	2.59E-14
ENSG00000126709	IFI6	2.931679	6.31E-17
ENSG00000100342	APOL1	2.884573	7.44E-13
ENSG00000121064	SCPEP1	2.782189	3.61E-15
ENSG00000128422	KRT17	2.745172	1.42E-13
ENSG00000112096	SOD2	2.708377	1.62E-23
ENSG00000185885	IFITM1	2.658544	4.89E-20
ENSG00000062038	CDH3	2.566323	3.29E-15
ENSG00000160932	LY6E	2.557552	9.42E-26
ENSG00000240065	PSMB9	2.549671	1.49E-20
ENSG00000110852	CLEC2B	2.484418	4.86E-15
ENSG00000131669	NINJ1	2.459842	1.29E-12
ENSG00000075618	FSCN1	2.395697	1.22E-14
ENSG00000069849	ATP1B3	2.354273	6.97E-22
ENSG00000204642	HLA-F	2.243527	1.08E-13
ENSG00000163565	IFI16	2.226203	1.87E-13
ENSG00000135821	GLUL	2.163651	2.21E-14
ENSG00000244509	APOBEC3C	2.043895	8.22E-13
ENSG00000154639	CXADR	2.017401	7.76E-15
ENSG00000122378	FAM213A	2.001455	2.30E-17

A.9 Oesophagus cluster O2 - squamous cells

gene_id	external_gene_name	logFC	PValue
ENSG00000198848	CES1	7.970745	2.14E-62
ENSG00000170477	KRT4	6.242512	1.89E-30
ENSG00000167768	KRT1	6.079779	3.17E-34
ENSG00000178934	LGALS7B	5.721497	9.70E-79
ENSG00000229807	XIST	5.592307	8.65E-39

gene_id	external_gene_name	logFC	PValue
ENSG00000244122	UGT1A7	5.048782	1.75E-33
ENSG00000167754	KLK5	4.851131	5.51E-17
ENSG00000166535	A2ML1	4.699258	3.23E-18
ENSG00000134760	DSG1	4.665446	3.96E-20
ENSG00000196344	ADH7	4.657118	5.37E-33
ENSG00000167759	KLK13	4.003741	2.12E-15
ENSG00000095303	PTGS1	3.785943	1.89E-15
ENSG00000151632	AKR1C2	3.485303	2.28E-21
ENSG00000121552	CSTA	3.465886	4.65E-27
ENSG00000234964	FABP5P7	2.839058	9.15E-33
ENSG00000170465	KRT6C	2.813417	3.51E-15

A.10 Oesophagus cluster O3 - columnar cells

gene_id	external_gene_name	logFC	PValue
ENSG00000243709	LEFTY1	7.04773	1.24E-61
ENSG00000169876	MUC17	6.968872	9.53E-53
ENSG00000163734	CXCL3	6.197434	3.76E-32
ENSG00000168631	DPCR1	6.180643	6.86E-24
ENSG00000165376	CLDN2	5.920001	1.74E-21
ENSG00000013588	GPRC5A	5.836924	6.71E-33
ENSG00000081041	CXCL2	5.835352	5.86E-27
ENSG00000166869	CHP2	5.698615	2.82E-31
ENSG00000104783	KCNN4	5.696447	2.60E-26
ENSG00000163739	CXCL1	5.645146	3.93E-22
ENSG00000132911	NMUR2	5.61364	2.38E-19
ENSG00000163586	FABP1	5.609402	7.32E-25
ENSG00000198203	SULT1C2	5.609366	7.59E-36
ENSG00000140297	GCNT3	5.540084	2.07E-29
ENSG00000118017	A4GNT	5.46366	6.81E-20
ENSG00000168955	TM4SF20	5.405423	5.51E-25
ENSG00000102837	OLFM4	5.297437	1.83E-28
ENSG00000095539	SEMA4G	5.245036	1.11E-25
ENSG00000101049	SGK2	5.169435	6.35E-20
ENSG00000186854	TRABD2A	5.116574	3.78E-18
ENSG00000137860	SLC28A2	5.063506	6.30E-20
ENSG00000204616	TRIM31	5.026925	3.16E-18
ENSG00000165140	FBP1	4.967006	4.34E-28
ENSG00000163817	SLC6A20	4.907507	9.21E-17
ENSG00000166920	C15orf48	4.897107	1.01E-19
ENSG00000132437	DDC	4.879197	4.29E-23
ENSG00000174358	SLC6A19	4.877767	1.37E-13
ENSG00000166825	ANPEP	4.876361	8.23E-26
ENSG00000162482	AKR7A3	4.83856	5.69E-25
ENSG00000148735	PLEKHS1	4.820675	3.73E-17
ENSG00000134240	HMGCS2	4.787414	1.65E-24
ENSG00000187908	DMBT1	4.78559	1.18E-25
ENSG00000010438	PRSS3	4.64903	9.40E-18
ENSG00000166391	MOGAT2	4.604964	2.20E-14
ENSG00000149150	SLC43A1	4.603235	2.08E-14
ENSG00000143375	CGN	4.59637	1.15E-16
ENSG00000143167	GPA33	4.556045	2.83E-17
ENSG00000156510	HKDC1	4.541407	8.33E-17
ENSG00000166866	MYO1A	4.472363	1.32E-17
ENSG00000118322	ATP10B	4.46254	1.11E-19
ENSG00000109511	ANXA10	4.447844	7.93E-19
ENSG00000105650	PDE4C	4.445967	1.55E-21
ENSG00000114378	HYAL1	4.441855	2.89E-13
ENSG00000196188	CTSE	4.433649	7.36E-18
ENSG00000151715	TMEM45B	4.419874	8.03E-23
ENSG00000141655	TNFRSF11A	4.401908	2.38E-19
ENSG00000119547	ONECUT2	4.379691	2.97E-18
ENSG00000021826	CPS1	4.184677	3.20E-14
ENSG00000099812	MISP	4.184404	3.28E-16

gene_id	external_gene_name	logFC	PValue
ENSG00000105289	TJP3	4.126298	1.58E-14
ENSG00000184012	TMPRSS2	4.103417	5.39E-16
ENSG00000173597	SULT1B1	4.045564	1.83E-13
ENSG00000066405	CLDN18	4.041514	4.70E-17
ENSG00000105388	CEACAM5	4.03265	8.13E-14
ENSG00000169894	MUC3A	3.981715	9.63E-20
ENSG00000086696	HSD17B2	3.95239	5.41E-14
ENSG00000173702	MUC13	3.861594	2.55E-15
ENSG00000117983	MUC5B	3.832446	1.56E-14
ENSG00000138193	PLCE1	3.829787	1.69E-14
ENSG00000079112	CDH17	3.814243	2.28E-15
ENSG00000164237	CMBL	3.774997	3.05E-19
ENSG00000149260	CAPN5	3.731458	2.54E-13
ENSG00000233041	PHGR1	3.667145	2.29E-13
ENSG00000196352	CD55	3.559487	2.72E-13
ENSG00000185499	MUC1	3.559059	2.16E-14
ENSG00000101076	HNF4A	3.517497	1.40E-13
ENSG00000077157	PPP1R12B	3.438232	2.46E-13
ENSG00000175538	KCNE3	3.396245	1.82E-13
ENSG00000181649	PHLDA2	2.551232	2.00E-14

A.11 Oesophagus cluster O4 - columnar cells

gene_id	external_gene_name	logFC	PValue
ENSG00000175899	A2M	8.445431	8.81E-49
ENSG00000198788	MUC2	8.021655	1.33E-67
ENSG00000122859	NEUROG3	7.682371	1.32E-39
ENSG00000130300	PLVAP	7.538088	9.61E-37
ENSG00000106331	PAX4	7.397652	2.58E-31
ENSG00000129167	TPH1	7.285501	2.92E-37
ENSG00000186474	KLK12	7.104196	1.23E-30
ENSG00000176435	CLEC14A	7.079766	9.19E-30
ENSG00000108551	RASD1	7.012534	8.09E-40
ENSG00000213088	ACKR1	6.893448	1.82E-27
ENSG00000179144	GIMAP7	6.744213	9.31E-32
ENSG00000240583	AQP1	6.716904	2.54E-25
ENSG00000163515	RETNLB	6.703482	2.78E-29
ENSG00000198719	DLL1	6.674991	3.95E-29
ENSG00000007908	SELE	6.663072	2.48E-28
ENSG00000179776	CDH5	6.55975	9.31E-27
ENSG00000174059	CD34	6.540212	8.75E-28
ENSG00000106991	ENG	6.496037	2.07E-22
ENSG00000204291	COL15A1	6.481498	4.75E-25
ENSG00000174175	SELP	6.282668	6.86E-28
ENSG00000131634	TMEM204	6.242608	1.42E-28
ENSG00000162618	ADGRL4	6.222831	9.57E-26
ENSG00000131477	RAMP2	6.176262	3.78E-28
ENSG00000133574	GIMAP4	6.134089	5.92E-28
ENSG00000134817	APLNR	6.100636	4.75E-22
ENSG00000188175	HEPACAM2	6.07849	9.44E-20
ENSG00000099998	GGT5	6.072979	1.19E-23
ENSG00000170989	S1PR1	5.905208	3.16E-23
ENSG00000127920	GNG11	5.793474	4.38E-21
ENSG00000100234	TIMP3	5.710226	1.63E-18
ENSG00000144485	HES6	5.697547	1.27E-19
ENSG00000149564	ESAM	5.678386	6.68E-20
ENSG00000060709	RIMBP2	5.548636	6.02E-17
ENSG00000184113	CLDN5	5.53069	3.57E-20
ENSG00000111341	MGP	5.484552	2.59E-23
ENSG00000122711	SPINK4	5.428672	2.36E-22
ENSG00000167748	KLK1	5.331629	6.95E-18
ENSG00000179914	ITLN1	5.271746	9.07E-19
ENSG00000105784	RUNDC3B	5.245376	1.91E-16
ENSG00000173198	CYSLTR1	5.183682	2.56E-19

gene_id	external_gene_name	logFC	PValue
ENSG00000106772	PRUNE2	5.175988	2.52E-15
ENSG00000163762	TM4SF18	5.167531	3.44E-18
ENSG00000163618	CADPS	5.165159	2.55E-14
ENSG00000213203	GIMAP1	5.163988	9.64E-18
ENSG00000110799	VWF	5.125987	8.69E-15
ENSG00000117707	PROX1	5.072906	1.23E-15
ENSG00000120708	TGFBI	4.977497	3.37E-14
ENSG00000124766	SOX4	3.743143	2.71E-25

A.12 Barrett's oesophagus cluster B1 - goblet cells

gene_id	external_gene_name	logFC	PValue
ENSG00000174562	KLK15	6.116711	1.71E-37
ENSG00000162998	FRZB	6.060249	9.82E-36
ENSG00000130035	GALNT8	6.046597	5.56E-45
ENSG00000116176	TPSG1	5.820541	1.84E-34
ENSG00000179914	ITLN1	5.539538	1.03E-40
ENSG00000158764	ITLN2	5.124621	5.48E-24
ENSG00000100302	RASD2	4.972009	2.72E-28
ENSG00000198788	MUC2	4.740967	1.84E-43
ENSG00000167748	KLK1	4.488499	3.48E-37
ENSG00000198719	DLL1	4.458206	1.05E-23
ENSG00000076864	RAP1GAP	4.375941	1.47E-36
ENSG00000142515	KLK3	4.354852	1.75E-40
ENSG00000174992	ZG16	4.340588	9.86E-19
ENSG00000122711	SPINK4	4.135448	1.77E-34
ENSG00000172482	AGXT	4.037768	9.73E-19
ENSG00000106178	CCL24	3.979341	1.59E-16
ENSG00000146205	ANO7	3.977226	2.06E-22
ENSG00000198488	B3GNT6	3.929162	2.00E-17
ENSG00000137634	NXPE4	3.905292	5.74E-16
ENSG00000184709	LRRC26	3.85318	6.24E-30
ENSG00000188833	ENTPD8	3.732462	7.96E-16
ENSG00000144485	HES6	3.675458	3.63E-22
ENSG00000197520	FAM177B	3.610659	1.67E-16
ENSG00000136854	STXBP1	3.433918	9.78E-17
ENSG00000107954	NEURL1	3.350787	2.99E-17
ENSG00000159921	GNE	3.328133	5.26E-18
ENSG00000070526	ST6GALNAC1	3.002982	1.23E-38
ENSG00000129538	RNASE1	2.743344	3.79E-20
ENSG00000115756	HPCAL1	2.698595	3.12E-20
ENSG00000148180	GSN	2.612085	2.92E-22
ENSG00000077232	DNAJC10	2.079836	1.80E-19
ENSG00000064787	BCAS1	2.068122	5.27E-16

A.13 Barrett's oesophagus cluster B2 - low proliferating Barrett's cells

gene_id	external_gene_name	logFC	PValue
ENSG00000133742	CA1	4.70099	6.87E-13
ENSG00000096088	PGC	4.241212	6.89E-15
ENSG00000165376	CLDN2	4.16321	9.90E-12
ENSG00000213996	TM6SF2	3.791935	1.71E-13
ENSG00000118515	SGK1	3.486023	2.07E-11
ENSG00000172367	PDZD3	3.246055	1.84E-09
ENSG00000182333	LIPF	3.125801	1.56E-09
ENSG00000167653	PSCA	2.995816	6.50E-08
ENSG00000198099	ADH4	2.947127	4.83E-08
ENSG00000104368	PLAT	2.679592	1.79E-08

gene_id	external_gene_name	logFC	PValue
ENSG00000163817	SLC6A20	2.521765	2.10E-07
ENSG00000088340	FER1L4	2.460193	1.73E-07
ENSG00000108602	ALDH3A1	2.331979	2.08E-09
ENSG00000166816	LDHD	2.283105	2.89E-07
ENSG00000150594	ADRA2A	2.208716	2.87E-07
ENSG00000142798	HSPG2	2.198878	1.77E-07
ENSG00000104267	CA2	2.186567	3.49E-10
ENSG00000205358	MT1H	2.151436	3.42E-08

A.14 Barrett's oesophagus cluster B3 - high proliferating Barrett's cells

gene_id	external_gene_name	logFC	PValue
ENSG00000137807	KIF23	6.454205	1.70E-54
ENSG00000169679	BUB1	6.286556	6.25E-55
ENSG00000066279	ASPM	5.955669	7.25E-40
ENSG00000178999	AURKB	5.885878	1.73E-40
ENSG00000157456	CCNB2	5.797809	4.90E-43
ENSG00000089685	BIRC5	5.779013	6.74E-50
ENSG00000117399	CDC20	5.757594	5.70E-37
ENSG00000166851	PLK1	5.726831	1.18E-37
ENSG00000168078	PBK	5.603938	2.27E-40
ENSG00000080986	NDC80	5.601276	9.96E-32
ENSG00000184357	HIST1H1B	5.583145	6.78E-27
ENSG00000088325	TPX2	5.575328	4.51E-35
ENSG00000137804	NUSAP1	5.570507	6.16E-42
ENSG00000126787	DLGAP5	5.553319	1.67E-35
ENSG00000090889	KIF4A	5.54907	5.20E-35
ENSG00000138180	CEP55	5.548501	8.98E-38
ENSG00000142945	KIF2C	5.489198	1.64E-27
ENSG00000156970	BUB1B	5.47614	1.80E-30
ENSG00000175063	UBE2C	5.433822	1.99E-30
ENSG00000121152	NCAPH	5.420157	1.73E-29
ENSG00000161888	SPC24	5.298594	2.33E-29
ENSG00000165304	MELK	5.279055	7.72E-30
ENSG00000145386	CCNA2	5.228285	1.34E-30
ENSG00000131747	TOP2A	5.156263	1.48E-27
ENSG00000135451	TROAP	5.15377	2.84E-24
ENSG00000171848	RRM2	5.129593	6.29E-30
ENSG00000100526	CDKN3	5.127608	1.21E-30
ENSG00000170312	CDK1	5.12067	9.68E-22
ENSG00000184661	CDCA2	5.11707	1.99E-24
ENSG00000134057	CCNB1	5.067426	3.35E-29
ENSG00000183856	IQGAP3	4.959021	1.38E-31
ENSG00000105011	ASF1B	4.946451	1.01E-24
ENSG00000011426	ANLN	4.846591	3.47E-21
ENSG00000123485	HJURP	4.830907	1.31E-20
ENSG00000111665	CDCA3	4.823706	9.63E-24
ENSG00000214357	NEURL1B	4.708263	5.25E-20
ENSG00000169607	CKAP2L	4.682474	9.86E-18
ENSG00000072571	HMMR	4.66544	7.60E-23
ENSG00000124575	HIST1H1D	4.499458	2.92E-20
ENSG00000198901	PRC1	4.474504	5.52E-23
ENSG00000148773	MKI67	4.413448	7.19E-22
ENSG000000051341	POLQ	4.381021	3.71E-17
ENSG00000138160	KIF11	4.374771	1.03E-19
ENSG00000164109	MAD2L1	4.334275	1.62E-20
ENSG00000144554	FANCD2	4.331343	1.56E-20
ENSG00000137812	CASC5	4.286263	2.95E-17
ENSG00000198826	ARHGAP11A	4.285264	2.98E-18
ENSG00000188486	H2AFX	4.192795	6.21E-25
ENSG00000076382	SPAG5	4.152973	3.48E-14

gene_id	external_gene_name	logFC	PValue
ENSG00000182481	KPNA2	4.133774	3.62E-22
ENSG00000087586	AURKA	4.048819	6.96E-17
ENSG00000164104	HMGB2	4.044088	3.84E-25
ENSG00000101057	MYBL2	4.042375	8.35E-16
ENSG00000138778	CENPE	3.967499	1.75E-14
ENSG00000203760	CENPW	3.874886	6.34E-16
ENSG00000164611	PTTG1	3.859149	1.06E-15
ENSG00000138182	KIF20B	3.765072	1.24E-14
ENSG00000167900	TK1	3.656272	5.08E-15
ENSG00000166803	KIAA0101	3.573956	8.43E-17
ENSG00000123975	CKS2	3.133694	4.20E-21
ENSG00000213347	MXD3	3.038391	5.48E-20
ENSG00000197061	HIST1H4C	2.07004	1.95E-16

A.15 Barrett's oesophagus cluster B4 - enteroendocrine cells

gene_id	external_gene_name	logFC	PValue
ENSG00000100604	CHGA	8.580235	4.47E-100
ENSG00000109472	CPE	8.061247	5.64E-92
ENSG00000116774	OLFML3	7.637461	6.94E-54
ENSG00000118271	TTR	7.636173	1.39E-73
ENSG00000133636	NTS	7.261163	9.90E-42
ENSG00000096395	MLN	7.238081	2.73E-49
ENSG00000089199	CHGB	7.148665	1.24E-66
ENSG00000104112	SCG3	7.058048	1.70E-55
ENSG00000079689	SCGN	7.009603	1.59E-54
ENSG00000102109	PCSK1N	6.887125	2.19E-47
ENSG00000163499	CRYBA2	6.781161	7.12E-43
ENSG00000131711	MAP1B	6.258657	9.59E-45
ENSG00000145362	ANK2	6.177585	4.91E-37
ENSG00000117152	RGS4	6.127877	3.48E-37
ENSG00000147488	ST18	6.116088	1.31E-38
ENSG00000163497	FEV	6.056995	8.49E-29
ENSG00000102466	FGF14	6.033789	1.09E-32
ENSG00000145321	GC	6.032149	2.09E-25
ENSG00000060709	RIMBP2	6.015067	9.39E-32
ENSG00000145198	VWA5B2	5.918535	1.29E-29
ENSG00000106331	PAX4	5.904379	1.08E-22
ENSG00000141837	CACNA1A	5.786284	2.04E-38
ENSG00000204889	KRT40	5.771165	6.80E-26
ENSG00000168280	KIF5C	5.677767	1.90E-34
ENSG00000180616	SSTR2	5.669346	6.95E-22
ENSG00000171951	SCG2	5.615967	5.37E-21
ENSG00000134198	TSPAN2	5.608405	7.37E-30
ENSG00000173826	KCNH6	5.543106	5.10E-24
ENSG00000109072	VTN	5.483937	1.95E-20
ENSG00000187045	TMPRSS6	5.464375	1.72E-24
ENSG00000185002	RFX6	5.424029	1.63E-22
ENSG00000180785	OR51E1	5.302617	3.52E-16
ENSG00000171303	KCNK3	5.241355	5.71E-24
ENSG00000126218	F10	5.228073	6.43E-22
ENSG00000136999	NOV	5.165861	5.73E-28
ENSG00000006071	ABCC8	5.138926	1.35E-18
ENSG00000147231	CXorf57	5.138375	4.31E-18
ENSG00000129167	TPH1	5.120725	1.63E-17
ENSG00000198829	SUCNR1	5.077061	3.32E-17
ENSG00000211445	GPX3	5.058402	1.05E-19
ENSG00000137033	IL33	5.039246	2.56E-19
ENSG00000086619	ERO1B	5.038371	1.09E-25
ENSG00000036565	SLC18A1	4.894239	5.44E-16
ENSG00000186567	CEACAM19	4.800527	4.90E-19
ENSG00000171476	HOPX	4.64967	2.36E-19

gene_id	external_gene_name	logFC	PValue
ENSG00000164638	SLC29A4	4.500216	6.41E-17
ENSG00000135919	SERPINE2	4.469322	2.16E-16
ENSG00000166959	MS4A8	4.208881	2.73E-24
ENSG00000145730	PAM	3.443027	1.26E-17
ENSG00000167552	TUBA1A	2.798841	3.79E-16

Appendix B

Differentially expressed genes in cell clusters from all tissues combined

Clustering analysis of single cells from an ensemble of all cells analysed by single cell RNA-seq from four patients with Barrett's oesophagus identified seven clusters and one control cluster (not shown). Highly significant upregulated genes in each cluster are shown in this appendix with a manual annotation of cell type. See Materials and Methods for explanation of how PValue thresholds were determined. 100 results with lowest PValues included where total results exceeds this.

