Ageing the Barrett's Lesion: A Study of Evolution to Oesophageal Adenocarcinoma

Richard John Hackett

Centre for Genomics and Computational Biology,

Bart's Cancer Institute,

Bart's and The London School of Medicine and Dentistry,

Queen Mary, University of London

Primary Supervisor: Dr Stuart A C McDonald

Secondary Supervisor: Professor Trevor A Graham

Thesis submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy

September 2022

Statement of Originality

I, Richard John Hackett, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Signature redacted

Date: 19th September 2022

Details of collaboration and publications:

Chapter 6

Dr Freddie Whiting assisted me in the design for a cell line experiment that tests the sequencing protocol presented in this thesis. He completed the cell culture on my behalf but everything else thereafter I completed