B.1 Non epithelial

gene_id	external_gene_name	logFC	PValue
ENSG00000132465	JCHAIN	9.421933	3.74E-182
ENSG00000211899	IGHM	8.984969	3.46E-127
ENSG00000211890	IGHA2	8.358677	7.29E-116
ENSG00000170476	MZB1	7.901268	1.32E-92
ENSG00000211949	IGHV3-23	7.677077	1.88E-112
ENSG00000211895	IGHA1	7.488205	6.33E-91
ENSG00000026751	SLAMF7	6.901355	6.54E-92
ENSG00000239855	IGKV1-6	6.414827	1.24E-82
ENSG00000137265	IRF4	6.182461	1.04E-65
ENSG00000211592	IGKC	6.137428	2.36E-85
ENSG00000130300	PLVAP	5.943188	4.95E-53
ENSG00000048462	TNFRSF17	5.696844	6.91E-91
ENSG00000204291	COL15A1	5.538907	1.12E-63
ENSG00000118308	LRMP	5.510801	1.12E-42
ENSG00000173198	CYSLTR1	5.297046	1.58E-65
ENSG00000230006	ANKRD36BP2	5.293929	2.26E-36
ENSG00000143297	FCRL5	5.285329	2.52E-41
ENSG00000007908	SELE	5.255194	1.35E-48
ENSG00000174175	SELP	5.025437	4.23E-56
ENSG00000162618	ADGRL4	4.931259	5.44E-50
ENSG00000099958	DERL3	4.802054	5.94E-26
ENSG00000179144	GIMAP7	4.766687	1.63E-29
ENSG00000167077	MEI1	4.718847	5.64E-26
ENSG00000176435	CLEC14A	4.635725	7.47E-25

gene_id	external_gene_name	logFC	PValue
ENSG00000213088	ACKR1	4.565352	1.38E-24
ENSG00000163762	TM4SF18	4.49778	3.00E-43
ENSG00000102096	PIM2	4.488458	1.96E-27
ENSG00000122862	SRGN	4.375402	8.58E-25
ENSG00000152689	RASGRP3	4.158154	1.01E-20
ENSG00000131477	RAMP2	4.042631	2.46E-23
ENSG00000133574	GIMAP4	3.911595	1.75E-19
ENSG00000105784	RUNDC3B	3.869083	1.19E-22
ENSG00000127920	GNG11	3.819933	1.60E-19
ENSG00000197943	PLCG2	3.805331	1.08E-15
ENSG00000187498	COL4A1	3.65177	4.30E-19
ENSG00000198624	CCDC69	3.650818	1.32E-16
ENSG00000213203	GIMAP1	3.648437	8.02E-20

B.2 Barrett's type

gene_id	external_gene_name	logFC	PValue
ENSG00000133742	CA1	5.041841	5.43E-47
ENSG00000243709	LEFTY1	4.913848	2.44E-89
ENSG00000132911	NMUR2	4.691247	3.04E-56
ENSG00000163739	CXCL1	3.570928	1.80E-30
ENSG00000086548	CEACAM6	3.445152	2.54E-37
ENSG00000183091	NEB	3.351777	3.66E-32
ENSG00000205277	MUC12	3.337496	3.92E-34
ENSG00000109610	SOD3	3.240966	2.00E-30
ENSG00000094755	GABRP	3.209838	1.66E-29
ENSG00000135374	ELF5	3.083779	7.12E-24
ENSG0000013588	GPRC5A	3.028671	1.65E-46
ENSG00000105388	CEACAM5	2.987728	1.44E-29
ENSG00000169876	MUC17	2.92779	6.02E-31
ENSG00000204019	CT83	2.852599	3.70E-27
ENSG00000198734	F5	2.834421	5.27E-26
ENSG00000196352	CD55	2.704958	1.73E-32
ENSG00000204616	TRIM31	2.646599	1.16E-22
ENSG00000108511	HOXB6	2.498172	1.89E-23
ENSG00000135480	KRT7	2.465271	3.70E-24
ENSG00000181649	PHLDA2	2.322317	7.43E-32
ENSG00000141655	TNFRSF11A	2.11044	3.88E-27
ENSG00000185275	CD24P4	2.077556	2.51E-32

B.3 Enterocytes

gene_id	external_gene_name	logFC	PValue
ENSG00000114113	RBP2	7.451906	6.70E-207
ENSG00000160868	CYP3A4	7.349499	1.33E-160
ENSG00000131142	CCL25	7.232882	2.14E-180
ENSG00000115850	LCT	7.16445	1.10E-134
ENSG00000084674	APOB	7.149958	1.14E-111
ENSG00000172689	MS4A10	6.978757	4.25E-134
ENSG00000091138	SLC26A3	6.852212	1.76E-125
ENSG00000138823	MTTP	6.612707	1.37E-152
ENSG00000110244	APOA4	6.562009	4.39E-79
ENSG00000163581	SLC2A2	6.358898	6.33E-97
ENSG00000105398	SULT2A1	6.302881	1.12E-107
ENSG00000110245	APOC3	6.064688	1.19E-68
ENSG00000197888	UGT2B17	5.992107	5.37E-129
ENSG00000196549	MME	5.905477	1.03E-110
ENSG00000007216	SLC13A2	5.728182	4.28E-84
ENSG00000163295	ALPI	5.530091	2.83E-56
ENSG00000135218	CD36	5.495565	1.02E-80

gene_id	external_gene_name	logFC	PValue
ENSG00000060566	CREB3L3	5.485824	3.65E-73
ENSG00000136872	ALDOB	5.462767	5.33E-79
ENSG00000172016	REG3A	5.315973	4.96E-64
ENSG00000088386	SLC15A1	5.312731	5.88E-64
ENSG00000021488	SLC7A9	5.307051	4.02E-74
ENSG00000086205	FOLH1	5.265366	4.56E-78
ENSG00000138792	ENPEP	5.207295	1.78E-75
ENSG00000135220	UGT2A3	5.201604	4.84E-84
ENSG00000124253	PCK1	5.191855	1.10E-61
ENSG00000122121	XPNPEP2	5.09951	3.21E-61
ENSG00000175329	ISX	5.037526	7.17E-64
ENSG00000244067	GSTA2	4.997029	1.02E-97
ENSG00000131482	G6PC	4.963312	6.96E-58
ENSG00000145626	UGT3A1	4.950971	5.86E-78
ENSG00000080293	SCTR	4.887538	9.01E-66
ENSG00000154646	TMPRSS15	4.865505	1.73E-44
ENSG00000100170	SLC5A1	4.853572	5.56E-80
ENSG00000117834	SLC5A9	4.851446	1.07E-54
ENSG00000138308	PLA2G12B	4.823566	5.87E-91
ENSG00000214782	MS4A18	4.79083	4.67E-55
ENSG00000141434	MEP1B	4.755083	2.19E-41
ENSG00000143595	AQP10	4.749562	4.10E-53
ENSG00000168748	CA7	4.693382	2.80E-37
ENSG00000130234	ACE2	4.658397	8.68E-52
ENSG00000169906	S100G	4.626627	1.02E-72
ENSG00000204740	MALRD1	4.616837	7.91E-52
ENSG00000163586	FABP1	4.574445	2.29E-58
ENSG00000142959	BEST4	4.570469	2.05E-40
ENSG00000156222	SLC28A1	4.570396	2.36E-43
ENSG00000148942	SLC5A12	4.563364	1.67E-56
ENSG00000196660	SLC30A10	4.544134	1.58E-55
ENSG00000186198	SLC51B	4.534903	1.01E-46
ENSG00000142583	SLC2A5	4.419469	1.38E-53
ENSG00000204978	ERICH4	4.418129	2.02E-49
ENSG00000146285	SCML4	4.417816	7.25E-58
ENSG00000138030	KHK	4.400471	8.72E-63
ENSG00000090402	SI	4.399162	1.31E-59
ENSG00000115386	REG1A	4.367393	4.31E-69
ENSG00000112299	VNN1	4.352571	3.83E-42
ENSG00000162460	TMEM82	4.347936	1.16E-44
ENSG00000170482	SLC23A1	4.311957	5.18E-49
ENSG00000118137	APOA1	4.299969	2.95E-36
ENSG00000164270	HTR4	4.274324	7.10E-46
ENSG00000118094	TREH	4.216181	1.50E-36
ENSG00000174885	NLRP6	4.191195	4.26E-37
ENSG00000112494	UNC93A	4.189221	1.19E-43
ENSG00000023839	ABCC2	4.185655	2.02E-34
ENSG00000186115	CYP4F2	4.179359	7.29E-34
ENSG00000196839	ADA	4.117844	7.67E-46
ENSG00000158516	CPA2	4.107826	1.10E-36
ENSG00000103485	QPRT	4.100034	1.69E-47
ENSG00000152785	BMP3	4.081186	5.06E-35
ENSG00000114771	AADAC	4.051868	1.26E-56
ENSG00000176919	C8G	4.031197	4.83E-31
ENSG00000187288	CIDEC	4.010202	5.55E-30
ENSG00000226306	NPY6R	3.998366	3.59E-33
ENSG00000163075	CFAP221	3.961188	1.48E-39
ENSG00000104808	DHDH	3.948601	1.80E-29
ENSG00000165828	PRAP1	3.916167	4.35E-62
ENSG00000138075	ABCG5	3.857158	8.92E-31
ENSG00000106404	CLDN15	3.854526	4.00E-56
ENSG00000234155	RP11-30P6.6	3.85351	2.60E-33
ENSG00000132517	SLC52A1	3.840354	6.80E-28
ENSG00000182938	OTOP3	3.825931	7.67E-28
ENSG00000182156	ENPP7	3.814502	5.70E-26
ENSG00000171234	UGT2B7	3.812045	6.42E-30
ENSG00000241224	FLJ22763	3.75206	1.83E-35

gene_id	external_gene_name	logFC	PValue
ENSG00000160870	CYP3A7	3.738423	1.45E-47
ENSG00000197181	PIWIL2	3.735174	1.50E-29
ENSG00000173237	C11orf86	3.713451	1.05E-23
ENSG00000187908	DMBT1	3.707047	8.55E-41
ENSG00000115255	REEP6	3.690568	6.29E-57
ENSG00000143921	ABCG8	3.664015	1.02E-27
ENSG00000166869	CHP2	3.661498	7.08E-45
ENSG00000232070	TMEM253	3.649892	7.60E-35
ENSG00000147872	PLIN2	3.640802	9.92E-46
ENSG00000015413	DPEP1	3.638251	3.88E-27
ENSG00000243955	GSTA1	3.595042	1.99E-49
ENSG00000109743	BST1	3.581256	2.04E-30
ENSG00000013583	HEBP1	3.580348	1.30E-50
ENSG00000176273	SLC35G1	3.561801	2.37E-33
ENSG00000071967	CYBRD1	3.557708	9.03E-41
ENSG00000100889	PCK2	3.541152	9.01E-46

B.4 Mucus neck

gene_id	external_gene_name	logFC	PValue
ENSG00000183607	GKN2	8.724955	7.76E-292
ENSG00000169605	GKN1	8.44435	4.54E-222
ENSG00000096088	PGC	7.03602	6.69E-159
ENSG00000229859	PGA3	6.865219	5.49E-145
ENSG00000229183	PGA4	6.699652	1.63E-168
ENSG00000182333	LIPF	5.828241	3.42E-91
ENSG00000163631	ALB	5.700624	1.22E-99
ENSG00000110195	FOLR1	5.45926	6.84E-66
ENSG00000168631	DPCR1	5.453797	3.37E-69
ENSG00000105641	SLC5A5	5.443257	1.55E-52
ENSG00000125999	BPIFB1	5.418066	1.62E-65
ENSG00000167653	PSCA	5.083306	1.14E-68
ENSG00000189377	CXCL17	4.802821	5.78E-76
ENSG00000174502	SLC26A9	4.473513	1.82E-41
ENSG00000108950	FAM20A	4.469103	9.75E-47
ENSG00000159197	KCNE2	4.190088	3.98E-40
ENSG00000138115	CYP2C8	4.147881	2.71E-35
ENSG00000169583	CLIC3	4.075786	1.19E-36
ENSG00000180251	SLC9A4	4.060023	2.37E-47
ENSG00000161055	SCGB3A1	4.033123	3.19E-32
ENSG00000140274	DUOXA2	3.958887	4.21E-27
ENSG00000196569	LAMA2	3.942977	1.16E-33
ENSG00000023445	BIRC3	3.90518	4.27E-35
ENSG00000150556	LYPD6B	3.857596	3.52E-45
ENSG00000100665	SERPINA4	3.844528	3.43E-31
ENSG00000135063	FAM189A2	3.83167	1.23E-31
ENSG00000099769	IGFALS	3.748755	7.89E-25
ENSG00000178597	PSAPL1	3.74061	7.24E-29
ENSG00000135312	HTR1B	3.698458	1.42E-29
ENSG00000232931	LINC00342	3.644277	3.89E-31
ENSG00000042980	ADAM28	3.632364	1.85E-57
ENSG00000174501	ANKRD36C	3.627034	2.11E-62
ENSG00000140279	DUOX2	3.588457	9.37E-28
ENSG00000107159	CA9	3.580643	9.68E-56
ENSG00000228278	ORM2	3.516321	7.99E-37
ENSG00000153446	C16orf89	3.512599	2.36E-23
ENSG00000088340	FER1L4	3.447901	8.62E-32
ENSG00000171243	SOSTDC1	3.447486	5.50E-31
ENSG00000127249	ATP13A4	3.443298	1.29E-26
ENSG00000161040	FBXL13	3.439338	5.27E-25
ENSG00000137033	IL33	3.436278	1.52E-30
ENSG00000160181	TFF2	3.388803	1.90E-35
ENSG00000227640	SOX21-AS1	3.370204	2.07E-31
ENSG00000130957	FBP2	3.315468	4.02E-25

gene_id	external_gene_name	logFC	PValue
ENSG00000171916	LGALS9C	3.289292	1.01E-34
ENSG00000188818	ZDHHC11	3.282736	1.78E-34
ENSG00000019102	VSIG2	3.24996	2.43E-52
ENSG00000066405	CLDN18	3.246942	8.44E-49
ENSG00000172461	FUT9	3.204493	4.94E-27
ENSG00000100918	REC8	3.143753	9.77E-27
ENSG00000196542	SPTSSB	3.077128	1.52E-23
ENSG00000214456	PLIN5	3.057562	4.62E-19
ENSG00000101842	VSIG1	2.949844	5.13E-29
ENSG00000174514	MFSD4	2.935216	3.34E-22
ENSG00000148346	LCN2	2.85469	2.39E-25
ENSG00000104267	CA2	2.823937	3.51E-46
ENSG00000177133	LINC00982	2.813375	8.16E-20
ENSG00000077044	DGKD	2.788255	4.80E-36
ENSG00000154065	ANKRD29	2.751123	1.95E-18
ENSG00000088992	TESC	2.732515	7.84E-29
ENSG00000196620	UGT2B15	2.712332	7.36E-20
ENSG00000179583	CIITA	2.68324	9.73E-19
ENSG00000108602	ALDH3A1	2.679518	6.93E-30
ENSG00000127528	KLF2	2.58573	5.90E-19
ENSG00000185499	MUC1	2.549278	4.31E-34
ENSG00000139044	B4GALNT3	2.527582	2.35E-18
ENSG00000015532	XYLT2	2.521867	3.02E-25
ENSG00000164379	FOXQ1	2.512827	6.68E-24
ENSG00000139144	PIK3C2G	2.501222	1.64E-21
ENSG00000215182	MUC5AC	2.473851	1.24E-28
ENSG00000090020	SLC9A1	2.458848	2.24E-29
ENSG00000185551	NR2F2	2.456636	8.72E-25
ENSG00000149577	SIDT2	2.454959	3.40E-27
ENSG00000006025	OSBPL7	2.384351	6.68E-21
ENSG00000196188	CTSE	2.359344	8.41E-23
ENSG00000164266	SPINK1	2.326811	1.74E-31
ENSG00000198203	SULT1C2	2.265204	7.76E-22
ENSG00000156966	B3GNT7	2.203842	4.08E-18
ENSG00000116260	QSOX1	2.199002	9.59E-30
ENSG00000203697	CAPN8	2.145405	1.81E-19
ENSG00000139629	GALNT6	2.004787	7.40E-19

B.5 Goblet

gene_id	external_gene_name	logFC	PValue
ENSG00000179914	ITLN1	7.718778	1.90E-208
ENSG00000016490	CLCA1	6.902937	3.31E-100
ENSG00000158764	ITLN2	6.842165	1.35E-124
ENSG00000163515	RETNLB	6.487551	4.98E-92
ENSG00000130035	GALNT8	6.35982	1.92E-139
ENSG00000164822	DEFA6	6.263345	2.29E-64
ENSG00000198788	MUC2	6.175601	1.05E-134
ENSG00000122711	SPINK4	5.899019	5.13E-111
ENSG00000167748	KLK1	5.632451	3.95E-99
ENSG00000106178	CCL24	5.556855	7.59E-82
ENSG00000164816	DEFA5	5.469085	2.54E-46
ENSG00000186474	KLK12	5.397581	1.36E-56
ENSG00000174562	KLK15	5.358123	3.56E-66
ENSG00000204983	PRSS1	5.27908	5.62E-43
ENSG00000172238	ATOH1	5.176765	1.56E-58
ENSG00000184709	LRRC26	5.113182	2.03E-85
ENSG00000198719	DLL1	4.970811	7.14E-55
ENSG00000198691	ABCA4	4.943643	5.21E-43
ENSG00000144485	HES6	4.939047	3.46E-68
ENSG00000174992	ZG16	4.849594	2.16E-47
ENSG00000116176	TPSG1	4.716237	9.84E-41
ENSG00000100302	RASD2	4.681937	7.68E-50
ENSG00000172482	AGXT	4.550742	2.15E-41

gene_id	external_gene_name	logFC	PValue
ENSG00000162998	FRZB	4.093138	5.78E-24
ENSG00000076864	RAP1GAP	3.779348	1.65E-39
ENSG00000142515	KLK3	3.737276	4.13E-88
ENSG00000167964	RAB26	3.737008	1.99E-29
ENSG00000204361	NXPE2	3.689732	9.62E-24
ENSG00000070526	ST6GALNAC1	3.674713	1.36E-71
ENSG00000122859	NEUROG3	3.674635	1.29E-16
ENSG00000197273	GUCA2A	3.667629	8.42E-21
ENSG00000159921	GNE	3.66313	1.13E-33
ENSG00000143001	TMEM61	3.64325	2.48E-27
ENSG00000198488	B3GNT6	3.609209	1.81E-23
ENSG00000107954	NEURL1	3.602475	1.45E-32
ENSG00000101443	WFDC2	3.590657	6.04E-36
ENSG00000128917	DLL4	3.554092	1.36E-21
ENSG00000007402	CACNA2D2	3.425551	7.19E-22
ENSG00000146205	ANO7	3.417925	1.05E-24
ENSG00000158292	GPR153	3.366403	5.51E-21
ENSG00000104728	ARHGEF10	3.339264	2.30E-26
ENSG00000120708	TGFB1	3.33268	2.56E-16
ENSG00000188833	ENTPD8	3.242893	1.48E-19
ENSG00000155254	MARVELD1	3.221283	1.24E-17
ENSG00000144596	GRIP2	3.201193	4.32E-18
ENSG00000108551	RASD1	3.186564	1.17E-18
ENSG00000130038	CRACR2A	3.121986	3.73E-19
ENSG00000160180	TFF3	3.0931	7.15E-40
ENSG00000139998	RAB15	3.040575	4.63E-22
ENSG00000084636	COL16A1	3.022294	2.69E-19
ENSG00000187068	C3orf70	2.932028	2.51E-16
ENSG00000104936	DMPK	2.85729	8.99E-19
ENSG00000124766	SOX4	2.807667	6.34E-42
ENSG00000117525	F3	2.794695	2.66E-23
ENSG00000124664	SPDEF	2.78935	2.66E-22
ENSG00000115756	HPCAL1	2.778971	2.79E-34
ENSG00000169684	CHRNA5	2.750594	3.18E-16
ENSG00000136378	ADAMTS7	2.723679	1.09E-15
ENSG00000138821	SLC39A8	2.672527	9.87E-17
ENSG00000175344	CHRNA7	2.608778	1.93E-15
ENSG00000197249	SERPINA1	2.59567	1.03E-19
ENSG00000143578	CREB3L4	2.51651	4.30E-17
ENSG00000136542	GALNT5	2.44382	2.07E-16
ENSG00000010219	DYRK4	2.385552	5.55E-18
ENSG00000078699	CBFA2T2	2.367787	1.34E-16
ENSG00000077232	DNAJC10	2.319473	4.70E-31
ENSG00000157613	CREB3L1	2.263263	1.81E-22
ENSG00000131149	GSE1	2.227406	5.46E-20
ENSG00000106537	TSPAN13	2.034008	9.47E-26

B.6 Squamous

gene_id	external_gene_name	logFC	PValue
ENSG00000140519	RHCG	9.904276	0
ENSG00000186847	KRT14	9.356956	0
ENSG00000205420	KRT6A	9.202039	0
ENSG00000167656	LY6D	9.125959	0
ENSG00000183347	GBP6	9.055132	0
ENSG00000178363	CALML3	8.970222	0
ENSG00000186081	KRT5	8.918592	3.06320700421573e-322
ENSG000000057149	SERPINB3	8.627009	0
ENSG00000124466	LYPD3	8.523737	0
ENSG00000230937	MIR205HG	8.377345	0
ENSG00000143546	S100A8	8.321828	0
ENSG00000134762	DSC3	8.302677	0
ENSG00000137975	CLCA2	8.183798	0
ENSG00000185479	KRT6B	8.159502	0

gene_id	external_gene_name	logFC	PValue
ENSG00000186832	KRT16	8.075337	0
ENSG00000163220	S100A9	8.053018	1.48E-239
ENSG00000170477	KRT4	7.981863	5.09E-180
ENSG00000171401	KRT13	7.960159	0
ENSG00000206073	SERPINB4	7.929602	1.26E-228
ENSG00000196754	S100A2	7.887767	0
ENSG00000170465	KRT6C	7.876662	0
ENSG00000143320	CRABP2	7.574124	1.74E-241
ENSG00000081277	PKP1	7.413914	0
ENSG00000171346	KRT15	7.25281	1.49E-191
ENSG00000197641	SERPINB13	7.202083	7.33E-302
ENSG00000205076	LGALS7	6.952302	0
ENSG00000137699	TRIM29	6.937645	9.00E-184
ENSG00000138755	CXCL9	6.828742	7.31E-126
ENSG00000154227	CERS3	6.738887	1.01E-289
ENSG00000141579	ZNF750	6.659798	9.06E-184
ENSG00000073282	TP63	6.539231	1.59E-164
ENSG00000178934	LGALS7B	6.524788	2.21E-173
ENSG00000196344	ADH7	6.404835	9.36E-114
ENSG00000163331	DAPL1	6.394497	2.82480451999666e-317
ENSG00000087494	PTHLH	6.391845	3.86E-117
ENSG00000134757	DSG3	6.365069	9.34E-148
ENSG00000173432	SAA1	6.355647	3.86E-114
ENSG00000174564	IL20RB	6.338349	2.16E-130
ENSG00000136235	GPNMB	6.330715	7.23E-119
ENSG00000110852	CLEC2B	6.306776	5.00488499237183e-321
ENSG00000129194	SOX15	6.240759	6.42E-212
ENSG00000153802	TMPRSS11D	6.132514	1.24E-163
ENSG00000138271	GPR87	6.128058	3.38E-171
ENSG00000187054	TMPRSS11A	6.083901	2.26E-140
ENSG00000147697	GSDMC	5.975054	5.73E-145
ENSG00000182326	C1S	5.934732	3.15E-81
ENSG00000158055	GRHL3	5.898779	1.80E-117
ENSG00000000971	CFH	5.873448	1.70E-107
ENSG00000179593	ALOX15B	5.777345	1.01E-89
ENSG00000131203	IDO1	5.774432	7.85E-70
ENSG00000121552	CSTA	5.769526	7.06E-146
ENSG00000129451	KLK10	5.710084	2.96E-83
ENSG00000169594	BNC1	5.673218	7.61E-113
ENSG00000125998	FAM83C	5.646908	1.02E-92
ENSG00000165272	AQP3	5.636515	1.52E-105
ENSG00000165794	SLC39A2	5.615932	9.69E-84
ENSG00000120471	TP53AIP1	5.613895	3.83E-117
ENSG00000188373	C10orf99	5.593266	1.29E-93
ENSG00000019549	SNAI2	5.584987	4.68E-101
ENSG00000159403	C1R	5.565986	3.44E-74
ENSG000000086570	FAT2	5.551557	1.90E-127
ENSG00000166535	A2ML1	5.546328	1.37E-81
ENSG00000186867	QRFPR	5.494763	2.52E-109
ENSG00000156535	CD109	5.493856	2.84E-141
ENSG00000134760	DSG1	5.481318	2.89E-82
ENSG00000169245	CXCL10	5.381474	1.42E-47
ENSG00000135046	ANXA1	5.316073	1.38E-98
ENSG00000178726	THBD	5.281648	2.19E-75
ENSG00000154451	GBP5	5.274563	2.46E-77
ENSG00000113430	IRX4	5.252191	2.83E-86
ENSG00000132746	ALDH3B2	5.219157	2.98E-63
ENSG00000167768	KRT1	5.215703	9.87E-75
ENSG00000143341	HMCN1	5.198693	1.87E-71
ENSG00000128510	CPA4	5.177767	8.86E-59
ENSG00000105131	EPHX3	5.139701	4.53E-57
ENSG00000167779	IGFBP6	5.115294	1.76E-69
ENSG00000109101	FOXP1	5.079901	4.23E-78
ENSG00000171517	LPAR3	5.035093	5.81E-93
ENSG00000088726	TMEM40	5.023491	4.94E-86
ENSG00000159231	CBR3	5.003785	1.05E-115
ENSG00000122861	PLAU	4.937906	1.14E-50

gene_id	external_gene_name	logFC	PValue
ENSG00000169469	SPRR1B	4.797652	2.48E-42
ENSG00000079931	MOXD1	4.780201	5.41E-48
ENSG00000187955	COL14A1	4.763014	8.73E-43
ENSG00000185614	FAM212A	4.739814	6.95E-66
ENSG00000175183	CSRP2	4.731907	4.32E-67
ENSG00000176014	TUBB6	4.681091	8.93E-61
ENSG00000128422	KRT17	4.68021	1.04E-135
ENSG00000025708	TYMP	4.669342	8.54E-66
ENSG00000077942	FBLN1	4.668536	9.85E-54
ENSG00000171056	SOX7	4.642381	3.41E-71
ENSG00000092295	TGM1	4.636942	1.00E-43
ENSG00000129151	BBOX1	4.591167	1.41E-56
ENSG00000095303	PTGS1	4.591135	2.65E-43
ENSG00000127863	TNFRSF19	4.589424	9.47E-49
ENSG00000180438	TPRXL	4.587297	7.28E-53
ENSG00000140459	CYP11A1	4.586511	4.21E-56
ENSG00000186891	TNFRSF18	4.567948	4.11E-42
ENSG00000137440	FGFBP1	4.561525	6.32E-48
ENSG00000188883	KLRG2	4.552344	5.48E-53

B.7 Enteroendocrine

gene_id	external_gene_name	logFC	PValue
ENSG00000118271	TTR	8.367249	1.76E-179
ENSG00000109072	VTN	7.607731	7.19E-111
ENSG00000096395	MLN	7.509709	3.69E-96
ENSG00000100604	CHGA	7.424808	1.81E-114
ENSG00000079689	SCGN	7.185621	8.53E-97
ENSG00000116774	OLFML3	7.020595	1.92E-80
ENSG00000163497	FEV	7.010955	2.06E-83
ENSG00000163499	CRYBA2	6.868707	5.83E-102
ENSG00000157017	GHRL	6.803236	5.32E-98
ENSG00000162992	NEUROD1	6.723244	6.91E-63
ENSG00000171951	SCG2	6.65464	8.38E-69
ENSG00000157005	SST	6.553441	2.61E-60
ENSG00000136698	CFC1	6.506376	2.02E-65
ENSG00000145321	GC	6.404441	2.33E-59
ENSG00000185002	RFX6	6.234455	1.92E-60
ENSG00000089199	CHGB	6.124399	1.79E-58
ENSG00000177669	MBOAT4	6.10364	3.84E-49
ENSG00000139219	COL2A1	5.861738	9.84E-64
ENSG00000036565	SLC18A1	5.804888	1.31E-45
ENSG00000133636	NTS	5.802587	4.45E-42
ENSG00000145198	VWA5B2	5.740606	2.95E-53
ENSG00000146013	GFRA3	5.71838	3.05E-43
ENSG00000180616	SSTR2	5.640363	1.03E-44
ENSG00000060709	RIMBP2	5.625286	1.30E-53
ENSG00000129167	TPH1	5.547619	3.11E-36
ENSG00000116183	PAPPA2	5.408763	4.92E-42
ENSG00000165507	C10orf10	5.3339	2.93E-40
ENSG00000171303	KCNK3	5.271627	7.28E-48
ENSG00000180785	OR51E1	5.185126	1.82E-28
ENSG00000171759	PAH	5.128956	5.69E-32
ENSG00000102466	FGF14	5.069056	1.50E-35
ENSG00000141837	CACNA1A	5.04839	1.72E-49
ENSG00000140287	HDC	5.018621	3.86E-21
ENSG00000006071	ABCC8	4.955759	2.64E-31
ENSG00000153002	CPB1	4.946263	5.81E-25
ENSG00000147488	ST18	4.938659	7.53E-42
ENSG00000173404	INSM1	4.86887	7.09E-32
ENSG00000173826	KCNH6	4.8322	2.37E-35
ENSG00000134198	TSPAN2	4.800292	1.71E-39
ENSG00000124780	KCNK17	4.798376	1.90E-36
ENSG00000161149	TUBA3FP	4.776498	8.04E-37

gene_id	external_gene_name	logFC	PValue
ENSG00000145087	STXBP5L	4.774441	1.22E-37
ENSG00000178695	KCTD12	4.669849	2.32E-41
ENSG00000164638	SLC29A4	4.641796	3.66E-36
ENSG00000126218	F10	4.640256	1.14E-27
ENSG00000151726	ACSL1	4.56189	5.79E-37
ENSG00000104112	SCG3	4.557601	5.63E-29
ENSG00000152580	IGSF10	4.521253	1.28E-26
ENSG00000086619	ERO1B	4.510906	2.36E-35
ENSG00000197594	ENPP1	4.504472	2.14E-27
ENSG00000004848	ARX	4.431105	4.14E-31
ENSG00000204889	KRT40	4.423877	2.60E-23
ENSG00000146674	IGFBP3	4.42334	3.09E-24
ENSG00000175426	PCSK1	4.368906	1.74E-22
ENSG00000102109	PCSK1N	4.330306	8.99E-22
ENSG00000152932	RAB3C	4.32934	8.78E-30
ENSG00000106331	PAX4	4.300139	1.07E-17
ENSG00000125851	PCSK2	4.254889	1.66E-19
ENSG00000120457	KCNJ5	4.236822	7.17E-24
ENSG00000139055	ERP27	4.234203	5.75E-20
ENSG00000211445	GPX3	4.221676	2.76E-22
ENSG00000198829	SUCNR1	4.221295	5.61E-24
ENSG00000162595	DIRAS3	4.210076	7.79E-21
ENSG00000187045	TMPRSS6	4.193042	9.82E-25
ENSG00000166922	SCG5	4.134306	1.47E-22
ENSG0000016082	ISL1	4.038579	1.27E-20
ENSG00000125780	TGM3	4.023416	1.61E-18
ENSG00000147231	CXorf57	4.008574	5.58E-21
ENSG00000184530	C6orf58	3.981329	3.50E-15
ENSG00000128536	CDHR3	3.974514	1.41E-22
ENSG00000166741	NNMT	3.972329	2.71E-18
ENSG00000139767	SRRM4	3.966946	2.46E-21
ENSG00000054356	PTPRN	3.953579	7.52E-18
ENSG00000133958	UNC79	3.887042	1.54E-19
ENSG00000095739	BAMBI	3.846703	2.99E-17
ENSG00000178171	AMER3	3.840884	2.24E-19
ENSG00000152804	HHEX	3.822758	7.83E-17
ENSG00000166959	MS4A8	3.815564	2.51E-25
ENSG00000115828	QPCT	3.79268	2.54E-15
ENSG00000179921	GPBAR1	3.765	5.26E-15
ENSG00000168243	GNG4	3.753364	1.31E-16
ENSG00000135447	PPP1R1A	3.749728	4.60E-20
ENSG00000176956	LY6H	3.736433	5.43E-16
ENSG00000153956	CACNA2D1	3.725323	1.55E-19
ENSG00000133519	ZDHHC8P1	3.708363	2.00E-18
ENSG00000172794	RAB37	3.706794	5.91E-17
ENSG00000100228	RAB36	3.695343	1.92E-14
ENSG00000110318	CEP126	3.690931	8.82E-22
ENSG00000117707	PROX1	3.689369	9.13E-19
ENSG00000204347	BTBD17	3.666124	6.19E-15
ENSG00000186567	CEACAM19	3.655208	1.54E-17
ENSG00000169507	SLC38A11	3.584756	1.49E-17
ENSG00000162552	WNT4	3.571072	1.87E-14
ENSG00000140323	DISP2	3.551259	1.12E-15
ENSG00000171476	HOPX	3.543589	2.85E-18
ENSG00000057593	F7	3.540642	1.12E-14
ENSG00000196557	CACNA1H	3.497437	9.37E-16
ENSG00000145730	PAM	3.436475	5.18E-27
ENSG00000077279	DCX	3.426925	3.72E-17
ENSG00000157542	KCNJ6	3.423162	1.68E-15

Appendix C

Classifiers used in `snv.predict`

`snv.predict` can be installed from github.com/mrrichowen/snv.predict. Full code available in Appendix D.

feature	derivation
alt_bqual	nuseq
alt_counts	nuseq
alt_distance	nuseq
alt_forward_counts	nuseq
alt_indel	nuseq
alt_mqual	nuseq
alt_reverse_counts	nuseq
ambiguous_counts	nuseq
deletion_counts	nuseq
entropy	nuseq
insertion_counts	nuseq
num_reads_with_indel	nuseq
ref_bqual	nuseq
ref_counts	nuseq
ref_distance	nuseq
ref_forward_counts	nuseq
ref_indel	nuseq
ref_mqual	nuseq
ref_reverse_counts	nuseq
total_bqual	nuseq
total_counts	nuseq
total_distance	nuseq
total_forward_counts	nuseq
total_indel	nuseq
total_mqual	nuseq
total_reverse_counts	nuseq
ref_genome_entropy	nuseq
gc_content	nuseq
forward_homopolymer_length	nuseq
reverse_homopolymer_length	nuseq
alt_counts_norm	snv.bench
alt_forward_counts_norm	snv.bench
alt_reverse_counts_norm	snv.bench
ref_counts_norm	snv.bench
ref_forward_counts_norm	snv.bench
ref_reverse_counts_norm	snv.bench
alt_bqual_ratio	snv.bench
alt_mqual_ratio	snv.bench
alt_distance_ratio	snv.bench
ref_bqual_ratio	snv.bench
ref_mqual_ratio	snv.bench
ref_distance_ratio	snv.bench
total_bqual_ratio	snv.bench
total_mqual_ratio	snv.bench

Table C.1: Table showing the classifiers extracted from the SNV caller using nuseq and snv.bench, respectively.

Appendix D

Functions for the prediction of nucleotide variation in single cell RNA-seq

D.1 snv.bench

Function for organising data into single matrix, computing new classifiers from the nuseq output and removing highly correlating classifiers. Results are used with downstream functions snv.performance and snv.cross.predict.

```
#' snv.bench
#'
#'This is a convenience function that takes the output files from nuseq-sc and
  combines them with a list of data frames with annotated snvs for input into other
  snv.predict functions. It will output a single snv.summary file concatenating
  all results and a vector of the low correlating classifiers recommended for
  subsequent analysis.
#'@param nuseq_classifiers List of named data frames as output from nuseq-sc.
#'@param truths List of data.frames with snv information. Easiest to use the
  convenience function 'convert_to_pos' and add a fifth column ('labels') with snvs
  annotated as true_positive (1), false_positive (0) and unknown (NA).
#'@param training_set If TRUE, only snvs with ground truth data will be output. This
  is useful if this data is to be used for training a model to be applied to a
  different dataset with snv.cross.predict.
#'@keywords
#'@export
#'@examples
#'snv.bench(nuseq_classifiers, truths, training_set=FALSE)

snv.bench <- function(nuseq_classifiers, truths, training_set=FALSE){

  require(caret)

  data <- truths

  fun <- function(x){
    x <- na.omit(x)
    x$mut <- paste(x$chrom, x$coord, x$alt, sep=":")
  }

  data <- lapply(data, fun)
```



```

a <- NULL
z <- NULL
for(i in 1:length(data)){
  a <- data[[i]]
  z <- c(z,a)
}

mut_filter <- z

data <- nuseq_classifiers

fun <- function(x){
  x <- na.omit(x)
  x <- as.data.frame(x, stringsAsFactors=F)
}

data <- lapply(data, fun)

fun <- function(x){

  #Normalise counts by total depth
  x$alt_counts_norm <- (x$alt_counts+0.01)/(x$total_counts+0.01)
  x$alt_forward_counts_norm <- (x$alt_forward_counts+0.01)/(x$total_forward_counts+0.01)
  x$alt_reverse_counts_norm <- (x$alt_reverse_counts+0.01)/(x$total_reverse_counts+0.01)
  x$ref_counts_norm <- (x$ref_counts+0.01)/(x$total_counts+0.01)
  x$ref_forward_counts_norm <- (x$ref_forward_counts+0.01)/(x$total_forward_counts+0.01)
  x$ref_reverse_counts_norm <- (x$ref_reverse_counts+0.01)/(x$total_reverse_counts+0.01)

  #Ratios
  x$alt_bqual_ratio <- (x$alt_bqual+0.01)/(sum(x$alt_bqual)+0.01)
  x$alt_mqual_ratio <- (x$alt_mqual+0.01)/(sum(x$alt_mqual)+0.01)
  x$alt_distance_ratio <- (x$alt_distance+0.01)/(sum(x$alt_distance)+0.01)

  x$ref_bqual_ratio <- (x$ref_bqual+0.01)/(sum(x$ref_bqual)+0.01)
  x$ref_mqual_ratio <- (x$ref_mqual+0.01)/(sum(as.numeric(x$ref_mqual))+0.01)
  x$ref_distance_ratio <- (x$ref_distance+0.01)/(sum(x$ref_distance)+0.01)

  x$total_bqual_ratio <- (x$total_bqual+0.01)/(sum(x$total_bqual)+0.01)
  x$total_mqual_ratio <- (x$total_mqual+0.01)/(sum(as.numeric(x$total_mqual))+0.01)
  x$total_distance_ratio <- (x$total_distance+0.01)/(sum(x$total_distance)+0.01)

  return(x)
}

x <- lapply(data, fun)
data <- x

#Make one matrix for all data and filter mutations
a <- NULL
z <- NULL
for(i in 1:length(x)){
  a <- x[[i]]
  a$sample_id <- names(x)[i]
  a$mutation <- paste(a$chrom,a$coord,a$alt,sep=":")
  a[,!apply(is.na(a[,1:ncol(a)-1]), 2, any)]
  a[a$mutation%in%mut_filter,"snp_class"] <- "true_positive"
  a[!grep("ERCC",a$chrom),"snp_class"] <- "false_positive"
  z <- rbind(z,a)
}

```

```

}

x <- z

data <- x

if(training_set==TRUE){
  data <- na.omit(data)
  write.table(data, file="snv.summary", quote=F, row.names=F, sep="\t")
}

if(training_set==FALSE){
  write.table(data, file="snv.summary", quote=F, row.names=F, sep="\t")
}

x <- data
x <- x[,c(5:(ncol(x)-3))]
print(colnames(x))
cm <- cor(x)
hc <- findCorrelation(cm, cutoff=0.75)

if(length(hc)<20){
  print("too many high correlating classifiers, using all")
}

if(length(hc)>19){
  print("highly correlated classifiers removed, remaining classifiers _
recommended for model")
  x <- x[,hc]
  classifiers <- colnames(x)
  write.table(classifiers, file="low_correlation_classifiers", quote=F,
    row.names=F, sep="\t")
}
}

```

D.2 snv.performance

Function designed to train and test models to check the performance of SNV calling.

```

#' snv.performance
#'
#' This function takes the output files from snv.benchmark, makes a snv prediction
#' model and tests it on hold out data. The output is a confusion matrix, a set of
#' snv predictions and a table for plotting a ROC curve (all applied to test data
#' only, AUC is output in STDOUT).
#' @param snv_data Output from snv.benchmark. A data frame containing all derived snv
#' classifiers with ground truth data appended.
#' @param test_index A vector to determine which samples to train the model on (1)
#' and then test (2). This can be user input to allow consistency between modelling
#' methods if required.
#' @param classifiers Defaults to use all classifiers. We recommend providing a
#' vector of the low correlating classifiers to use in building the prediction model
#' . This is already output from snv.bench.
#' @param threshold Defaults to 0.5. Values 0 to 1 can be input to tune predictions
#' for specificity or sensitivity.
#' @param model.method Defaults to standard random forest, other options are "tuned_
#' randomforest", "svm", and "tuned_svm".
#' @keywords
#' @export
#' @examples
#' snv.performance(snv_data, classifiers, test_index=NULL, threshold=NULL, model.method="
#' tuned_randomforest")

```

```

snv.performance <- function(snv_data, classifiers=NULL, test_index=NULL, threshold=NULL,
  model.method="tuned_randomforest"){

  require(ROCR)
  require(ggplot2)
  require(cowplot)
  require(caret)

  if(is.null(threshold)){
    threshold <- 0.5
  }

  data <- na.omit(snv_data)
  dups <- !duplicated(data$mutation)

  data <- data[dups,]

  if(is.null(test_index)){
    set.seed(2)
    ind <- sample(2, nrow(data), replace = TRUE, prob=c(0.5, 0.5))
    write.table(ind, file="test_index", quote=F, row.names=F, sep="\t")
  }

  if(!is.null(test_index)){
    ind <- test_index
    ind <- ind[dups,]
  }

  x <- data[ind==1,]
  names <- x[,c(1:4, ncol(x)-2, ncol(x)-1)]
  x <- x[,5:(ncol(x)-3)]
  x <- x[,classifiers[,1]]
  y <- as.factor(names$snp_class)
  x <- as.matrix(x)

  #run model
  if(model.method=="randomforest"){
    require(randomForest)
    model <- randomForest(x,y)
  }

  if(model.method=="tuned_randomforest"){
    require(randomForest)
    bestmtry <- NULL
    mtry <- NULL
    for(i in seq(500,2500,500)){
      bestmtry <- as.data.frame(tuneRF(x, y, stepFactor=1.5, trace=F, improve=0.05, ntree=i)
    )
      bestmtry$ntree <- i
      mtry <- rbind(mtry, bestmtry)
    }
    mtry <- mtry[which.min(mtry$OOBError),]
    model <- randomForest(x,y, ntree=bestmtry$ntree, mtry=mtry$mtry)
  }

  if(model.method=="svm"){
    require(e1071)
    model <- svm(x,y, probability=T)
  }

  if(model.method=="tuned_svm"){
    require(e1071)
    svm.tune <- tune(svm,x,y, ranges = list(gamma = 2^(-8:1), cost = 2^(0:4)), tunecontrol
      = tune.control(sampling = "fix"))
    model <- svm(x,y, cost=svm.tune[[1]][[2]], gamma=svm.tune[[1]][[1]], probability=TRUE)
  }

```

```

#Predict with test data
x <- data[ind==2,]
names <- x[,c(1:4,ncol(x)-2,ncol(x)-1)]
x <- x[,5:(ncol(x)-3)]

if(is.null(classifiers)){
  classifiers <- colnames(x)
}

x <- x[,classifiers[,1]]
y <- as.factor(names$snp_class)
x <- as.matrix(x)

model.prob <- predict(model, type="prob", newdata=x, probability=TRUE)

if(grepl("randomforest",model.method)==TRUE){
  model.pred <- prediction(model.prob[,2], names$snp_class)
}

if(grepl("svm",model.method)==TRUE){
  model.pred <- prediction(attr(model.prob, "probabilities")[,1], y)
}

model.cm <- confusionMatrix(as.factor(model.pred@predictions[[1]] > threshold),y=="true
  _positive")

z <- as.data.frame(model.prob)
z$snp_class <- names$snp_class
z$label <- "test"

model.perf <- performance(model.pred,"tpr","fpr")
auc <- performance(model.pred, measure = "auc")
print(auc)

plot <- as.data.frame(cbind(model.perf@x.values[[1]],model.perf@y.values[[1]]))
colnames(plot) <- c("FPR","TPR")

write.table(z, file="predictions",row.names=F,quote=F,sep="\t")
write.table(plot, file="plot_ROC",row.names=F,quote=F,sep="\t")
write.table(model.cm[[2]], file="confusion_matrix",quote=F,sep="\t")
}

```

D.3 snv.cross.predict

Function for making a SNV prediction model with one data set, and then applying it to another data set.

```

#' snv.cross.predict
#'
#' This function makes a snv prediction model on one set of data and then applies it
  to another, typically one without the ground truth information required to build
  a cell type specific model.
#' @param train_data Output from snv.bench. A data frame containing all derived snv
  classifiers with ground truths.
#' @param test_data Output from snv.bench. A data frame containing all derived snv
  classifiers (where training_set=FALSE, though not essential).
#' @param classifiers Output from snv.bench. A vector to keep low correlating
  classifiers only, all used if NULL. If applied to the training data, it will need
  to be used here.
#' @param threshold Defaults to 0.5. Values 0 to 1 can be input to tune predictions
  for specificity or sensitivity.
#' @keywords
#' @export

```

```

#' @examples
#' snv.predict(train_data, test_data, classifiers, threshold)

snv.cross.predict <- function(train_data, test_data, classifiers=NULL, threshold=NULL){

  require(caret)
  require(ROCR)
  require(randomForest)

  x <- na.omit(train_data)
  x <- x[!duplicated(x$mutation),]
  names <- x[,c(1:4, ncol(x)-2, ncol(x)-1)]
  x <- x[,5:(ncol(x)-3)]

  if(is.null(classifiers)){
    classifiers <- as.data.frame(colnames(x), stringsAsFactors = F)
  }

  x <- x[, classifiers[,1]]
  y <- as.factor(names$snp_class)
  x <- as.matrix(x)

  bestmtry <- NULL
  mtry <- NULL
  for(i in seq(500, 2500, 500)){
    bestmtry <- as.data.frame(tuneRF(x, y, stepFactor=1.5, trace=F, plot=FALSE, improve
      =0.05, ntree=i))
    bestmtry$ntree <- i
    mtry <- rbind(mtry, bestmtry)
  }

  mtry <- mtry[which.min(mtry$OOBError),]
  model <- randomForest(x, y, ntree=mtry$ntree, mtry=mtry$mtry)

  x <- na.omit(test_data)
  x <- x[!duplicated(x$mutation),]
  names <- x[,c(1:4, ncol(x)-2, ncol(x)-1)]
  x <- x[,5:(ncol(x)-3)]
  x <- x[, classifiers[,1]]
  x <- as.matrix(x)

  model.prob <- predict(model, type="prob", newdata=x, probability=TRUE)
  z <- as.data.frame(model.prob)
  z <- cbind(names, model.prob)

  write.table(z, file="cross_predictions", row.names=F, quote=F, sep="\t")
}

```

D.4 snv.predict

Function to predict SNVs from data incorporating module to compute additional classifiers from nuseq and remove highly correlating classifiers.

```

#' snv.predict
#'
#' This function predicts snvs in single cell RNA-seq data by creating a model based
#' on true positives provided by the user and false positives from synthetic spike-
#' ins (currently only ERCCs are supported). It will create a directory in the
#' working directory called snv.predict.results containing a table of snv
#' predictions.
#' @param nuseq_classifiers List of named data frames as output from nuseq-sc.

```

```

#' @param truths List of named data.frames with snv information. Easiest to use the
#' convenience function 'convert_to_pos' and add a fifth column with snvs annotated
#' as true_positive (1), false_positive (0) and unknown (NA).
#' @keywords
#' @export
#' @examples
#' snv.predict(classifiers, truths)

snv.predict <- function(nuseq_classifiers, truths){

  require(caret)
  require(ROCR)
  require(randomForest)

  data <- nuseq_classifiers

  fun <- function(x){
    x <- na.omit(x)
  }

  data <- lapply(data, fun)

  print("loading_ground_truths_and_checking_samples_match")

  x <- truths

  a <- NULL
  z <- NULL
  for(i in 1:length(x)){
    a <- x[[i]]
    z <- rbind(z, a)
  }

  gt <- z

  print("generating_new_classifiers")

  fun <- function(x){

    #Normalise counts by total depth
    x$alt_counts_norm <- (x$alt_counts+0.01)/(x$total_counts+0.01)
    x$alt_forward_counts_norm <- (x$alt_forward_counts+0.01)/(x$total_forward_counts
    +0.01)
    x$alt_reverse_counts_norm <- (x$alt_reverse_counts+0.01)/(x$total_reverse_counts
    +0.01)
    x$ref_counts_norm <- (x$ref_counts+0.01)/(x$total_counts+0.01)
    x$ref_forward_counts_norm <- (x$ref_forward_counts+0.01)/(x$total_forward_counts
    +0.01)
    x$ref_reverse_counts_norm <- (x$ref_reverse_counts+0.01)/(x$total_reverse_counts
    +0.01)

    #Ratios
    x$alt_bqual_ratio <- (x$alt_bqual+0.01)/(sum(x$alt_bqual)+0.01)
    x$alt_mqual_ratio <- (x$alt_mqual+0.01)/(sum(x$alt_mqual)+0.01)
    x$alt_distance_ratio <- (x$alt_distance+0.01)/(sum(x$alt_distance)+0.01)

    x$ref_bqual_ratio <- (x$ref_bqual+0.01)/(sum(x$ref_bqual)+0.01)
    x$ref_mqual_ratio <- (x$ref_mqual+0.01)/(sum(as.numeric(x$ref_mqual))+0.01)
    x$ref_distance_ratio <- (x$ref_distance+0.01)/(sum(x$ref_distance)+0.01)

    x$total_bqual_ratio <- (x$total_bqual+0.01)/(sum(x$total_bqual)+0.01)
    x$total_mqual_ratio <- (x$total_mqual+0.01)/(sum(as.numeric(x$total_mqual))+0.01)
    x$total_distance_ratio <- (x$total_distance+0.01)/(sum(x$total_distance)+0.01)

    x <- as.data.frame(x, stringsAsFactors=F)

    return(x)
  }
}

```

```

x <- lapply(data, fun)

#Make training matrix
a <- NULL
z <- NULL
for(i in 1:length(x)){
  a <- x[[i]]
  a$sample_id <- names(x)[i]
  z <- rbind(z,a)
}
cm <- z

print("deduplicating_and_adding_ground_truths")

x <- paste(cm$chrom,cm$coord,cm$alt, sep=":")
gt <- na.omit(gt)
gt <- gt[gt$label==1,]
y <- paste(gt$chrom,gt$coord,gt$alt, sep=":")
calls <- x%in%y
cm[calls==TRUE,"snp_class"] <- "true_positive"
cm[grep("ERCC",cm$chrom),"snp_class"] <- "false_positive"
dups <- !duplicated(paste(cm$chrom,cm$coord,cm$alt, sep=":"))
data <- cm
cm <- cm[dups,]

print("filtering_classifiers")

cor.m <- cor(cm[,c(5:(ncol(cm)-2))])
hc <- findCorrelation(cor.m, cutoff=0.75)

if(length(hc)<20){
  print("too_many_high_correlating_classifiers,_using_all")
  hc <- colnames(cm[,c(5:(ncol(x)-2))])
}

if(length(hc)>19){
  print("highly_correlated_classifiers_removed,_printing_remaining_classifiers_used_for_modelling")
  print(colnames(cm[,hc+4]))
}

#Train the model
cm <- na.omit(cm)
y <- as.factor(cm$snp_class)
x <- cm[,hc+4]

print("tuning_model")

bestmtry <- NULL
mtry <- NULL
for(i in seq(500,2500,500)){
  bestmtry <- as.data.frame(tuneRF(x, y, stepFactor=1.5,trace=F,plot=FALSE, improve
    =0.05, ntree=i))
  bestmtry$ntree <- i
  mtry <- rbind(mtry,bestmtry)
}
mtry <- mtry[which.min(mtry$OOBError),]
model <- randomForest(x,y,ntree=mtry$ntree,mtry=mtry$mtry)

print("predicting_snvs")
x <- as.matrix(data[,5:(ncol(data)-2)])

model.prob <- predict(model, type="prob", newdata=x, probability=TRUE)

x <- as.data.frame(cbind(data[,c(1:4,ncol(data))],model.prob))

print("saving_outputs")

```

```
write.table(x, file=paste0(getwd(), "/predicted_snps"), row.names=F, quote=F, sep="\t")
write.table(plot, file=paste0(getwd(), "/plot_ROC"), row.names=F, quote=F, sep="\t")

print("output_data_saved")

}
```


Total approximate word count (excluding references, appendices, diagrams and tables): 43000
Total pages: 262

References

- Abdel-Latif, M. M. M., J. O’Riordan, H. J. Windle, E. Carton, N. Ravi, D. Kelleher, and J. V. Reynolds
2004. Nf-kappab activation in esophageal adenocarcinoma: relationship to barrett’s metaplasia, survival, and response to neoadjuvant chemoradiotherapy. *Annals of surgery*, 239:491–500.
- Abdo, J., D. K. Agrawal, and S. K. Mittal
2017. Basis for molecular diagnostics and immunotherapy for esophageal cancer. *Expert review of anticancer therapy*, 17:33–45.
- Ai, X., T. Kitazawa, A.-T. Do, M. Kusche-Gullberg, P. A. Labosky, and C. P. Emerson
2007. Sulf1 and sulf2 regulate heparan sulfate-mediated gdnf signaling for esophageal innervation. *Development (Cambridge, England)*, 134:3327–3338.
- Aikou, S., J. Aida, K. Takubo, Y. Yamagata, Y. Seto, M. Kaminishi, and S. Nomura
2013. Columnar metaplasia in a surgical mouse model of gastro-esophageal reflux disease is not derived from bone marrow-derived cell. *Cancer science*, 104(9):1154–1161.
- Ajani, J. A., W. Rodriguez, G. Bodoky, V. Moiseyenko, M. Lichinitser, V. Gorbunova, I. Vynnychenko, A. Garin, I. Lang, and S. Falcon
2010. Multicenter phase iii comparison of cisplatin/s-1 with cisplatin/infusional fluorouracil in advanced gastric or gastroesophageal adenocarcinoma study: the flags trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28:1547–1553.
- Alaynick, W. A., J. M. Way, S. A. Wilson, W. G. Benson, L. Pei, M. Downes, R. Yu, J. W. Jonker, J. A. Holt, D. K. Rajpal, H. Li, J. Stuart, R. McPherson, K. S. Remlinger, C.-Y. Chang, D. P. McDonnell, R. M. Evans, and A. N. Billin
2010. Errgamma regulates cardiac, gastric, and renal potassium homeostasis. *Molecular endocrinology (Baltimore, Md.)*, 24:299–309.
- Alexandrov, L. B., P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, and M. R. Stratton
2015. Clock-like mutational processes in human somatic cells. *Nature genetics*, 47:1402–1407.

- Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Brresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjrd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, M. Imielinsk, N. Jger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lpez-Otn, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valds-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, A. P. C. G. Initiative, I. B. C. Consortium, I. M.-S. Consortium, I. Ped-Brain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton
2013. Signatures of mutational processes in human cancer. *Nature*, 500:415–421.
- Allison, P. R. and A. S. Johnstone
1953. The oesophagus lined with gastric mucous membrane. *Thorax*, 8:87–101.
- Anders, S., P. T. Pyl, and W. Huber
2015. Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31:166–169.
- Angermueller, C., S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. Smallwood, C. P. Ponting, T. Voet, G. Kelsey, O. Stegle, and W. Reik
2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13:229–232.
- Arnold, K., A. Sarkar, M. A. Yram, J. M. Polo, R. Bronson, S. Sengupta, M. Seandel, N. Geijsen, and K. Hochedlinger
2011. Sox2(+) adult stem and progenitor cells are important for tissue regeneration and survival of mice. *Cell stem cell*, 9:317–329.
- Artavanis-Tsakonas, S., M. D. Rand, and R. J. Lake
1999. Notch signaling: cell fate control and signal integration in development. *Science (New York, N.Y.)*, 284:770–776.
- Association, A. G., S. J. Spechler, P. Sharma, R. F. Souza, J. M. Inadomi, and N. J. Shaheen
2011. American gastroenterological association medical position statement on the management of barrett’s esophagus. *Gastroenterology*, 140:1084–1091.
- Attar, M., E. Sharma, S. Li, C. Bryer, L. Cubitt, J. Broxholme, H. Lockstone, J. Kinchen, A. Simmons, P. Piazza, et al.
2018. A practical solution for preserving single cells for rna sequencing. *Scientific reports*, 8(1):2151.

- Attwood, S. E., L.-A. Harrison, S. L. Preston, and J. A. Jankowski
2008. Esophageal adenocarcinoma in "mice and men": back to basics! *The American journal of gastroenterology*, 103:2367–2372.
- Attwood, S. E., T. C. Smyrk, T. R. DeMeester, S. S. Mirvish, H. J. Stein, and R. A. Hinder
1992. Duodeno-esophageal reflux and the development of esophageal adenocarcinoma in rats. *Surgery*, 111:503–510.
- Aubin, J., U. Dry, M. Lemieux, P. Chailier, and L. Jeannotte
2002. Stomach regional specification requires *hoxa5*-driven mesenchymal-epithelial signaling. *Development (Cambridge, England)*, 129:4075–4087.
- Bacher, R. and C. Kendzierski
2016. Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17(1):63.
- Bailey, T., L. Biddlestone, N. Shepherd, H. Barr, P. Warner, and J. Jankowski
1998. Altered cadherin and catenin complexes in the barrett's esophagus-dysplasia-adenocarcinoma sequence: correlation with disease progression and dedifferentiation. *The American journal of pathology*, 152:135–144.
- Baker, A.-M., B. Cereser, S. Melton, A. G. Fletcher, M. Rodriguez-Justo, P. J. Tadrous, A. Humphries, G. Elia, S. A. C. McDonald, N. A. Wright, B. D. Simons, M. Jansen, and T. A. Graham
2014. Quantification of crypt and stem cell evolution in the normal and neoplastic human colon. *Cell reports*, 8:940–947.
- Bandura, D. R., V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner
2009. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical chemistry*, 81:6813–6822.
- Bang, Y.-J., E. Van Cutsem, A. Feyereislova, H. C. Chung, L. Shen, A. Sawaki, F. Lordick, A. Ohtsu, Y. Omuro, T. Satoh, G. Aprile, E. Kulikov, J. Hill, M. Lehle, J. Rschoff, Y.-K. Kang, and T. T. Investigators
2010. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of her2-positive advanced gastric or gastro-oesophageal junction cancer (toga): a phase 3, open-label, randomised controlled trial. *Lancet (London, England)*, 376:687–697.
- Baran-Gale, J., T. Chandra, and K. Kirschner
2017. Experimental design for single-cell rna sequencing. *Briefings in functional genomics*.

- Barbera, M., M. di Pietro, E. Walker, C. Brierley, S. MacRae, B. D. Simons, P. H. Jones, J. Stingl, and R. C. Fitzgerald
2015. The human squamous oesophagus has widespread capacity for clonal expansion from cells at diverse stages of differentiation. *Gut*, 64:11–19.
- Barker, N., M. Huch, P. Kujala, M. van de Wetering, H. J. Snippert, J. H. van Es, T. Sato, D. E. Stange, H. Begthel, M. van den Born, E. Danenberg, S. van den Brink, J. Korving, A. Abo, P. J. Peters, N. Wright, R. Poulsom, and H. Clevers
2010. Lgr5(+ve) stem cells drive self-renewal in the stomach and build long-lived gastric units in vitro. *Cell stem cell*, 6:25–36.
- Barker, N., J. H. van Es, J. Kuipers, P. Kujala, M. van den Born, M. Cozijnsen, A. Haegebarth, J. Korving, H. Begthel, P. J. Peters, and H. Clevers
2007. Identification of stem cells in small intestine and colon by marker gene lgr5. *Nature*, 449:1003–1007.
- Barker, N., A. van Oudenaarden, and H. Clevers
2012. Identifying the stem cell of the intestinal crypt: strategies and pitfalls. *Cell stem cell*, 11:452–460.
- Barrett, M. T., C. A. Sanchez, L. J. Prevo, D. J. Wong, P. C. Galipeau, T. G. Paulson, P. S. Rabinovitch, and B. J. Reid
1999. Evolution of neoplastic cell lineages in barrett oesophagus. *Nature genetics*, 22:106–109.
- Barrett, N.
1950. Chronic peptic ulcer of the oesophagus and oesophagitis. *British Journal of Surgery*, 38(150):175–182.
- Barrett, N. R.
1957. The lower esophagus lined by columnar epithelium. *Surgery*, 41:881–894.
- Barriga, F. M., E. Montagni, M. Mana, M. Mendez-Lago, X. Hernando-Momblona, M. Sevillano, A. Guillaumet-Adkins, G. Rodriguez-Esteban, S. J. A. Buczacki, M. Gut, H. Heyn, D. J. Winton, O. H. Yilmaz, C. S.-O. Attolini, I. Gut, and E. Batlle
2017. Mex3a marks a slowly dividing subpopulation of lgr5+ intestinal stem cells. *Cell stem cell*, 20:801–816.e7.
- Bartman, A. E., M. P. Buisine, J. P. Aubert, G. A. Niehans, N. W. Toribara, Y. S. Kim, E. J. Kelly, J. E. Crabtree, and S. B. Ho
1998. The muc6 secretory mucin gene is expressed in a wide variety of epithelial tissues. *The Journal of pathology*, 186:398–405.
- Bendall, S. C., K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er
2014. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157:714–725.

- Bennett, C., P. Moayyedi, D. A. Corley, J. DeCaestecker, Y. Falck-Ytter, G. Falk, N. Vakil, S. Sanders, M. Vieth, J. Inadomi, D. Aldulaimi, K.-Y. Ho, R. Odze, S. J. Meltzer, E. Quigley, S. Gittens, P. Watson, G. Zaninotto, P. G. Iyer, L. Alexandre, Y. Ang, J. Callaghan, R. Harrison, R. Singh, P. Bhandari, R. Bisschops, B. Geramizadeh, P. Kaye, S. Krishnadath, M. B. Fennerty, H. Manner, K. S. Nason, O. Pech, V. Konda, K. Ragunath, I. Rahman, Y. Romero, R. Sampliner, P. D. Siersema, J. Tack, T. C. K. Tham, N. Trudgill, D. S. Weinberg, J. Wang, K. Wang, J. Y. Y. Wong, S. Attwood, P. Malfertheiner, D. MacDonald, H. Barr, M. K. Ferguson, J. Jankowski, and B. C. Consortium
2015. Bob cat: A large-scale review and delphi consensus for management of barrett’s esophagus with no dysplasia, indefinite for, or low-grade dysplasia. *The American journal of gastroenterology*, 110:662–82; quiz 683.
- Bjerknes, M. and H. Cheng
2002. Multipotential stem cells in adult mouse gastric epithelium. *American journal of physiology. Gastrointestinal and liver physiology*, 283:G767–G777.
- Bjerknes, M. and H. Cheng
2010. Cell lineage metastability in gfi1-deficient mouse intestinal epithelium. *Developmental biology*, 345:49–63.
- Blainey, P. C. and S. R. Quake
2011. Digital mda for enumeration of total nucleic acid contamination. *Nucleic acids research*, 39:e19.
- Blazeby, J. M., J. R. Farndon, J. Donovan, and D. Alderson
2000. A prospective longitudinal study examining the quality of life of patients with esophageal carcinoma. *Cancer*, 88:1781–1787.
- Brahmer, J. R., C. G. Drake, I. Wollner, J. D. Powderly, J. Picus, W. H. Sharfman, E. Stankevich, A. Pons, T. M. Salay, T. L. McMiller, M. M. Gilson, C. Wang, M. Selby, J. M. Taube, R. Anders, L. Chen, A. J. Korman, D. M. Pardoll, I. Lowy, and S. L. Topalian
2010. Phase i study of single-agent anti-programmed death-1 (mdx-1106) in refractory solid tumors: safety, clinical activity, pharmacodynamics, and immunologic correlates. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28:3167–3175.
- Buenrostro, J. D., B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf
2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523:486–490.
- Cancer Genome Atlas Research Network
2014. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513:202–209.

Cancer Genome Atlas Research Network

2017. Integrated genomic characterization of oesophageal carcinoma. *Nature*, 541(7636):169–175.

Castillo, D., S. Puig, M. Iglesias, A. Seoane, C. de Bols, V. Munitiz, P. Parrilla, L. Comerma, R. Poulsom, K. K. Krishnadath, L. Grande, and M. Pera

2012. Activation of the bmp4 pathway and early expression of cdx2 characterize non-specialized columnar metaplasia in a human model of barrett’s esophagus. *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract*, 16:227–37; discussion 237.

Catenacci, D. V. T., N. C. Tebbutt, I. Davidenko, A. M. Murad, S.-E. Al-Batran, D. H. Ilson, S. Tjulandin, E. Gotovkin, B. Karaszewska, I. Bondarenko, M. A. Tejani, A. A. Udrea, M. Tehfe, F. De Vita, C. Turkington, R. Tang, A. Ang, Y. Zhang, T. Hoang, R. Sidhu, and D. Cunningham

2017. Rilotumumab plus epirubicin, cisplatin, and capecitabine as first-line therapy in advanced met-positive gastric or gastro-oesophageal junction cancer (rilomet-1): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet. Oncology*, 18:1467–1482.

Chandrakesan, P., J. Yao, D. Qu, R. May, N. Weygant, Y. Ge, N. Ali, S. M. Sureban, M. Gude, K. Vega, E. Bannerman-Menson, L. Xia, M. Bronze, G. An, and C. W. Houchen

2017. Dclk1, a tumor stem cell marker, regulates pro-survival signaling and self-renewal of intestinal tumor cells. *Molecular cancer*, 16:30.

Chandrasoma, P. T., R. Der, P. Dalton, G. Kobayashi, Y. Ma, J. Peters, and T. De-meester

2001. Distribution and significance of epithelial types in columnar-lined esophagus. *The American journal of surgical pathology*, 25:1188–1193.

Chaves, P., C. Cruz, A. Dias Pereira, A. Suspiro, J. C. M. de Almeida, C. N. Leito, and J. Soares

2005. Gastric and intestinal differentiation in barrett’s metaplasia and associated adenocarcinoma. *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus*, 18:383–387.

Chen, J., Q. Zhou, Y. Wang, and K. Ning

2016. Single-cell snp analyses and interpretations based on rna-seq data for colon cancer research. *Scientific reports*, 6:34420.

Chen, K. H., A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang

2015. Rna imaging. spatially resolved, highly multiplexed rna profiling in single cells. *Science (New York, N.Y.)*, 348:aaa6090.

Cheow, L. F., E. T. Courtois, Y. Tan, R. Viswanathan, Q. Xing, R. Z. Tan, D. S. W.

- Tan, P. Robson, Y.-H. Loh, S. R. Quake, and W. F. Burkholder
2016. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature methods*, 13:833–836.
- Choi, M. Y., A. I. Romer, Y. Wang, M. P. Wu, S. Ito, A. B. Leiter, and R. A. Shivdasani
2008. Requirement of the tissue-restricted homeodomain transcription factor nkx6.3 in differentiation of gastrin-producing g cells in the stomach antrum. *Molecular and cellular biology*, 28:3208–3218.
- Chung, W., H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, and W.-Y. Park
2017. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications*, 8:15081.
- Clark, R. J., M. P. Craig, S. Agrawal, and M. Kadakia
2018a. microrna involvement in the onset and progression of barrett’s esophagus: a systematic review. *Oncotarget*, 9:8179–8196.
- Clark, S. J., R. Argelaguet, C.-A. Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, O. Stegle, and W. Reik
2018b. scnm-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature communications*, 9:781.
- Clemons, N. J., D. H. Wang, D. Croagh, A. Tikoo, C. M. Fennell, C. Murone, A. M. Scott, D. N. Watkins, and W. A. Phillips
2012. Sox9 drives columnar differentiation of esophageal squamous epithelium: a possible role in the pathogenesis of barrett’s esophagus. *American journal of physiology. Gastrointestinal and liver physiology*, 303:G1335–G1346.
- Coad, R. A., A. C. Woodman, P. J. Warner, H. Barr, N. A. Wright, and N. A. Shepherd
2005. On the histogenesis of barrett’s oesophagus and its associated squamous islands: a three-dimensional study of their morphological relationship with native oesophageal gland ducts. *The Journal of pathology*, 206:388–394.
- Conroy, T., M.-P. Galais, J.-L. Raoul, O. Bouch, S. Gourgou-Bourgade, J.-Y. Douillard, P.-L. Etienne, V. Boige, I. Martel-Lafay, P. Michel, C. Llacer-Moscardo, E. Franois, G. Crhange, M. B. Abdelghani, B. Juzyna, L. Bedenne, A. Adenis, F. F. de Cancrologie Digestive, and U.-G. Group
2014. Definitive chemoradiotherapy with folfox versus fluorouracil and cisplatin in patients with oesophageal cancer (prodige5/accord17): final results of a randomised, phase 2/3 trial. *The Lancet. Oncology*, 15:305–314.
- Cooper, J., Y. Ding, J. Song, and K. Zhao
2017. Genome-wide mapping of dnase i hypersensitive sites in rare cell populations using single-cell dnase sequencing. *Nature protocols*, 12:2342–2354.

- Crawford, S. E., V. Stellmach, J. E. Murphy-Ullrich, S. M. Ribeiro, J. Lawler, R. O. Hynes, G. P. Boivin, and N. Bouck
1998. Thrombospondin-1 is a major activator of *tgf-beta1* in vivo. *Cell*, 93:1159–1170.
- Crosby, T., C. N. Hurt, S. Falk, S. Gollins, J. Staffurth, R. Ray, J. A. Bridgewater, J. I. Geh, D. Cunningham, J. Blazeby, R. Roy, T. Maughan, G. Griffiths, and S. Mukherjee
2017. Long-term results and recurrence patterns from scope-1: a phase ii/iii randomised trial of definitive chemoradiotherapy +/- cetuximab in oesophageal cancer. *British journal of cancer*, 116:709–716.
- Crosby, T. D. L., A. E. Brewster, A. Borley, L. Perschky, P. Kehagioglou, J. Court, and T. S. Maughan
2004. Definitive chemoradiation in patients with inoperable oesophageal carcinoma. *British journal of cancer*, 90:70–75.
- CRUK
2016. <http://www.cancerresearchuk.org/about-cancer/oesophageal-cancer>. Accessed Mar 2017.
- Cunliffe, R. N., F. R. Rose, J. Keyte, L. Abberley, W. C. Chan, and Y. R. Mahida
2001. Human defensin 5 is stored in precursor form in normal paneth cells and is expressed by some villous epithelial cells and by metaplastic paneth cells in the colon in inflammatory bowel disease. *Gut*, 48:176–185.
- Cunningham, D., A. F. Okines, and S. Ashley
2010. Capecitabine and oxaliplatin for advanced esophagogastric cancer. *New England Journal of Medicine*, 362(9):858–859.
- Darmanis, S., C. J. Gallant, V. D. Marinescu, M. Niklasson, A. Segerman, G. Flamentourakis, S. Fredriksson, E. Assarsson, M. Lundberg, S. Nelander, B. Westermarck, and U. Landegren
2016. Simultaneous multiplexed measurement of rna and proteins in single cells. *Cell reports*, 14:380–389.
- Das, K. M., Y. Kong, M. Bajpai, D. Kulkarni, X. Geng, P. Mishra, D. Banerjee, and K. Hirshfield
2011. Transformation of benign barrett’s epithelium by repeated acid and bile exposure over 65 weeks: a novel in vitro model. *International journal of cancer*, 128:274–282.
- Daud, A. I., J. D. Wolchok, C. Robert, W.-J. Hwu, J. S. Weber, A. Ribas, F. S. Hodi, A. M. Joshua, R. Kefford, P. Hersey, R. Joseph, T. C. Gangadhar, R. Dronca, A. Patnaik, H. Zarour, C. Roach, G. Toland, J. K. Lunceford, X. N. Li, K. E. Emanuel, M. Dolled-Filhart, S. P. Kang, S. Ebbinghaus, and O. Hamid
2016. Programmed death-ligand 1 expression and response to the anti-programmed

- death 1 antibody pembrolizumab in melanoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 34:4102–4109.
- Davelaar, A. L., D. Straub, N. S. Buttar, P. Fockens, and K. K. Krishnadath
2015. Active matrix metalloproteases are expressed early on and are high during the barrett’s esophagus malignancy sequence. *Scandinavian journal of gastroenterology*, 50:321–332.
- Davenport, C., U. Diekmann, I. Budde, N. Detering, and O. Naujok
2016. Anterior-posterior patterning of definitive endoderm generated from human embryonic stem cells depends on the differential signaling of retinoic acid, wnt-, and bmp-signaling. *Stem cells (Dayton, Ohio)*, 34:2635–2647.
- Del Portillo, A., S. M. Lagana, Y. Yao, T. Uehara, N. Jhala, T. Ganguly, P. Nagy, J. Gutierrez, A. Luna, J. Abrams, Y. Liu, R. Brand, J. L. Sepulveda, G. W. Falk, and A. R. Sepulveda
2015. Evaluation of mutational testing of preneoplastic barrett’s mucosa by next-generation sequencing of formalin-fixed, paraffin-embedded endoscopic samples for detection of concurrent dysplasia and adenocarcinoma in barrett’s esophagus. *The Journal of molecular diagnostics : JMD*, 17:412–419.
- DeNardi, F. G. and R. H. Riddell
1991. The normal esophagus. *The American journal of surgical pathology*, 15:296–309.
- Dent, J., H. El-Serag, M. Wallander, and S. Johansson
2005. Epidemiology of gastro-oesophageal reflux disease: a systematic review. *Gut*, 54(5):710–717.
- Dey, S. S., L. Kester, B. Spanjaard, M. Bienko, and A. van Oudenaarden
2015. Integrated genome and transcriptome sequencing of the same cell. *Nature biotechnology*, 33:285–289.
- Dhupkar, P., N. Gordon, J. Stewart, and E. S. Kleinerman
2018. Anti-pd-1 therapy redirects macrophages from an m2 to an m1 phenotype inducing regression of os lung metastases. *Cancer medicine*, 7:2654–2664.
- Ding, B., L. Zheng, Y. Zhu, N. Li, H. Jia, R. Ai, A. Wildberg, and W. Wang
2015. Normalization and noise reduction for single cell rna-seq experiments. *Bioinformatics (Oxford, England)*, 31:2225–2227.
- Ding, J., A. Bashashati, A. Roth, A. Oloumi, K. Tse, T. Zeng, G. Haffari, M. Hirst, M. A. Marra, A. Condon, S. Aparicio, and S. P. Shah
2012. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics (Oxford, England)*, 28:167–175.
- 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.

- Dong, J., M. F. Buas, P. Gharahkhani, B. J. Kendall, L. Onstad, S. Zhao, L. A. Anderson, A. H. Wu, W. Ye, N. C. Bird, L. Bernstein, W.-H. Chow, M. D. Gammon, G. Liu, C. Caldas, P. D. Pharoah, H. A. Risch, P. G. Iyer, B. J. Reid, L. J. Hardie, J. Lagergren, N. J. Shaheen, D. A. Corley, R. C. Fitzgerald, Stomach, O. C. S. S. consortium, D. C. Whiteman, T. L. Vaughan, and A. P. Thrift
2018. Determining risk of barrett’s esophagus and esophageal adenocarcinoma based on epidemiologic factors and genetic variants. *Gastroenterology*, 154:1273–1281.e3.
- Doup, D. P., M. P. Alcolea, A. Roshan, G. Zhang, A. M. Klein, B. D. Simons, and P. H. Jones
2012. A single progenitor population switches behavior to maintain and repair esophageal epithelium. *Science (New York, N.Y.)*, 337:1091–1093.
- Drost, J., R. van Boxtel, F. Blokzijl, T. Mizutani, N. Sasaki, V. Sasselli, J. de Ligt, S. Behjati, J. E. Grolleman, T. van Wezel, S. Nik-Zainal, R. P. Kuiper, E. Cuppen, and H. Clevers
2017. Use of crispr-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science (New York, N.Y.)*, 358:234–238.
- Du, A., K. W. McCracken, E. R. Walp, N. A. Terry, T. J. Klein, A. Han, J. M. Wells, and C. L. May
2012. Arx is required for normal enteroendocrine cell development in mice and humans. *Developmental biology*, 365:175–188.
- Duan, S., G. Yuan, X. Liu, R. Ren, J. Li, W. Zhang, J. Wu, X. Xu, L. Fu, Y. Li, J. Yang, W. Zhang, R. Bai, F. Yi, K. Suzuki, H. Gao, C. R. Esteban, C. Zhang, J. C. Izpisua Belmonte, Z. Chen, X. Wang, T. Jiang, J. Qu, F. Tang, and G.-H. Liu
2015. Pten deficiency reprogrammes human neural stem cells towards a glioblastoma stem cell-like phenotype. *Nature communications*, 6:10068.
- Dulak, A. M., P. Stojanov, S. Peng, M. S. Lawrence, C. Fox, C. Stewart, S. Bandla, Y. Imamura, S. E. Schumacher, E. Shefler, A. McKenna, S. L. Carter, K. Cibulskis, A. Sivachenko, G. Saksena, D. Voet, A. H. Ramos, D. Auclair, K. Thompson, C. Sougnez, R. C. Onofrio, C. Guiducci, R. Beroukhim, Z. Zhou, L. Lin, J. Lin, R. Reddy, A. Chang, R. Landrenau, A. Pennathur, S. Ogino, J. D. Luketich, T. R. Golub, S. B. Gabriel, E. S. Lander, D. G. Beer, T. E. Godfrey, G. Getz, and A. J. Bass
2013. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature genetics*, 45:478–486.
- DuPage, M., C. Mazumdar, L. M. Schmidt, A. F. Cheung, and T. Jacks
2012. Expression of tumour-specific antigens underlies cancer immunoediting. *Nature*, 482:405–409.

- Du, A., M. Robinson, and C. Soneson
2018. A systematic performance evaluation of clustering methods for single-cell rna-seq data [version 1; referees: 1 approved with reservations]. *F1000Research*, 7(1141).
- Dvorak, K., M. Chavarria, C. M. Payne, L. Ramsey, C. Crowley-Weber, B. Dvorakova, B. Dvorak, H. Bernstein, H. Holubec, R. E. Sampliner, C. Bernstein, A. Prasad, S. B. Green, and H. Garewal
2007a. Activation of the interleukin-6/stat3 antiapoptotic pathway in esophageal cells by bile acids and low ph: relevance to barrett’s esophagus. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 13:5305–5313.
- Dvorak, K., C. M. Payne, M. Chavarria, L. Ramsey, B. Dvorakova, H. Bernstein, H. Holubec, R. E. Sampliner, N. Guy, A. Condon, C. Bernstein, S. B. Green, A. Prasad, and H. S. Garewal
2007b. Bile acids in combination with low ph induce oxidative stress and oxidative dna damage: relevance to the pathogenesis of barrett’s oesophagus. *Gut*, 56:763–771.
- Dvorakova, K., C. M. Payne, L. Ramsey, H. Holubec, R. Sampliner, J. Dominguez, B. Dvorak, H. Bernstein, C. Bernstein, A. Prasad, R. Fass, H. Cui, and H. Garewal
2004. Increased expression and secretion of interleukin-6 in patients with barrett’s esophagus. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 10:2020–2028.
- Edsgrd, D., B. Reinius, and R. Sandberg
2016. scphaser: haplotype inference using single-cell rna-seq data. *Bioinformatics (Oxford, England)*, 32:3038–3040.
- Ellis, F. H., X. Xu, M. H. Kulke, J. LoCicero, and M. Loda
2001. Malignant transformation of the esophageal mucosa is enhanced in p27 knock-out mice. *The Journal of thoracic and cardiovascular surgery*, 122:809–814.
- Enge, M., H. E. Arda, M. Mignardi, J. Beausang, R. Bottino, S. K. Kim, and S. R. Quake
2017. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell*, 171:321–330.e14.
- Fan, H. C., G. K. Fu, and S. P. A. Fodor
2015. Expression profiling. combinatorial labeling of single cells for gene expression cytometry. *Science (New York, N.Y.)*, 347:1258367.
- Farlik, M., N. C. Sheffield, A. Nuzzo, P. Datlinger, A. Schnegger, J. Klughammer, and C. Bock
2015. Single-cell dna methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell reports*, 10:1386–1397.

- Farr, R., H. van Malenstein, R. De Vos, K. Geboes, I. Depoortere, P. Vanden Berghe, F. Fornari, K. Blondeau, V. Mertens, J. Tack, and D. Sifrim
2008. Short exposure of oesophageal mucosa to bile acids, both in acidic and weakly acidic conditions, can impair mucosal integrity and provoke dilated intercellular spaces. *Gut*, 57:1366–1374.
- Fein, M., J. H. Peters, N. Baril, M. McGarvey, P. Chandrasoma, D. Shibata, P. W. Laird, and K. A. Skinner
1999. Loss of function of *trp53*, but not *apc*, leads to the development of esophageal adenocarcinoma in mice with jejunoesophageal reflux. *The Journal of surgical research*, 83:48–55.
- Findlay, J. M., F. Castro-Giner, S. Makino, E. Rayner, C. Kartsonaki, W. Cross, M. Kovac, D. Ulahannan, C. Palles, R. S. Gillies, et al.
2016. Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nature communications*, 7.
- Fitzgerald, R., B. Onwuegbusi, M. Bajaj-Elliott, I. Saeed, W. Burnham, and M. Farthing
2002. Diversity in the oesophageal phenotypic response to gastro-oesophageal reflux: immunological determinants. *Gut*, 50(4):451–459.
- Fitzgerald, R. C.
2006. Molecular basis of barrett’s oesophagus and oesophageal adenocarcinoma. *Gut*, 55:1810–1820.
- Fitzgerald, R. C., M. di Pietro, K. Ragunath, Y. Ang, J.-Y. Kang, P. Watson, N. Trudgill, P. Patel, P. V. Kaye, S. Sanders, M. O’Donovan, E. Bird-Lieberman, P. Bhandari, J. A. Jankowski, S. Attwood, S. L. Parsons, D. Loft, J. Lagergren, P. Moayyedi, G. Lyratzopoulos, J. de Caestecker, and B. S. of Gastroenterology
2014. British society of gastroenterology guidelines on the diagnosis and management of barrett’s oesophagus. *Gut*, 63:7–42.
- Florell, S. R., C. M. Coffin, J. A. Holden, J. W. Zimmermann, J. W. Gerwels, B. K. Summers, D. A. Jones, and S. A. Leachman
2001. Preservation of rna for functional genomic studies: a multidisciplinary tumor bank protocol. *Modern pathology*, 14(2):116.
- Fuchs, C. S., T. Doi, R. W. Jang, K. Muro, T. Satoh, M. Machado, W. Sun, S. I. Jalal, M. A. Shah, J.-P. Metges, M. Garrido, T. Golan, M. Mandala, Z. A. Wainberg, D. V. Catenacci, A. Ohtsu, K. Shitara, R. Geva, J. Bleeker, A. H. Ko, G. Ku, P. Philip, P. C. Enzinger, Y.-J. Bang, D. Levitan, J. Wang, M. Rosales, R. P. Dalal, and H. H. Yoon
2018. Safety and efficacy of pembrolizumab monotherapy in patients with previously treated advanced gastric and gastroesophageal junction cancer: Phase 2 clinical keynote-059 trial. *JAMA oncology*, 4:e180013.

- Gao, S., L. Yan, R. Wang, J. Li, J. Yong, X. Zhou, Y. Wei, X. Wu, X. Wang, X. Fan, et al.
2018. Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell rna-sequencing. *Nature cell biology*, 20(6):721.
- Garcia, E., A. Hayden, C. Birts, E. Britton, A. Cowie, K. Pickard, M. Mellone, C. Choh, M. Derouet, P. Duriez, F. Noble, M. J. White, J. N. Primrose, J. C. Strefford, M. Rose-Zerilli, G. J. Thomas, Y. Ang, A. D. Sharrocks, R. C. Fitzgerald, T. J. Underwood, and O. consortium
2016. Authentication and characterisation of a new oesophageal adenocarcinoma cell line: Mfd-1. *Scientific reports*, 6:32417.
- Garrison E, M. G.
2012. Haplotype-based variant detection from short-read sequencing. *arXiv*.
- Gawad, C., W. Koh, and S. R. Quake
2014. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences of the United States of America*, 111:17947–17952.
- Gehring, J. S., B. Fischer, M. Lawrence, and W. Huber
2015. Somaticsignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics (Oxford, England)*, 31:3673–3675.
- Gerbe, F., B. Brulin, L. Makrini, C. Legraverend, and P. Jay
2009. Dcamkl-1 expression identifies tuft cells rather than stem cells in the adult mouse intestinal epithelium. *Gastroenterology*, 137:2179–80; author reply 2180–1.
- Ghaleb, A. M. and V. W. Yang
2017. Krppel-like factor 4 (klf4): What we currently know. *Gene*, 611:27–37.
- Glickman, J. N., A. Yang, A. Shahsafaei, F. McKeon, and R. D. Odze
2001. Expression of p53-related protein p63 in the gastrointestinal tract and in esophageal metaplastic and neoplastic disorders. *Human pathology*, 32:1157–1165.
- Glover, E. J., M. M. Leland, E. J. Dick, and G. B. Hubbard
2008. Gastroesophageal reflux disease in baboons (papio sp.): a new animal model. *Journal of medical primatology*, 37:18–25.
- Goldman, A., H. D. R. Chen, H. B. Roesly, K. A. Hill, M. E. Tome, B. Dvorak, H. Bernstein, and K. Dvorak
2011. Characterization of squamous esophageal cells resistant to bile acids at acidic ph: implication for barrett’s esophagus pathogenesis. *American journal of physiology. Gastrointestinal and liver physiology*, 300:G292–G302.
- Goodman, A. M., S. Kato, L. Bazhenova, S. P. Patel, G. M. Frampton, V. Miller, P. J. Stephens, G. A. Daniels, and R. Kurzrock
2017. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Molecular cancer therapeutics*, 16:2598–2608.

- Grainger, S., J. G. A. Savory, and D. Lohnes
2010. Cdx2 regulates patterning of the intestinal epithelium. *Developmental biology*, 339:155–165.
- Gregson, E. M., J. Bornschein, and R. C. Fitzgerald
2016. Genetic progression of barrett’s oesophagus to oesophageal adenocarcinoma. *British journal of cancer*, 115(4):403.
- Griffin, M. and E. C. Sweeney
1987. The relationship of endocrine cells, dysplasia and carcinoembryonic antigen in barrett’s mucosa to adenocarcinoma of the oesophagus. *Histopathology*, 11:53–62.
- Grün, D., A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden
2015. Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251.
- Grün, D., M. J. Muraro, J.-C. Boisset, K. Wiebrands, A. Lyubimova, G. Dharmadhikari, M. van den Born, J. van Es, E. Jansen, H. Clevers, et al.
2016. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19(2):266–277.
- Guillaumet-Adkins, A., G. Rodríguez-Esteban, E. Mereu, M. Mendez-Lago, D. A. Jaitin, A. Villanueva, A. Vidal, A. Martinez-Marti, E. Felip, A. Vivancos, et al.
2017. Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome biology*, 18(1):45.
- Guo, F., L. Li, J. Li, X. Wu, B. Hu, P. Zhu, L. Wen, and F. Tang
2017. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell research*, 27:967–988.
- Haber, A. L., M. Biton, N. Rogel, R. H. Herbst, K. Shekhar, C. Smillie, G. Burgin, T. M. Delorey, M. R. Howitt, Y. Katz, I. Tirosh, S. Beyaz, D. Dionne, M. Zhang, R. Raychowdhury, W. S. Garrett, O. Rozenblatt-Rosen, H. N. Shi, O. Yilmaz, R. J. Xavier, and A. Regev
2017. A single-cell survey of the small intestinal epithelium. *Nature*, 551:333–339.
- Haidry, R. J., M. A. Butt, J. M. Dunn, A. Gupta, G. Lipman, H. L. Smart, P. Bhandari, L. Smith, R. Willert, G. Fullarton, M. Di Pietro, C. Gordon, I. Penman, H. Barr, P. Patel, N. Kapoor, J. Hoare, R. Narayanasamy, Y. Ang, A. Veitch, K. Ragunath, M. Novelli, L. B. Lovat, and U. R. Registry
2015. Improvement over time in outcomes for patients undergoing endoscopic therapy for barrett’s oesophagus-related neoplasia: 6-year experience from the first 500 patients treated in the uk patient registry. *Gut*, 64:1192–1199.
- Hamada, H., C. Meno, D. Watanabe, and Y. Saijoh
2002. Establishment of vertebrate left–right asymmetry. *Nature Reviews Genetics*, 3(2):103.

- Han, K. Y., K.-T. Kim, J.-G. Joung, D.-S. Son, Y. J. Kim, A. Jo, H.-J. Jeon, H.-S. Moon, C. E. Yoo, W. Chung, H. H. Eum, S. Kim, H. K. Kim, J. E. Lee, M.-J. Ahn, H.-O. Lee, D. Park, and W.-Y. Park
2018a. Sidr: simultaneous isolation and parallel sequencing of genomic dna and total rna from single cells. *Genome research*, 28:75–87.
- Han, X., R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, D. Huang, Y. Xu, W. Huang, M. Jiang, X. Jiang, J. Mao, Y. Chen, C. Lu, J. Xie, Q. Fang, Y. Wang, R. Yue, T. Li, H. Huang, S. H. Orkin, G.-C. Yuan, M. Chen, and G. Guo
2018b. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172:1091–1107.e17.
- Hao, J., B. Liu, C. S. Yang, and X. Chen
2009. Gastroesophageal reflux leads to esophageal cancer in a surgical model with mice. *BMC gastroenterology*, 9:59.
- Harris-Johnson, K. S., E. T. Domyan, C. M. Vezina, and X. Sun
2009. beta-catenin promotes respiratory progenitor identity in mouse foregut. *Proceedings of the National Academy of Sciences of the United States of America*, 106:16287–16292.
- Hayward, J.
1961. The lower end of the oesophagus. *Thorax*, 16:36–41.
- Hicks, S. C., F. W. Townes, M. Teng, and R. A. Irizarry
2017. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*.
- Hofmann, A. F. and L. R. Hagey
2008. Bile acids: chemistry, pathochemistry, biology, pathobiology, and therapeutics. *Cellular and molecular life sciences : CMLS*, 65:2461–2483.
- Horn, C. C., B. A. Kimball, H. Wang, J. Kaus, S. Dienel, A. Nagy, G. R. Gathright, B. J. Yates, and P. L. R. Andrews
2013. Why can’t rodents vomit? a comparative behavioral, anatomical, and physiological study. *PloS one*, 8:e60537.
- Hou, Y., H. Guo, C. Cao, X. Li, B. Hu, P. Zhu, X. Wu, L. Wen, F. Tang, Y. Huang, and J. Peng
2016. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell research*, 26:304–319.
- Hou, Y., K. Wu, X. Shi, F. Li, L. Song, H. Wu, M. Dean, G. Li, S. Tsang, R. Jiang, X. Zhang, B. Li, G. Liu, N. Bedekar, N. Lu, G. Xie, H. Liang, L. Chang, T. Wang, J. Chen, Y. Li, X. Zhang, H. Yang, X. Xu, L. Wang, and J. Wang
2015. Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. *GigaScience*, 4:37. Original DateCompleted: 20150807.

- Houghton, J., C. Stoicov, S. Nomura, A. B. Rogers, J. Carlson, H. Li, X. Cai, J. G. Fox, J. R. Goldenring, and T. C. Wang
2004. Gastric cancer originating from bone marrow-derived cells. *Science (New York, N.Y.)*, 306:1568–1571.
- Hu, Y., K. Huang, Q. An, G. Du, G. Hu, J. Xue, X. Zhu, C.-Y. Wang, Z. Xue, and G. Fan
2016. Simultaneous profiling of transcriptome and dna methylome from a single cell. *Genome biology*, 17:88.
- Hu, Y., V. A. Williams, O. Gellersen, C. Jones, T. J. Watson, and J. H. Peters
2007. The pathogenesis of barretts esophagus: secondary bile acids upregulate intestinal differentiation factor cdx2 expression in esophageal cells. *Journal of Gastrointestinal Surgery*, 11(7):827–834.
- Huang, L., F. Ma, A. Chapman, S. Lu, and X. S. Xie
2015. Single-cell whole-genome amplification and sequencing: Methodology and applications. *Annual review of genomics and human genetics*, 16:79–102.
- Hutchinson, L., B. Stenstrom, D. Chen, B. Piperdi, S. Levey, S. Lyle, T. C. Wang, and J. Houghton
2010. Human barrett’s adenocarcinoma of the esophagus, associated myofibroblasts, and endothelium can arise from bone marrow-derived cells after allogeneic stem cell transplant. *Stem cells and development*, 20(1):11–17.
- Hyland, P. L., N. Hu, M. Rotunno, H. Su, C. Wang, L. Wang, R. M. Pfeiffer, B. Gherman, C. Giffen, C. Dykes, S. M. Dawsey, C. C. Abnet, K. M. Johnson, R. D. Acosta, P. E. Young, B. D. Cash, and P. R. Taylor
2014. Global changes in gene expression of barrett’s esophagus compared to normal squamous esophagus and gastric cardia tissues. *PloS one*, 9:e93219.
- Iinuma, H., R. Fukushima, T. Inaba, J. Tamura, T. Inoue, E. Ogawa, M. Horikawa, Y. Ikeda, N. Matsutani, K. Takeda, K. Yoshida, T. Tsunoda, T. Ikeda, Y. Nakamura, and K. Okinaga
2014. Phase i clinical study of multiple epitope peptide vaccine combined with chemoradiation therapy in esophageal cancer patients. *Journal of translational medicine*, 12:84.
- Islam, S., A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Linnerberg, and S. Linnarsson
2014. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11:163–166.
- Isomoto, H., A. Wang, Y. Mizuta, Y. Akazawa, K. Ohba, K. Omagari, M. Miyazaki, K. Murase, T. Hayashi, K. Inoue, et al.
2003. Elevated levels of chemokines in esophageal mucosa of patients with reflux esophagitis. *The American journal of gastroenterology*, 98(3):551–556.

- Jaitin, D. A., E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit
2014. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science (New York, N.Y.)*, 343:776–779.
- Jang, B. G., B. L. Lee, and W. H. Kim
2015. Intestinal stem cell markers in the intestinal metaplasia of stomach and barrett’s esophagus. *PloS one*, 10:e0127300.
- Janjigian, Y. Y., J. Bendell, E. Calvo, J. W. Kim, P. A. Ascierto, P. Sharma, P. A. Ott, K. Peltola, D. Jaeger, J. Evans, F. de Braud, I. Chau, C. T. Harbison, C. Dorange, M. Tschaika, and D. T. Le
2018. Checkmate-032 study: Efficacy and safety of nivolumab and nivolumab plus ipilimumab in patients with metastatic esophagogastric cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, P. JCO2017766212.
- Jayewickreme, C. D. and R. A. Shivdasani
2015. Control of stomach smooth muscle development and intestinal rotation by transcription factor barx1. *Developmental biology*, 405:21–32.
- Jenny, M., C. Uhl, C. Roche, I. Duluc, V. Guillermin, F. Guillemot, J. Jensen, M. Kedinger, and G. Gradwohl
2002. Neurogenin3 is differentially required for endocrine cell fate specification in the intestinal and gastric epithelium. *The EMBO journal*, 21:6338–6347.
- Ji, R.-R., S. D. Chasalow, L. Wang, O. Hamid, H. Schmidt, J. Cogswell, S. Alaparthi, D. Berman, M. Jure-Kunkel, N. O. Siemers, J. R. Jackson, and V. Shahabi
2012. An immune-active tumor microenvironment favors clinical response to ipilimumab. *Cancer immunology, immunotherapy : CII*, 61:1019–1031.
- Jiang, L., F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver
2011. Synthetic spike-in standards for rna-seq experiments. *Genome research*, 21:1543–1551.
- Jiang, M., H. Li, Y. Zhang, Y. Yang, R. Lu, K. Liu, S. Lin, X. Lan, H. Wang, H. Wu, J. Zhu, Z. Zhou, J. Xu, D.-K. Lee, L. Zhang, Y.-C. Lee, J. Yuan, J. A. Abrams, T. C. Wang, A. R. Sepulveda, Q. Wu, H. Chen, X. Sun, J. She, X. Chen, and J. Que
2017. Transitional basal cells at the squamous-columnar junction generate barrett’s oesophagus. *Nature*, 550:529–533.
- Kageyama, S., H. Ikeda, Y. Miyahara, N. Imai, M. Ishihara, K. Saito, S. Sugino, S. Ueda, T. Ishikawa, S. Kokura, H. Naota, K. Ohishi, T. Shiraishi, N. Inoue, M. Tanabe, T. Kidokoro, H. Yoshioka, D. Tomura, I. Nukaya, J. Mineno, K. Takesako, N. Katayama, and H. Shiku
2015. Adoptive transfer of mage-a4 t-cell receptor gene-transduced lymphocytes

- in patients with recurrent esophageal cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 21:2268–2277.
- Kageyama, S., H. Wada, K. Muro, Y. Niwa, S. Ueda, H. Miyata, S. Takiguchi, S. H. Sugino, Y. Miyahara, H. Ikeda, N. Imai, E. Sato, T. Yamada, M. Osako, M. Ohnishi, N. Harada, T. Hishida, Y. Doki, and H. Shiku
2013. Dose-dependent effects of ny-eso-1 protein vaccine complexed with cholesteryl pullulan (chp-ny-eso-1) on immune responses and survival benefits of esophageal cancer patients. *Journal of translational medicine*, 11:246.
- Kandulski, A. and P. Malfertheiner
2012. Gastroesophageal reflux disease from reflux episodes to mucosal inflammation. *Nature Reviews Gastroenterology and Hepatology*, 9(1):15–22.
- Kang, J. H., E. H. Lee, S. W. Park, and I. Y. Chung
2011. Muc5ac expression through bidirectional communication of notch and epidermal growth factor receptor pathways. *Journal of immunology (Baltimore, Md. : 1950)*, 187:222–229.
- Kang, L., X. Liu, Z. Gong, H. Zheng, J. Wang, Y. Li, H. Yang, J. Hardwick, H. Dai, R. T. P. Poon, N. P. Lee, M. Mao, Z. Peng, and R. Chen
2015. Genome-wide identification of rna editing in hepatocellular carcinoma. *Genomics*, 105:76–82.
- Kang, Y.-K., N. Boku, T. Satoh, M.-H. Ryu, Y. Chao, K. Kato, H. C. Chung, J.-S. Chen, K. Muro, W. K. Kang, K.-H. Yeh, T. Yoshikawa, S. C. Oh, L.-Y. Bai, T. Tamura, K.-W. Lee, Y. Hamamoto, J. G. Kim, K. Chin, D.-Y. Oh, K. Minashi, J. Y. Cho, M. Tsuda, and L.-T. Chen
2017. Nivolumab in patients with advanced gastric or gastro-oesophageal junction cancer refractory to, or intolerant of, at least two previous chemotherapy regimens (ono-4538-12, attraction-2): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet (London, England)*, 390:2461–2471.
- Kanigowska, P., Y. Shen, Y. Zheng, S. Rosser, and Y. Cai
2016. Smart dna fabrication using sound waves: applying acoustic dispensing technologies to synthetic biology. *Journal of laboratory automation*, 21(1):49–56.
- Karthaus, W. R., P. J. Iaquinta, J. Drost, A. Gracanin, R. van Boxtel, J. Wongvipat, C. M. Dowling, D. Gao, H. Begthel, N. Sachs, R. G. J. Vries, E. Cuppen, Y. Chen, C. L. Sawyers, and H. C. Clevers
2014. Identification of multipotent luminal progenitor cells in human prostate organoid cultures. *Cell*, 159:163–175.
- Kastelein, F., K. Biermann, E. W. Steyerberg, J. Verheij, M. Kalisvaart, L. H. J. Looijenga, H. A. Stoop, L. Walter, E. J. Kuipers, M. C. W. Spaander, M. J. Bruno, and P. study group
2013a. Aberrant p53 protein expression is associated with an increased risk of neoplastic progression in patients with barrett’s oesophagus. *Gut*, 62:1676–1683.

- Kastelein, F., M. C. W. Spaander, E. W. Steyerberg, K. Biermann, V. E. Valkhoff, E. J. Kuipers, M. J. Bruno, and P. S. Group
2013b. Proton pump inhibitors reduce the risk of neoplastic progression in patients with barrett's esophagus. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 11:382–388.
- 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 (Cambridge, England)*, 129:2619–2628.
- Kauer, W. K., J. H. Peters, T. R. DeMeester, H. Feussner, A. P. Ireland, H. J. Stein, and R. J. Siewert
1997. Composition and concentration of bile acid reflux into the esophagus of patients with gastroesophageal reflux disease. *Surgery*, 122:874–881.
- Kavanagh, M. E., K. E. O'Sullivan, C. O'Hanlon, J. N. O'Sullivan, J. Lysaght, and J. V. Reynolds
2014. The esophagitis to adenocarcinoma sequence; the role of inflammation. *Cancer letters*, 345:182–189.
- Kazumori, H., S. Ishihara, M. A. Rumi, Y. Kadowaki, and Y. Kinoshita
2006. Bile acids directly augment caudal related homeobox gene cdx2 expression in oesophageal keratinocytes in barretts epithelium. *Gut*, 55(1):16–25.
- Kazumori, H., S. Ishihara, Y. Takahashi, Y. Amano, and Y. Kinoshita
2011. Roles of kruppel-like factor 4 in oesophageal epithelial cells in barrett's epithelium development. *Gut*, 60:608–617.
- Kelsey, G., O. Stegle, and W. Reik
2017. Single-cell epigenomics: Recording the past and predicting the future. *Science (New York, N.Y.)*, 358:69–75.
- Kerkhof, M., H. van Dekken, E. W. Steyerberg, G. A. Meijer, A. H. Mulder, A. de Brune, A. Driessen, F. J. ten Kate, J. G. Kusters, E. J. Kuipers, P. D. Siersema, and C. study group
2007. Grading of dysplasia in barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology*, 50:920–927.
- Kim, B.-M., G. Buchner, I. Miletich, P. T. Sharpe, and R. A. Shivdasani
2005. The stomach mesenchymal transcription factor barx1 specifies gastric epithelial identity through inhibition of transient wnt signaling. *Developmental cell*, 8:611–622.
- Kim, T.-H., B.-M. Kim, J. Mao, S. Rowan, and R. A. Shivdasani
2011. Endodermal hedgehog signals modulate notch pathway activity in the developing digestive tract mesenchyme. *Development (Cambridge, England)*, 138:3225–3233.

- Kim, T.-H., F. Li, I. Ferreiro-Neira, L.-L. Ho, A. Luyten, K. Nalapareddy, H. Long, M. Verzi, and R. A. Shivdasani
2014. Broadly permissive intestinal chromatin underlies lateral inhibition and cell plasticity. *Nature*, 506:511–515.
- Kim, T.-H. and R. A. Shivdasani
2011. Notch signaling in stomach epithelial stem cell homeostasis. *The Journal of experimental medicine*, 208:677–688.
- Kim, T. S., C. L. Liu, M. Yassour, J. Holik, N. Friedman, S. Buratowski, and O. J. Rando
2010. Rna polymerase mapping during stress responses reveals widespread nonproductive transcription in yeast. *Genome biology*, 11(7):R75.
- Kiselev, V. Y., K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al.
2017. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483.
- Klein, A. M., L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner
2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161:1187–1201.
- Kobayashi, S., H. Yamada-Okabe, M. Suzuki, O. Natori, A. Kato, K. Matsubara, Y. Jau Chen, M. Yamazaki, S. Funahashi, K. Yoshida, E. Hashimoto, Y. Watanabe, H. Mutoh, M. Ashihara, C. Kato, T. Watanabe, T. Yoshikubo, N. Tamaoki, T. Ochiya, M. Kuroda, A. J. Levine, and T. Yamazaki
2012. Lgr5-positive colon cancer stem cells interconvert with drug-resistant lgr5-negative cells and are capable of tumor reconstitution. *Stem cells (Dayton, Ohio)*, 30:2631–2644.
- Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson
2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22:568–576.
- Koike, T. and S. Yasugi
1999. In vitro analysis of mesenchymal influences on the differentiation of stomach epithelial cells of the chicken embryo. *Differentiation; research in biological diversity*, 65:13–25.
- Kokubu, H., T. Ohtsuka, and R. Kageyama
2008. Mash1 is required for neuroendocrine cell development in the glandular stomach. *Genes to cells : devoted to molecular & cellular mechanisms*, 13:41–51.
- Kollmann, D., D. Ignatova, J. Jedamzik, Y.-T. Chang, G. Jomrich, A. Baierl, D. Kazakov, M. Michal, L. E. French, W. Hoetzenecker, T. Schatton, R. Asari,

- M. Preusser, M. Gnant, E. Guenova, and S. F. Schoppmann
2018. Pd-l1 expression is an independent predictor of favorable outcome in patients with localized esophageal adenocarcinoma. *Oncoimmunology*, 7:e1435226.
- Konturek, J. W., W. Bielanski, S. J. Konturek, J. Bogdal, and J. Oleksy
1989. Distribution and release of epidermal growth factor in man. *Gut*, 30:1194–1200.
- Krause, W. J.
2000. Brunner’s glands: a structural, histochemical and pathological profile. *Progress in histochemistry and cytochemistry*, 35:259–367.
- Krausova, M. and V. Korinek
2014. Wnt signaling in adult intestinal stem cells and cancer. *Cellular signalling*, 26:570–579.
- Krishnadath, K. K. and K. K. Wang
2015. Molecular pathogenesis of barrett esophagus: current evidence. *Gastroenterology clinics of North America*, 44:233–247.
- Krishnamurthy, S. and Y. Dayal
1995. Pancreatic metaplasia in barrett’s esophagus. an immunohistochemical study. *The American journal of surgical pathology*, 19:1172–1180.
- Kundaje, A., W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shores, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis
2015. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330.
- Kurian, L., I. Sancho-Martinez, E. Nivet, A. Aguirre, K. Moon, C. Pendaries, C. Volle-Challier, F. Bono, J.-M. Herbert, J. Pulecio, et al.
2013. Conversion of human fibroblasts to angioblast-like progenitor cells. *Nature methods*, 10(1):77–83.

- Kwon, M.-C., B.-K. Koo, Y.-Y. Kim, S.-H. Lee, N.-S. Kim, J.-H. Kim, and Y.-Y. Kong
2009. Essential role of cr6-interacting factor 1 (crif1) in e74-like factor 3 (elf3)-mediated intestinal development. *The Journal of biological chemistry*, 284:33634–33641.
- Laczka, D., F. Wang, F. B. Johnson, N. Jhala, A. Rosztczy, G. G. Ginsberg, G. W. Falk, A. K. Rustgi, and J. P. Lynch
2017. Modeling esophagitis using human three-dimensional organotypic culture system. *The American journal of pathology*, 187:1787–1799.
- Lamble, S., E. Batty, M. Attar, D. Buck, R. Bowden, G. Lunter, D. Crook, B. El-Fahmawi, and P. Piazza
2013. Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC biotechnology*, 13(1):104.
- Lao-Sirieix, P., A. Boussioutas, S. R. Kadri, M. O'Donovan, I. Debiram, M. Das, L. Harihar, and R. C. Fitzgerald
2009. Non-endoscopic screening biomarkers for barrett's oesophagus: from microarray analysis to the clinic. *Gut*, 58:1451–1459.
- Larsson, L. I., O. D. Madsen, P. Serup, J. Jonsson, and H. Edlund
1996. Pancreatic-duodenal homeobox 1 -role in gastric endocrine patterning. *Mechanisms of development*, 60:175–184.
- Larsson, L. I., L. St-Onge, D. M. Hougaard, B. Sosa-Pineda, and P. Gruss
1998. Pax 4 and 6 regulate gastrointestinal endocrine cell development. *Mechanisms of development*, 79:153–159.
- Lavery, D. L., P. Martinez, L. J. Gay, B. Cereser, M. R. Novelli, M. Rodriguez-Justo, S. L. Meijer, T. A. Graham, S. A. C. McDonald, N. A. Wright, and M. Jansen
2016. Evolution of oesophageal adenocarcinoma from metaplastic columnar epithelium without goblet cells in barrett's oesophagus. *Gut*, 65:907–913.
- Lavery, D. L., A. M. Nicholson, R. Poulsom, R. Jeffery, A. Hussain, L. J. Gay, J. A. Jankowski, S. S. Zeki, H. Barr, R. Harrison, et al.
2014. The stem cell organisation, and the proliferative and gene expression profile of barrett's epithelium, replicates pyloric-type gastric glands. *Gut*, Pp. gutjnl–2013.
- Lawrence, M. S., P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Corts, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler,

- K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz
2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499:214–218.
- Lee, C. S., N. Perreault, J. E. Brestelli, and K. H. Kaestner
2002. Neurogenin 3 is essential for the proper specification of gastric enteroendocrine cells and the maintenance of gastric epithelial cell identity. *Genes & development*, 16:1488–1497.
- Lee, E. R. and C. P. Leblond
1985. Dynamic histology of the antral epithelium in the mouse stomach: II. ultrastructure and renewal of isthmal cells. *The American journal of anatomy*, 172:205–224.
- Leedham, S. J., S. L. Preston, S. A. C. McDonald, G. Elia, P. Bhandari, D. Poller, R. Harrison, M. R. Novelli, J. A. Jankowski, and N. A. Wright
2008. Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human barrett’s oesophagus. *Gut*, 57:1041–1048.
- Levrat, M., R. Lambert, and G. Kirschbaum
1962. Esophagitis produced by reflux of duodenal contents in rats. *The American journal of digestive diseases*, 7:564–573.
- Li, H.
2011. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27:2987–2993.
- Li, H. and R. Durbin
2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25:1754–1760.
- Li, X., H. E. Francies, M. Secrier, J. Perner, A. Miremedi, N. Galeano-Dalmau, W. J. Barendt, L. Letchford, G. M. Leyden, E. K. Goffin, A. Barthorpe, H. Lightfoot, E. Chen, J. Gilbert, A. Noorani, G. Devonshire, L. Bower, A. Grantham, S. MacRae, N. Grehan, D. C. Wedge, R. C. Fitzgerald, and M. J. Garnett
2018. Organoid cultures recapitulate esophageal adenocarcinoma heterogeneity providing a model for clonality studies and precision therapeutics. *Nature communications*, 9:2983.
- Li, X., P. C. Galipeau, T. G. Paulson, C. A. Sanchez, J. Arnaudo, K. Liu, C. L. Sather, R. L. Kostadinov, R. D. Odze, M. K. Kuhner, C. C. Maley, S. G. Self, T. L. Vaughan, P. L. Blount, and B. J. Reid
2014. Temporal and spatial evolution of somatic chromosomal alterations: a case-cohort study of barrett’s esophagus. *Cancer prevention research (Philadelphia, Pa.)*, 7:114–127.

- Li, X., T. G. Paulson, P. C. Galipeau, C. A. Sanchez, K. Liu, M. K. Kuhner, C. C. Maley, S. G. Self, T. L. Vaughan, B. J. Reid, and P. L. Blount
2015. Assessment of esophageal adenocarcinoma risk using somatic chromosome alterations in longitudinal samples in barrett's esophagus. *Cancer prevention research (Philadelphia, Pa.)*, 8:845–856.
- Litingtung, Y., L. Lei, H. Westphal, and C. Chiang
1998. Sonic hedgehog is essential to foregut development. *Nature genetics*, 20:58–61.
- Long, J. D. and R. C. Orlando
1999. Esophageal submucosal glands: structure and function. *The American journal of gastroenterology*, 94(10):2818–2824.
- Lordick, F., Y.-K. Kang, H.-C. Chung, P. Salman, S. C. Oh, G. Bodoky, G. Kurteva, C. Volovat, V. M. Moiseyenko, V. Gorbunova, J. O. Park, A. Sawaki, I. Celik, H. Gtte, H. Meleznkov, M. Moehler, A. I. Onkologie, and E. Investigators
2013. Capecitabine and cisplatin with or without cetuximab for patients with previously untreated advanced gastric cancer (expand): a randomised, open-label phase 3 trial. *The Lancet. Oncology*, 14:490–499.
- Lörinc, E., L. Mellblom, and S. Öberg
2015. The immunophenotypic relationship between the submucosal gland unit, columnar metaplasia and squamous islands in the columnar-lined oesophagus. *Histopathology*, 67(6):792–798.
- Lörinc, E. and S. Öberg
2015. Hyperplasia of the submucosal glands of the columnar-lined oesophagus. *Histopathology*, 66(5):726–731.
- Luksza, M., N. Riaz, V. Makarov, V. P. Balachandran, M. D. Hellmann, A. Solovyov, N. A. Rizvi, T. Merghoub, A. J. Levine, T. A. Chan, J. D. Wolchok, and B. D. Greenbaum
2017. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*, 551:517–520.
- Luo, W., S. Li, B. Peng, Y. Ye, X. Deng, and K. Yao
2013. Embryonic stem cells markers sox2, oct4 and nanog expression and their correlations with epithelial-mesenchymal transition in nasopharyngeal carcinoma. *PloS one*, 8:e56324.
- Maag, J. L. V., O. M. Fisher, A. Levert-Mignon, D. C. Kaczorowski, M. L. Thomas, D. J. Hussey, D. I. Watson, A. Wettstein, Y. V. Bobryshev, M. Edwards, M. E. Dinger, and R. V. Lord
2017. Novel aberrations uncovered in barrett's esophagus and esophageal adenocarcinoma using whole transcriptome sequencing. *Molecular cancer research : MCR*, 15:1558–1569.

- Macaulay, I. C., W. Haerty, P. Kumar, Y. I. Li, T. X. Hu, M. J. Teng, M. Goolam, N. Saurat, P. Coupland, L. M. Shirley, M. Smith, N. Van der Aa, R. Banerjee, P. D. Ellis, M. A. Quail, H. P. Swerdlow, M. Zernicka-Goetz, F. J. Livesey, C. P. Ponting, and T. Voet
2015. G&t-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods*, 12:519–522.
- Macke, R. A., K. S. Nason, K.-i. Mukaisho, T. Hattori, T. Fujimura, S. Sasaki, K. Oyama, T. Miyashita, T. Ohta, K. Miwa, M. K. Gibson, A. Zaidi, U. Malhotra, A. Atasoy, T. Foxwell, and B. Jobe
2011. Barrett’s esophagus and animal models. *Annals of the New York Academy of Sciences*, 1232:392–400.
- Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll
2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161:1202–1214.
- Maley, C. C., P. C. Galipeau, J. C. Finley, V. J. Wongsurawat, X. Li, C. A. Sanchez, T. G. Paulson, P. L. Blount, R.-A. Risques, P. S. Rabinovitch, and B. J. Reid
2006. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature genetics*, 38:468–473.
- Maley, C. C., P. C. Galipeau, X. Li, C. A. Sanchez, T. G. Paulson, and B. J. Reid
2004. Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in barrett’s esophagus. *Cancer research*, 64:3414–3427.
- Mari, L., F. Milano, K. Parikh, D. Straub, V. Everts, K. K. Hoeben, P. Fockens, N. S. Buttar, and K. K. Krishnadath
2014. A psmad/cdx2 complex is essential for the intestinalization of epithelial metaplasia. *Cell reports*, 7:1197–1210.
- Martinez, P., M. R. Timmer, C. T. Lau, S. Calpe, M. D. C. Sancho-Serra, D. Straub, A.-M. Baker, S. L. Meijer, F. J. W. T. Kate, R. C. Mallant-Hent, A. H. J. Naber, A. H. A. M. van Oijen, L. C. Baak, P. Scholten, C. J. M. Bhmer, P. Fockens, J. J. G. H. M. Bergman, C. C. Maley, T. A. Graham, and K. K. Krishnadath
2016. Dynamic clonal equilibrium and predetermined cancer risk in barrett’s oesophagus. *Nature communications*, 7:12158.
- Maude, S. L., N. Frey, P. A. Shaw, R. Aplenc, D. M. Barrett, N. J. Bunin, A. Chew, V. E. Gonzalez, Z. Zheng, S. F. Lacey, Y. D. Mahnke, J. J. Melenhorst, S. R. Rheingold, A. Shen, D. T. Teachey, B. L. Levine, C. H. June, D. L. Porter, and S. A. Grupp
2014. Chimeric antigen receptor t cells for sustained remissions in leukemia. *The New England journal of medicine*, 371:1507–1517.

- May, R., T. E. Riehl, C. Hunt, S. M. Sureban, S. Anant, and C. W. Houchen
2008. Identification of a novel putative gastrointestinal stem cell and adenoma stem cell marker, doublecortin and cam kinase-like-1, following radiation injury and in adenomatous polyposis coli/multiple intestinal neoplasia mice. *Stem cells (Dayton, Ohio)*, 26:630–637.
- McCauley, H. A. and G. Guasch
2015. Three cheers for the goblet cell: maintaining homeostasis in mucosal epithelia. *Trends in molecular medicine*, 21:492–503.
- McCauley, H. A., C.-Y. Liu, A. C. Attia, K. A. Wikenheiser-Brokamp, Y. Zhang, J. A. Whitsett, and G. Guasch
2014. Tgf signaling inhibits goblet cell differentiation via spdef in conjunctival epithelium. *Development (Cambridge, England)*, 141:4628–4639.
- McDonald, S. A., D. Lavery, N. A. Wright, and M. Jansen
2015. Barrett oesophagus: lessons on its origins from the lesion itself. *Nature Reviews Gastroenterology & Hepatology*, 12(1):50–60.
- McDonald, S. A. C., L. C. Greaves, L. Gutierrez-Gonzalez, M. Rodriguez-Justo, M. Deheragoda, S. J. Leedham, R. W. Taylor, C. Y. Lee, S. L. Preston, M. Lovell, T. Hunt, G. Elia, D. Oukrif, R. Harrison, M. R. Novelli, I. Mitchell, D. L. Stoker, D. M. Turnbull, J. A. Z. Jankowski, and N. A. Wright
2008. Mechanisms of field cancerization in the human stomach: the expansion and spread of mutated gastric stem cells. *Gastroenterology*, 134:500–510.
- McPherson, A., A. Roth, E. Laks, T. Masud, A. Bashashati, A. W. Zhang, G. Ha, J. Biele, D. Yap, A. Wan, L. M. Prentice, J. Khattra, M. A. Smith, C. B. Nielsen, S. C. Mullaly, S. Kalloger, A. Karnezis, K. Shumansky, C. Siu, J. Rosner, H. L. Chan, J. Ho, N. Melnyk, J. Senz, W. Yang, R. Moore, A. J. Mungall, M. A. Marra, A. Bouchard-Ct, C. B. Gilks, D. G. Huntsman, J. N. McAlpine, S. Aparicio, and S. P. Shah
2016. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature genetics*, 48:758–767.
- Means, A. L., Y. Xu, A. Zhao, K. C. Ray, and G. Gu
2008. A ck19(creert) knockin mouse line allows for conditional dna recombination in epithelial cells in multiple endodermal organs. *Genesis (New York, N.Y. : 2000)*, 46:318–323.
- Metsuyanin, S., O. Harari-Steinberg, E. Buzhor, D. Omer, N. Pode-Shakked, H. Ben-Hur, R. Halperin, D. Schneider, and B. Dekel
2009. Expression of stem cell markers in the human fetal kidney. *PLoS One*, 4(8):e6709.
- Miyata, N., T. Azuma, S. Hozawa, H. Higuchi, A. Yokoyama, A. Kabashima,

- T. Igarashi, K. Saeki, and T. Hibi
2012. Transforming growth factor and ras/mek/erk signaling regulate the expression level of a novel tumor suppressor lefty. *Pancreas*, 41:745–752.
- Miyoshi, K., J. M. Shillingford, F. Le Provost, F. Gounari, R. Bronson, H. von Boehmer, M. M. Taketo, R. D. Cardiff, L. Hennighausen, and K. Khazaie
2002. Activation of beta -catenin signaling in differentiated mammary secretory cells induces transdifferentiation into epidermis and squamous metaplasias. *Proceedings of the National Academy of Sciences of the United States of America*, 99:219–224.
- Moffitt, J. R., J. Hao, D. Bambach-Mukku, T. Lu, C. Dulac, and X. Zhuang
2016. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proceedings of the National Academy of Sciences of the United States of America*, 113:14456–14461.
- Mooijman, D., S. S. Dey, J.-C. Boisset, N. Crosetto, and A. van Oudenaarden
2016. Single-cell 5hmc sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature biotechnology*, 34:852–856.
- Mori-Akiyama, Y., M. van den Born, J. H. van Es, S. R. Hamilton, H. P. Adams, J. Zhang, H. Clevers, and B. de Crombrughe
2007. Sox9 is required for the differentiation of paneth cells in the intestinal epithelium. *Gastroenterology*, 133:539–546.
- Morris, L. G. T., N. Riaz, A. Desrichard, Y. enbabaolu, A. A. Hakimi, V. Makarov, J. S. Reis-Filho, and T. A. Chan
2016. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, 7:10051–10063.
- Morrison, C., S. Pabla, J. M. Conroy, M. K. Nesline, S. T. Glenn, D. Dressman, A. Papanicolau-Sengos, B. Burgher, J. Andreas, V. Giamo, M. Qin, Y. Wang, F. L. Lenzo, A. Omilian, W. Bshara, M. Zibelman, P. Ghatalia, K. Dragnev, K. Shirai, K. G. Madden, L. J. Tafe, N. Shah, D. Kasuganti, L. de la Cruz-Merino, I. Araujo, Y. Saenger, M. Bogardus, M. Villalona-Calero, Z. Diaz, R. Day, M. Eisenberg, S. M. Anderson, I. Puzanov, L. Galluzzi, M. Gardner, and M. S. Ernstoff
2018. Predicting response to checkpoint inhibitors in melanoma beyond pd-l1 and mutational burden. *Journal for immunotherapy of cancer*, 6:32.
- Muraro, M. J., G. Dharmadhikari, D. Grn, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti, E. J. P. de Koning, and A. van Oudenaarden
2016. A single-cell transcriptome atlas of the human pancreas. *Cell systems*, 3:385–394.e3.
- Murillo, G. H., N. You, X. Su, W. Cui, M. P. Reilly, M. Li, K. Ning, and X. Cui
2016. Multigems: detection of snvs from multiple samples using model selection on high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 32:1486–1492.

- Mustata, R. C., T. Van Loy, A. Lefort, F. Libert, S. Strollo, G. Vassart, and M.-I. Garcia
2011. Lgr4 is required for paneth cell differentiation and maintenance of intestinal stem cells ex vivo. *EMBO reports*, 12:558–564.
- Muoz, J., D. E. Stange, A. G. Schepers, M. van de Wetering, B.-K. Koo, S. Itzkovitz, R. Volckmann, K. S. Kung, J. Koster, S. Radulescu, K. Myant, R. Versteeg, O. J. Sansom, J. H. van Es, N. Barker, A. van Oudenaarden, S. Mohammed, A. J. R. Heck, and H. Clevers
2012. The lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. *The EMBO journal*, 31:3079–3091.
- Nakamura, S., K. Hayashi, Y. Imaoka, Y. Kitamura, Y. Akazawa, K. Tabata, R. Groen, T. Tsuchiya, N. Yamasaki, T. Nagayasu, and J. Fukuoka
2017. Intratumoral heterogeneity of programmed cell death ligand-1 expression is common in lung cancer. *PloS one*, 12:e0186192.
- Narita, N., M. Bielinska, and D. B. Wilson
1997. Wild-type endoderm abrogates the ventral developmental defects associated with gata-4 deficiency in the mouse. *Developmental biology*, 189:270–274.
- Nehra, D., P. Howell, C. P. Williams, J. K. Pye, and J. Beynon
1999. Toxic bile acids in gastro-oesophageal reflux disease: influence of gastric acidity. *Gut*, 44:598–602.
- Nishikura, K.
2010. Functions and regulation of rna editing by adar deaminases. *Annual review of biochemistry*, 79:321–349.
- 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. *Experimental cell research*, 316:452–465.
- Nomura, S., H. Esumi, C. Job, and S. S. Tan
1998. Lineage and clonal development of gastric glands. *Developmental biology*, 204:124–135.
- Nones, K., N. Waddell, N. Wayte, A.-M. Patch, P. Bailey, F. Newell, O. Holmes, J. L. Fink, M. C. J. Quinn, Y. H. Tang, G. Lampe, K. Quek, K. A. Loffler, S. Manning, S. Idrisoglu, D. Miller, Q. Xu, N. Waddell, P. J. Wilson, T. J. C. Bruxner, A. N. Christ, I. Harliwong, C. Nourse, E. Nourbakhsh, M. Anderson, S. Kazakoff, C. Leonard, S. Wood, P. T. Simpson, L. E. Reid, L. Krause, D. J. Hussey, D. I. Watson, R. V. Lord, D. Nancarrow, W. A. Phillips, D. Gotley, B. M. Smithers, D. C. Whiteman, N. K. Hayward, P. J. Campbell, J. V. Pearson, S. M. Grimmond, and A. P. Barbour
2014. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature communications*, 5:5224.

- Nosho, K., Y. Baba, N. Tanaka, K. Shima, M. Hayashi, J. A. Meyerhardt, E. Giovannucci, G. Dranoff, C. S. Fuchs, and S. Ogino
2010. Tumour-infiltrating t-cell subsets, molecular changes in colorectal cancer, and prognosis: cohort study and literature review. *The Journal of pathology*, 222:350–366.
- O'Brien, T. D., P. Jia, J. Xia, U. Saxena, H. Jin, H. Vuong, P. Kim, Q. Wang, M. J. Aryee, M. Mino-Kenudson, J. A. Engelman, L. P. Le, A. J. Iafrate, R. S. Heist, W. Pao, and Z. Zhao
2015. Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer. *Methods (San Diego, Calif.)*, 83:118–127.
- Offman, J., F. Pesola, and P. Sasieni
2018. Trends and projections in adenocarcinoma and squamous cell carcinoma of the oesophagus in england from 1971 to 2037. *British journal of cancer*, 118:1391–1398.
- Ohigashi, Y., M. Sho, Y. Yamada, Y. Tsurui, K. Hamada, N. Ikeda, T. Mizuno, R. Yoriki, H. Kashizuka, K. Yane, F. Tsushima, N. Otsuki, H. Yagita, M. Azuma, and Y. Nakajima
2005. Clinical significance of programmed death-1 ligand-1 and programmed death-1 ligand-2 expression in human esophageal cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 11:2947–2953.
- Ohtsu, A., J. A. Ajani, Y.-X. Bai, Y.-J. Bang, H.-C. Chung, H.-M. Pan, T. Sahmoud, L. Shen, K.-H. Yeh, K. Chin, K. Muro, Y. H. Kim, D. Ferry, N. C. Tebbutt, S.-E. Al-Batran, H. Smith, C. Costantini, S. Rizvi, D. Lebwohl, and E. Van Cutsem
2013. Everolimus for previously treated advanced gastric cancer: results of the randomized, double-blind, phase iii granite-1 study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 31:3935–3943.
- Ohtsu, A., M. A. Shah, E. Van Cutsem, S. Y. Rha, A. Sawaki, S. R. Park, H. Y. Lim, Y. Yamada, J. Wu, B. Langer, M. Starnawski, and Y.-K. Kang
2011. Bevacizumab in combination with chemotherapy as first-line therapy in advanced gastric cancer: a randomized, double-blind, placebo-controlled phase iii study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 29:3968–3976.
- Orlando, R. C.
2010. The integrity of the esophageal mucosa. balance between offensive and defensive mechanisms. *Best practice & research. Clinical gastroenterology*, 24:873–882.
- Orloff, M., C. Peterson, X. He, S. Ganapathi, B. Heald, Y.-r. Yang, G. Bebek, T. Romigh, J. H. Song, W. Wu, S. David, Y. Cheng, S. J. Meltzer, and C. Eng
2011. Germline mutations in *msr1*, *ascc1*, and *cthr1* in patients with barrett esophagus and esophageal adenocarcinoma. *JAMA*, 306:410–419.

- Ormsby, A. H., M. F. Vaezi, J. E. Richter, J. R. Goldblum, T. W. Rice, G. W. Falk, and T. L. Gramlich
2000. Cytokeratin immunoreactivity patterns in the diagnosis of short-segment barrett's esophagus. *Gastroenterology*, 119(3):683–690.
- Ortmann, C. A., D. G. Kent, J. Nangalia, Y. Silber, D. C. Wedge, J. Grinfeld, E. J. Baxter, C. E. Massie, E. Papaemmanuil, S. Menon, A. L. Godfrey, D. Dimitropoulou, P. Guglielmelli, B. Bellosillo, C. Besses, K. Dhner, C. N. Harrison, G. S. Vassiliou, A. Vannucchi, P. J. Campbell, and A. R. Green
2015. Effect of mutation order on myeloproliferative neoplasms. *The New England journal of medicine*, 372:601–612.
- Ottesen, E. A., J. W. Hong, S. R. Quake, and J. R. Leadbetter
2006. Microfluidic digital pcr enables multigene analysis of individual environmental bacteria. *Science (New York, N.Y.)*, 314:1464–1467.
- Owen, R. P., M. J. White, D. T. Severson, B. Braden, A. Bailey, R. Goldin, L. M. Wang, N. Maynard, A. Green, P. Piazza, D. Buck, M. Middleton, C. Ponting, B. Schuster-Boeckler, and X. Lu
2018. Single cell rna-seq reveals profound transcriptional similarity between barretts esophagus and esophageal glands. *bioRxiv*.
- Palles, C., L. Chegwiddden, X. Li, J. M. Findlay, G. Farnham, F. Castro Giner, M. P. Peppelenbosch, M. Kovac, C. L. Adams, H. Prenen, S. Briggs, R. Harrison, S. Sanders, D. MacDonald, C. Haigh, A. Tucker, S. Love, M. Nanji, J. de Caestecker, D. Ferry, B. Rathbone, J. Hapeshi, H. Barr, P. Moayyedi, P. Watson, B. Zietek, N. Maroo, L. Gay, T. Underwood, L. Boulter, H. McMurtry, D. Monk, P. Patel, K. Ragunath, D. Al Dulaimi, I. Murray, K. Koss, A. Veitch, N. Trudgill, C. Nwokolo, B. Rembacken, P. Atherfold, E. Green, Y. Ang, E. J. Kuipers, W. Chow, S. Paterson, S. Kadri, I. Beales, C. Grimley, P. Mullins, C. Beckett, M. Farrant, A. Dixon, S. Kelly, M. Johnson, S. Wajed, A. Dhar, E. Sawyer, R. Roylance, L. Onstad, M. D. Gammon, D. A. Corley, N. J. Shaheen, N. C. Bird, L. J. Hardie, B. J. Reid, W. Ye, G. Liu, Y. Romero, L. Bernstein, A. H. Wu, A. G. Casson, R. Fitzgerald, D. C. Whiteman, H. A. Risch, D. M. Levine, T. L. Vaughan, A. P. Verhaar, J. van den Brande, E. L. Toxopeus, M. C. Spaander, B. P. L. Wijnhoven, L. J. W. van der Laan, K. Krishnadath, C. Wijmenga, G. Trynka, R. McManus, J. V. Reynolds, J. O'Sullivan, P. MacMathuna, S. A. McGarrigle, D. Kelleher, S. Vermeire, I. Cleynen, R. Bisschops, I. Tomlinson, and J. Jankowski
2015. Polymorphisms near *tbx5* and *gdf7* are associated with increased risk for barrett's esophagus. *Gastroenterology*, 148:367–378.
- Pardoll, D. M.
2012. The blockade of immune checkpoints in cancer immunotherapy. *Nature reviews. Cancer*, 12:252–264.
- Parekh, S., C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann

2016. The impact of amplification on differential expression analyses by rna-seq. *Scientific reports*, 6:25533.
- Patel, A. P., I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suv, A. Regev, and B. E. Bernstein
2014. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, 344:1396–1401.
- Paterson, A. L., P. Lao-Sirieix, M. O’Donovan, I. Debiram-Beecham, M. di Pietro, A. Miremadi, S. E. Attwood, F. M. Walter, P. D. Sasieni, R. C. Fitzgerald, BEST, and B. study groups
2017. Range of pathologies diagnosed using a minimally invasive capsule sponge to evaluate patients with reflux symptoms. *Histopathology*, 70:203–210.
- Peranzoni, E., J. Lemoine, L. Vimeux, V. Feuillet, S. Barrin, C. Kantari-Mimoun, N. Bercovici, M. Gurin, J. Biton, H. Ouakrim, F. Rgnier, A. Lupo, M. Alifano, D. Damotte, and E. Donnadieu
2018. Macrophages impede cd8 t cells from reaching tumor cells and limit the efficacy of anti-pd-1 treatment. *Proceedings of the National Academy of Sciences of the United States of America*, 115:E4041–E4050.
- Peterson, V. M., K. X. Zhang, N. Kumar, J. Wong, L. Li, D. C. Wilson, R. Moore, T. K. McClanahan, S. Sadekova, and J. A. Klappenbach
2017. Multiplexed quantification of proteins and transcripts in single cells. *Nature biotechnology*, 35:936–939.
- Picelli, S., Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg
2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096.
- Picelli, S., O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg
2014. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171.
- Pinto, D., A. Gregorieff, H. Begthel, and H. Clevers
2003. Canonical wnt signals are essential for homeostasis of the intestinal epithelium. *Genes & development*, 17:1709–1713.
- Poplin, R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks
2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.

- Pott, S.
2017. Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *eLife*, 6.
- Potten, C. S., L. Kovacs, and E. Hamilton
1974. Continuous labelling studies on mouse skin and intestine. *Cell and tissue kinetics*, 7:271–283.
- Powell, A., M. Parappilly, Y.-H. Lo, T. Lantz, and N. Shroyer
2016. Examination of the development of gastrointestinal brunners glands: Stem cells and differentiation factors. *The FASEB Journal*, 30(1_supplement):748.1–748.1.
- Qiao, X. T., J. W. Ziel, W. McKimpson, B. B. Madison, A. Todisco, J. L. Merchant, L. C. Samuelson, and D. L. Gumucio
2007. Prospective identification of a multilineage progenitor in murine stomach epithelium. *Gastroenterology*, 133:1989–1998.
- Qiu, S., S. Luo, O. Evgrafov, R. Li, G. P. Schroth, P. Levitt, J. A. Knowles, and K. Wang
2012. Single-neuron rna-seq: technical feasibility and reproducibility. *Frontiers in genetics*, 3:124.
- Quante, M., G. Bhagat, J. A. Abrams, F. Marache, P. Good, M. D. Lee, Y. Lee, R. Friedman, S. Asfaha, Z. Dubeykovskaya, et al.
2012. Bile acid and inflammation activate gastric cardia stem cells in a mouse model of barrett-like metaplasia. *Cancer cell*, 21(1):36–51.
- Quante, M., F. Marrache, J. R. Goldenring, and T. C. Wang
2010. Tff2 mrna transcript expression marks a gland progenitor cell of the gastric oxyntic mucosa. *Gastroenterology*, 139:2018–2027.e2.
- Que, J., T. Okubo, J. R. Goldenring, K.-T. Nam, R. Kurotani, E. E. Morrissey, O. Taranova, L. H. Pevny, and B. L. M. Hogan
2007. Multiple dose-dependent roles for sox2 in the patterning and differentiation of anterior foregut endoderm. *Development (Cambridge, England)*, 134:2521–2531.
- Raggi, M., R. Langer, M. Feith, H. Friess, M. Schauer, and J. Theisen
2010. Successful evaluation of a new animal model using mice for esophageal adenocarcinoma. *Langenbeck's archives of surgery*, 395:347–350.
- Raghoebir, L., E. R. M. Bakker, J. C. Mills, S. Swagemakers, M. B.-v. Kempen, A. B.-d. Munck, S. Driegen, D. Meijer, F. Grosveld, D. Tibboel, R. Smits, and R. J. Rottier
2012. Sox2 redirects the developmental fate of the intestinal epithelium toward a premature gastric phenotype. *Journal of molecular cell biology*, 4:377–385.

- Ramsey, V. G., J. M. Doherty, C. C. Chen, T. S. Stappenbeck, S. F. Konieczny, and J. C. Mills
2007. The maturation of mucus-secreting gastric epithelial progenitors into digestive-enzyme secreting zymogenic cells requires mist1. *Development (Cambridge, England)*, 134:211–222.
- Ramskld, D., S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg
2012. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nature biotechnology*, 30:777–782.
- Raymond, D. P., C. W. Seder, C. D. Wright, M. J. Magee, A. S. Kosinski, S. D. Cassivi, E. L. Grogan, S. H. Blackmon, M. S. Allen, B. J. Park, W. R. Burfeind, A. C. Chang, M. M. DeCamp, D. W. Wormuth, F. G. Fernandez, and B. D. Kozower
2016. Predictors of major morbidity or mortality after resection for esophageal cancer: A society of thoracic surgeons general thoracic surgery database risk adjustment model. *The Annals of thoracic surgery*, 102:207–214.
- Reddy, K. M., J. I. Chang, J. M. Shi, and B. U. Wu
2016. Risk of gastric cancer among patients with intestinal metaplasia of the stomach in a us integrated health care system. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 14:1420–1425.
- Reenan, R. A.
2005. Molecular determinants and guided evolution of species-specific rna editing. *Nature*, 434:409–413.
- Reveiller, M., S. Ghatak, L. Toia, I. Kalatskaya, L. Stein, M. D’Souza, Z. Zhou, S. Bandla, W. E. Gooding, T. E. Godfrey, and J. H. Peters
2012. Bile exposure inhibits expression of squamous differentiation genes in human esophageal epithelial cells. *Annals of surgery*, 255:1113–1120.
- Rishniw, M., P. Rodriguez, J. Que, Z. D. Burke, D. Tosh, H. Chen, and X. Chen
2011. Molecular aspects of esophageal development. *Annals of the New York Academy of Sciences*, 1232:309–315.
- Rizvi, N. A., M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher, and T. A. Chan
2015. Cancer immunology. mutational landscape determines sensitivity to pd-1 blockade in non-small cell lung cancer. *Science (New York, N.Y.)*, 348:124–128.
- Rizzetto, S., A. A. Eltahla, P. Lin, R. Bull, A. R. Lloyd, J. W. Ho, V. Venturi, and F. Luciani
2017. Impact of sequencing depth and read length on single cell rna sequencing data of t cells. *Scientific Reports*, 7(1):12781.

- 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.
- Rodriguez, P., S. Da Silva, L. Oxburgh, F. Wang, B. L. M. Hogan, and J. Que
2010. Bmp signaling in the development of the mouse esophagus and forestomach. *Development (Cambridge, England)*, 137:4171–4176.
- Roman, S., A. Ptr, A. Thpot, A. Hautefeuille, J.-Y. Scoazec, F. Mion, and P. Hainaut
2007. Downregulation of p63 upon exposure to bile salts and acid in normal and cancer esophageal cells in culture. *American journal of physiology. Gastrointestinal and liver physiology*, 293:G45–G53.
- Rosenberg, A. B., C. M. Roco, R. A. Muscat, A. Kuchina, P. Sample, Z. Yao, L. T. Graybuck, D. J. Peeler, S. Mukherjee, W. Chen, S. H. Pun, D. L. Sellers, B. Tasic, and G. Seelig
2018. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (New York, N.Y.)*, 360:176–182.
- Ross-Innes, C. S., J. Becq, A. Warren, R. K. Cheetham, H. Northen, M. O’Donovan, S. Malhotra, M. di Pietro, S. Ivakhno, M. He, et al.
2015a. Whole-genome sequencing provides new insights into the clonal architecture of barrett’s esophagus and esophageal adenocarcinoma. *Nature genetics*, 47(9):1038–1046.
- Ross-Innes, C. S., I. Debiram-Beecham, M. O’Donovan, E. Walker, S. Varghese, P. Lao-Sirieix, L. Lovat, M. Griffin, K. Ragunath, R. Haidry, S. S. Sami, P. Kaye, M. Novelli, B. Disep, R. Ostler, B. Aigret, B. V. North, P. Bhandari, A. Haycock, D. Morris, S. Attwood, A. Dhar, C. Rees, M. D. D. Rutter, P. D. Sasieni, R. C. Fitzgerald, and B. S. Group
2015b. Evaluation of a minimally invasive cell sampling device coupled with assessment of trefoil factor 3 expression for diagnosing barrett’s esophagus: a multi-center case-control study. *PLoS medicine*, 12:e1001780.
- Rotem, A., O. Ram, N. Shores, R. A. Sperling, A. Goren, D. A. Weitz, and B. E. Bernstein
2015. Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology*, 33:1165–1172.
- Rumiato, E., E. Boldrin, S. Malacrida, S. Realdon, M. Fassan, T. Morbin, G. Battaglia, A. Amadori, M. Rugge, and D. Saggioro
2017. Detection of genetic alterations in cfdna as a possible strategy to monitor the neoplastic progression of barrett’s esophagus. *Translational research : the journal of laboratory and clinical medicine*, 190:16–24.e1.
- Saito, T., H. Wada, M. Yamasaki, H. Miyata, H. Nishikawa, E. Sato, S. Kageyama, H. Shiku, M. Mori, and Y. Doki
2014. High expression of mage-a4 and mhc class i antigens in tumor cells and

- induction of mage-a4 immune responses are prognostic markers of chp-mage-a4 cancer vaccine. *Vaccine*, 32:5901–5907.
- Saliba, A.-E., A. J. Westermann, S. A. Gorski, and J. Vogel
2014. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860.
- Salit, M. L., L. Jiang, F. Schlesinger, B. Oliver, C. Davis, Y. Zhang, R. Li, and T. R. Gingeras
2011. Synthetic spike-in standards for rna-seq experiments. *Genome Research*, 21(9).
- Samuels-Lev, Y., D. J. O'Connor, D. Bergamaschi, G. Trigiante, J. K. Hsieh, S. Zhong, I. Campargue, L. Naumovski, T. Crook, and X. Lu
2001. Aspp proteins specifically stimulate the apoptotic function of p53. *Molecular cell*, 8:781–794.
- Sandmann, S., A. O. de Graaf, M. Karimi, B. A. van der Reijden, E. Hellström-Lindberg, J. H. Jansen, and M. Dugas
2017. Evaluating variant calling tools for non-matched next-generation sequencing data. *Scientific reports*, 7:43169.
- Satija, R., J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev
2015. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33:495–502.
- Sato, T., D. E. Stange, M. Ferrante, R. G. J. Vries, J. H. Van Es, S. Van den Brink, W. J. Van Houdt, A. Pronk, J. Van Gorp, P. D. Siersema, and H. Clevers
2011. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and barrett's epithelium. *Gastroenterology*, 141:1762–1772.
- Sato, T., R. G. Vries, H. J. Snippert, M. van de Wetering, N. Barker, D. E. Stange, J. H. van Es, A. Abo, P. Kujala, P. J. Peters, and H. Clevers
2009. Single lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature*, 459:262–265.
- Schneider, C. A., W. S. Rasband, and K. W. Eliceiri
2012. Nih image to imagej: 25 years of image analysis. *Nature methods*, 9:671–675.
- Secrier, M., X. Li, N. de Silva, M. D. Eldridge, G. Contino, J. Bornschein, S. MacRae, N. Grehan, M. O'Donovan, A. Miremedi, T.-P. Yang, L. Bower, H. Chettouh, J. Crawte, N. Galeano-Dalmau, A. Grabowska, J. Saunders, T. Underwood, N. Waddell, A. P. Barbour, B. Nutzinger, A. Achilleos, P. A. W. Edwards, A. G. Lynch, S. Tavar, R. C. Fitzgerald, O. C. Clinical, and M. S. O. Consortium
2016. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature genetics*, 48:1131–1141.

- Segerstolpe, ., A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andrasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, D. M. Smith, M. Kasper, C. mml, and R. Sandberg
2016. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, 24:593–607.
- Severson, D., R. Owen, M. White, X. Lu, and B. Schuster-Böckler
2018. Bearscc determines robustness of single-cell clusters using simulated technical replicates. *Nature communications*, 9(1):1187.
- Shapiro, J., J. J. B. van Lanschot, M. C. C. M. Hulshof, P. van Hagen, M. I. van Berge Henegouwen, B. P. L. Wijnhoven, H. W. M. van Laarhoven, G. A. P. Nieuwenhuijzen, G. A. P. Hospers, J. J. Bonenkamp, M. A. Cuesta, R. J. B. Blaisse, O. R. C. Busch, F. J. W. Ten Kate, G.-J. M. Creemers, C. J. A. Punt, J. T. M. Plukker, H. M. W. Verheul, E. J. S. Bilgen, H. van Dekken, M. J. C. van der Sangen, T. Rozema, K. Biermann, J. C. Beukema, A. H. M. Piet, C. M. van Rij, J. G. Reinders, H. W. Tilanus, E. W. Steyerberg, A. van der Gaast, and C. study group
2015. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (cross): long-term results of a randomised controlled trial. *The Lancet. Oncology*, 16:1090–1098.
- Shawihdi, M., E. Thompson, N. Kapoor, G. Powell, R. P. Sturgess, N. Stern, M. Roughton, M. G. Pearson, and K. Bodger
2014. Variation in gastroscopy rate in english general practice and outcome for oesophagogastric cancer: retrospective analysis of hospital episode statistics. *Gut*, 63:250–261.
- Shearer, C., J. Going, L. Neilson, C. Mackay, and R. Stuart
2005. Cytokeratin 7 and 20 expression in intestinal metaplasia of the distal oesophagus: relationship to gastro-oesophageal reflux disease. *Histopathology*, 47(3):268–275.
- Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin
2001. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29:308–311.
- Silberg, D. G., J. Sullivan, E. Kang, G. P. Swain, J. Moffett, N. J. Sund, S. D. Sackett, and K. H. Kaestner
2002. Cdx2 ectopic expression induces gastric intestinal metaplasia in transgenic mice. *Gastroenterology*, 122:689–696.
- Smallwood, S. A., H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey
2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11:817–820.

- Smith, K. A., E. Nol, I. Thurlings, H. Rehmann, S. Chocron, and J. Bakkers
2011. Bmp and nodal independently regulate lefty1 expression to maintain unilateral nodal activity during left-right axis specification in zebrafish. *PLoS genetics*, 7:e1002289.
- Solaymani-Dodaran, M., R. Logan, J. West, T. Card, and C. Coupland
2004. Risk of oesophageal cancer in barretts oesophagus and gastro-oesophageal reflux. *Gut*, 53(8):1070–1074.
- Sottocornola, R., C. Royer, V. Vives, L. Tordella, S. Zhong, Y. Wang, I. Ratnayaka, M. Shipman, A. Cheung, C. Gaston-Massuet, P. Ferretti, Z. Molnr, and X. Lu
2010. Aspp2 binds par-3 and controls the polarity and proliferation of neural progenitors during cns development. *Developmental cell*, 19:126–137.
- Souza, R. F., K. Shewmake, L. S. Terada, and S. J. Spechler
2002. Acid exposure activates the mitogen-activated protein kinase pathways in barrett’s esophagus. *Gastroenterology*, 122:299–307.
- Stachler, M. D., N. D. Camarda, C. Deitrick, A. Kim, A. T. Agoston, R. D. Odze, J. L. Hornick, A. Nag, A. R. Thorner, M. Ducar, A. Noffsinger, R. H. Lash, M. Redston, S. L. Carter, J. M. Davison, and A. J. Bass
2018. Detection of mutations in barrett’s esophagus before progression to high-grade dysplasia or adenocarcinoma. *Gastroenterology*, 155:156–167.
- Stachler, M. D., A. Taylor-Weiner, S. Peng, A. McKenna, A. T. Agoston, R. D. Odze, J. M. Davison, K. S. Nason, M. Loda, I. Leshchiner, et al.
2015. Paired exome analysis of barrett’s esophagus and adenocarcinoma. *Nature genetics*, 47(9):1047–1055.
- Stange, D. E., B.-K. Koo, M. Huch, G. Sibbel, O. Basak, A. Lyubimova, P. Kujala, S. Bartfeld, J. Koster, J. H. Geahlen, P. J. Peters, J. H. van Es, M. van de Wetering, J. C. Mills, and H. Clevers
2013. Differentiated troy+ chief cells act as reserve stem cells to generate all lineages of the stomach epithelium. *Cell*, 155:357–368.
- Stevens, T. J., D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O’Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sans, M. G. S. Palayret, B. Lehner, L. Di Croce, A. Wutz, B. Hendrich, D. Klennerman, and E. D. Laue
2017. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544:59–64.
- Stoeckius, M., C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert
2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14:865–868.

- Street, K., D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit
2018. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19:477.
- Svensson, V., K. N. Natarajan, L.-H. Ly, R. J. Miragaia, C. Labalette, I. C. Macaulay, A. Cvejic, and S. A. Teichmann
2017. Power analysis of single-cell rna-sequencing experiments. *Nature methods*, 14:381–387.
- Takaoka, K., H. Nishimura, and H. Hamada
2017. Both nodal signalling and stochasticity select for prospective distal visceral endoderm in mouse embryos. *Nature communications*, 8:1492.
- Tamagawa, Y., N. Ishimura, G. Uno, M. Aimi, N. Oshima, T. Yuki, S. Sato, S. Ishihara, and Y. Kinoshita
2016. Bile acids induce delta-like 1 expression via cdx2-dependent pathway in the development of barrett’s esophagus. *Laboratory investigation; a journal of technical methods and pathology*, 96:325–337.
- Tamagawa, Y., N. Ishimura, G. Uno, T. Yuki, H. Kazumori, S. Ishihara, Y. Amano, and Y. Kinoshita
2012. Notch signaling pathway and cdx2 expression in the development of barrett’s esophagus. *Laboratory investigation; a journal of technical methods and pathology*, 92:896–909.
- Tandonnet, S. and T. T. Torres
2017. Traditional versus 3 rna-seq in a non-model species. *Genomics data*, 11:9–16.
- Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani
2009. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6:377–382.
- Theodosiou, N. A. and C. J. Tabin
2005. Sox9 and nkx2.5 determine the pyloric sphincter epithelium under the control of bmp signaling. *Developmental biology*, 279:481–490.
- Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kuttyavin, B. Lajoie, B.-K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E.

- Crawford, and J. A. Stamatoyannopoulos
2012. The accessible chromatin landscape of the human genome. *Nature*, 489:75–82.
- Tirosh, I., B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.-R. Lin, O. Cohen, P. Shah, D. Lu, A. S. Genshaft, T. K. Hughes, C. G. K. Ziegler, S. W. Kazer, A. Gaillard, K. E. Kolb, A.-C. Villani, C. M. Johannesssen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty, D. T. Frederick, J. Jan-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, and L. A. Garraway
2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science (New York, N.Y.)*, 352:189–196.
- Tordella, L., S. Koch, V. Salter, A. Pagotto, J. B. Doondeea, S. M. Feller, I. Ratnayaka, S. Zhong, R. D. Goldin, G. Lozano, F. D. McKeon, M. Tavassoli, F. Fritzsche, G. F. Huber, M. Rssle, H. Moch, and X. Lu
2013. Aspp2 suppresses squamous cell carcinoma via rela/p65-mediated repression of p63. *Proceedings of the National Academy of Sciences of the United States of America*, 110:17969–17974.
- Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn
2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32:381–386.
- Treutlein, B., D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake
2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, 509:371–375.
- Udager, A. M., A. Prakash, D. A. Saenz, M. Schinke, T. Moriguchi, P. Y. Jay, K.-C. Lim, J. D. Engel, and D. L. Gumucio
2014. Proper development of the outer longitudinal smooth muscle of the mouse pylorus requires nkx2-5 and gata3. *Gastroenterology*, 146:157–165.e10.
- Uhlén, M., L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, . Sivertsson, C. Kampf, E. Sjstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigartyo, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontn
2015. Proteomics. tissue-based map of the human proteome. *Science (New York, N.Y.)*, 347:1260419.
- Underwood, T. J., M. F. Derouet, M. J. White, F. Noble, K. A. Moutasim, E. Smith, P. A. Drew, G. J. Thomas, J. N. Primrose, and J. P. Blaydes
2010. A comparison of primary oesophageal squamous epithelial cells with het-1a in organotypic culture. *Biology of the cell*, 102:635–644.

- Underwood, T. J., A. L. Hayden, M. Derouet, E. Garcia, F. Noble, M. J. White, S. Thirdborough, A. Mead, N. Clemons, M. Mellone, C. Uzoho, J. N. Primrose, J. P. Blaydes, and G. J. Thomas
2015. Cancer-associated fibroblasts predict poor outcome and promote periostin-dependent invasion in oesophageal adenocarcinoma. *The Journal of pathology*, 235:466–477.
- Urschel, J. D. and H. Vasan
2003. A meta-analysis of randomized controlled trials that compared neoadjuvant chemoradiation and surgery to surgery alone for resectable esophageal cancer. *American journal of surgery*, 185:538–543.
- Vaezi, M. F. and J. E. Richter
1996. Role of acid and duodenogastroesophageal reflux in gastroesophageal reflux disease. *Gastroenterology*, 111:1192–1199.
- Vakil, N., S. V. Van Zanten, P. Kahrilas, J. Dent, and R. Jones
2006. The montreal definition and classification of gastroesophageal reflux disease: a global evidence-based consensus. *The American journal of gastroenterology*, 101(8):1900–1920.
- van de Wetering, M., H. E. Francies, J. M. Francis, G. Bounova, F. Iorio, A. Pronk, W. van Houdt, J. van Gorp, A. Taylor-Weiner, L. Kester, A. McLaren-Douglas, J. Blokker, S. Jaksani, S. Bartfeld, R. Volckman, P. van Sluis, V. S. W. Li, S. Seepo, C. Sekhar Pedamallu, K. Cibulskis, S. L. Carter, A. McKenna, M. S. Lawrence, L. Lichtenstein, C. Stewart, J. Koster, R. Versteeg, A. van Oudenaarden, J. Saez-Rodriguez, R. G. J. Vries, G. Getz, L. Wessels, M. R. Stratton, U. McDermott, M. Meyerson, M. J. Garnett, and H. Clevers
2015. Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell*, 161:933–945.
- Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo
2013. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43:11.10.1–11.1033.
- van der Flier, L. G., A. Haegebarth, D. E. Stange, M. van de Wetering, and H. Clevers
2009. Olfm4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. *Gastroenterology*, 137:15–17.
- Van Der Maaten, L.
2014. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research*, 15(1):3221–3245.
- Van Gelder, R. N., M. E. von Zastrow, A. Yool, W. C. Dement, J. D. Barchas, and J. H. Eberwine
1990. Amplified rna synthesized from limited quantities of heterogeneous cdna.

Proceedings of the National Academy of Sciences of the United States of America, 87:1663–1667.

Van Nieuwenhove, Y., H. Destordeur, and G. Willems

2001. Spatial distribution and cell kinetics of the glands in the human esophageal mucosa. *European journal of morphology*, 39(3):163–168.

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, . 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 (Cambridge, England)*, 139:488–497.

Varghese, S., R. Newton, C. S. Ross-Innes, P. Lao-Sirieix, K. K. Krishnadath, M. O'Donovan, M. Novelli, L. Wernisch, J. Bergman, and R. C. Fitzgerald

2015. Analysis of dysplasia in patients with barrett's esophagus based on expression pattern of 90 genes. *Gastroenterology*, 149:1511–1518.e5.

Vega, M. E., V. Giroux, M. Natsuizaka, M. Liu, A. J. Klein-Szanto, D. B. Stairs, H. Nakagawa, K. K. Wang, T. C. Wang, J. P. Lynch, and A. K. Rustgi

2014. Inhibition of notch signaling enhances transdifferentiation of the esophageal squamous epithelium towards a barrett's-like metaplasia via klf4. *Cell cycle (Georgetown, Tex.)*, 13:3857–3866.

Verzi, M. P., A. H. Khan, S. Ito, and R. A. Shivdasani

2008. Transcription factor foxq1 controls mucin gene expression and granule content in mouse stomach surface mucous cells. *Gastroenterology*, 135:591–600.

Vives, V., J. Su, S. Zhong, I. Ratnayaka, E. Slee, R. Goldin, and X. Lu

2006. Aspp2 is a haploinsufficient tumor suppressor that cooperates with p53 to suppress tumor growth. *Genes & development*, 20:1262–1267.

von Furstenberg, R. J., J. Li, C. Stolarчук, R. Feder, A. Campbell, L. Kruger, L. M. Gonzalez, A. T. Blikslager, D. M. Cardona, S. J. McCall, S. J. Henning, and K. S. Garman

2017. Porcine esophageal submucosal gland culture model shows capacity for proliferation and differentiation. *Cellular and molecular gastroenterology and hepatology*, 4:385–404.

Waddell, T., I. Chau, D. Cunningham, D. Gonzalez, A. F. C. Okines, A. Frances, C. Okines, A. Wotherspoon, C. Saffery, G. Middleton, J. Wadsley, D. Ferry, W. Mansoor, T. Crosby, F. Coxon, D. Smith, J. Waters, T. Iveson, S. Falk, S. Slater, C. Peckitt, and Y. Barbachano

2013. Epirubicin, oxaliplatin, and capecitabine with or without panitumumab for patients with previously untreated advanced oesophagogastric cancer (real3): a randomised, open-label phase 3 trial. *The Lancet. Oncology*, 14:481–489.

- Wan, H., K. H. Kaestner, S.-L. Ang, M. Ikegami, F. D. Finkelman, M. T. Stahlman, P. C. Fulkerson, M. E. Rothenberg, and J. A. Whitsett
2004. Foxa2 regulates alveolarization and goblet cell hyperplasia. *Development (Cambridge, England)*, 131:953–964.
- Wang, D. H., N. J. Clemons, T. Miyashita, A. J. Dupuy, W. Zhang, A. Szczepny, I. M. Corcoran-Schwartz, D. L. Wilburn, E. A. Montgomery, J. S. Wang, N. A. Jenkins, N. A. Copeland, J. W. Harmon, W. A. Phillips, and D. N. Watkins
2010. Aberrant epithelial-mesenchymal hedgehog signaling characterizes barrett’s metaplasia. *Gastroenterology*, 138:1810–1822.
- Wang, L., S. Wang, and W. Li
2012. Rseqc: quality control of rna-seq experiments. *Bioinformatics*, 28(16):2184–2185.
- Wang, R. N., J. Green, Z. Wang, Y. Deng, M. Qiao, M. Peabody, Q. Zhang, J. Ye, Z. Yan, S. Denduluri, O. Idowu, M. Li, C. Shen, A. Hu, R. C. Haydon, R. Kang, J. Mok, M. J. Lee, H. L. Luu, and L. L. Shi
2014a. Bone morphogenetic protein (bmp) signaling in development and human diseases. *Genes & diseases*, 1:87–105.
- Wang, X., H. Ouyang, Y. Yamamoto, P. A. Kumar, T. S. Wei, R. Dagher, M. Vincent, X. Lu, A. M. Bellizzi, K. Y. Ho, et al.
2011. Residual embryonic cells as precursors of a barrett’s-like metaplasia. *Cell*, 145(7):1023–1035.
- Wang, X., S. Yang, X. Zhao, H. Guo, X. Ling, L. Wang, C. Fan, J. Yu, and S. Zhou
2014b. Oct3 and sox2 promote the transformation of barrett’s esophagus to adenocarcinoma by regulating the formation of tumor stem cells. *Oncology reports*, 31:1745–1753.
- Wapenaar, M. C., A. J. Monsuur, J. Poell, R. vant 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.
- Washimi, K., T. Yokose, M. Yamashita, T. Kageyama, K. Suzuki, M. Yoshihara, Y. Miyagi, H. Hayashi, and S. Tsuji
2012. Specific expression of human intelectin-1 in malignant pleural mesothelioma and gastrointestinal goblet cells. *PloS one*, 7:e39889.
- Watari, J., K. Moriichi, H. Tanabe, R. Sato, M. Fujiya, H. Miwa, K. M. Das, and Y. Kohgo
2009. Differences in genetic instability and cellular phenotype among barrett’s, cardiac, and gastric intestinal metaplasia in a japanese population with helicobacter pylori. *Histopathology*, 55:261–269.

- Weaver, J. M. J., C. S. Ross-Innes, N. Shannon, A. G. Lynch, T. Forshew, M. Barbera, M. Murtaza, C.-A. J. Ong, P. Lao-Sirieix, M. J. Dunning, L. Smith, M. L. Smith, C. L. Anderson, B. Carvalho, M. O'Donovan, T. J. Underwood, A. P. May, N. Grehan, R. Hardwick, J. Davies, A. Oloumi, S. Aparicio, C. Caldas, M. D. Eldridge, P. A. W. Edwards, N. Rosenfeld, S. Tavar, R. C. Fitzgerald, and O. consortium
2014. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nature genetics*, 46:837–843.
- White, N. M., M. Gabril, G. Ejeckam, M. Mathews, J. Fardy, F. Kamel, J. Dor, and G. M. Yousef
2008. Barrett's esophagus and cardiac intestinal metaplasia: two conditions within the same spectrum. *Canadian journal of gastroenterology = Journal canadien de gastroenterologie*, 22:369–375.
- Whiteman, H. J., M. E. Weeks, S. E. Dowen, S. Barry, J. F. Timms, N. R. Lemoine, and T. Crnogorac-Jurcevic
2007. The role of s100p in the invasion of pancreatic cancer cells is mediated through cytoskeletal changes and regulation of cathepsin d. *Cancer research*, 67:8633–8642.
- Wickham, H.
2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wolchok, J. D., V. Chiarion-Sileni, R. Gonzalez, P. Rutkowski, J.-J. Grob, C. L. Cowey, C. D. Lao, J. Wagstaff, D. Schadendorf, P. F. Ferrucci, M. Smylie, R. Dummer, A. Hill, D. Hogg, J. Haanen, M. S. Carlino, O. Bechter, M. Maio, I. Marquez-Rodas, M. Guidoboni, G. McArthur, C. Lebb, P. A. Ascierto, G. V. Long, J. Cebon, J. Sosman, M. A. Postow, M. K. Callahan, D. Walker, L. Rollin, R. Bhore, F. S. Hodi, and J. Larkin
2017. Overall survival with combined nivolumab and ipilimumab in advanced melanoma. *The New England journal of medicine*, 377:1345–1356.
- Wong, D. J., T. G. Paulson, L. J. Prevo, P. C. Galipeau, G. Longton, P. L. Blount, and B. J. Reid
2001. p16(ink4a) lesions are common, early abnormalities that undergo clonal expansion in barrett's metaplastic epithelium. *Cancer research*, 61:8284–8289.
- Wong, N. A. C. S., J. Wilding, S. Bartlett, Y. Liu, B. F. Warren, J. Piris, N. Maynard, R. Marshall, and W. F. Bodmer
2005. Cdx1 is an important molecular mediator of barrett's metaplasia. *Proceedings of the National Academy of Sciences of the United States of America*, 102:7565–7570.
- Wu, C., Y. Zhu, J. Jiang, J. Zhao, X.-G. Zhang, and N. Xu
2006. Immunohistochemical localization of programmed death-1 ligand-1 (pd-l1) in gastric carcinoma and its clinical significance. *Acta histochemica*, 108:19–24.

- Xiang, C. C., E. Mezey, M. Chen, S. Key, L. Ma, and M. J. Brownstein
2004. Using dsp, a reversible cross-linker, to fix tissue sections for immunostaining, microdissection and expression profiling. *Nucleic acids research*, 32(22):e185–e185.
- Yamamoto, Y., X. Wang, D. Bertrand, F. Kern, T. Zhang, M. Duleba, S. Srivastava, C. C. Khor, Y. Hu, L. H. Wilson, et al.
2016. Mutational spectrum of barretts stem cells suggests paths to initiation of a precancerous lesion. *Nature communications*, 7:10380.
- Ye, D. Z. and K. H. Kaestner
2009. Foxa1 and foxa2 control the differentiation of goblet and enteroendocrine l- and d-cells in mice. *Gastroenterology*, 137:2052–2062.
- Young, M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack
2010. Gene ontology analysis for rna-seq: accounting for selection bias. *Genome biology*, 11:R14.
- Yuan, G.-C., L. Cai, M. Elowitz, T. Enver, G. Fan, G. Guo, R. Irizarry, P. Kharchenko, J. Kim, S. Orkin, J. Quackenbush, A. Saadatpour, T. Schroeder, R. Shivdasani, and I. Tirosh
2017. Challenges and emerging directions in single-cell analysis. *Genome biology*, 18:84.
- Zafar, H., Y. Wang, L. Nakhleh, N. Navin, and K. Chen
2016. Monovar: single-nucleotide variant detection in single cells. *Nature methods*, 13:505–507.
- Zagari, R. M., L. H. Eusebi, S. Rabitti, L. Cristofori, A. Vestito, N. Pagano, and F. Bazzoli
2016. Prevalence of upper gastrointestinal endoscopic findings in the community: A systematic review of studies in unselected samples of subjects. *Journal of gastroenterology and hepatology*, 31(9):1527–1538.
- Zeisel, A., A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Jurus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson
2015. Brain structure. cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science (New York, N.Y.)*, 347:1138–1142.
- Zhou, G., Y.-G. Sun, H.-B. Wang, W.-Q. Wang, X.-W. Wang, and D.-C. Fang
2009. Acid and bile salt up-regulate bmp4 expression in human esophageal epithelium cells. *Scandinavian journal of gastroenterology*, 44:926–932.
- Zhu, C., Y. Gao, H. Guo, B. Xia, J. Song, X. Wu, H. Zeng, K. Kee, F. Tang, and C. Yi
2017. Single-cell 5-formylcytosine landscapes of mammalian early embryos and escs at single-base resolution. *Cell stem cell*, 20:720–731.e5.

- Ziegenhain, C., B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard
2017. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65:631–643.e4.
- Zorn, A. M. and J. M. Wells
2009. Vertebrate endoderm development and organ formation. *Annual review of cell and developmental biology*, 25:221–251.
- Zou, H., J. R. Molina, J. J. Harrington, N. K. Osborn, K. K. Klatt, Y. Romero, L. J. Burgart, and D. A. Ahlquist
2005. Aberrant methylation of secreted frizzled-related protein genes in esophageal adenocarcinoma and barrett’s esophagus. *International journal of cancer*, 116:584–591.
- Zkny, J. and D. Duboule
1999. Hox genes and the making of sphincters. *Nature*, 401:761–762.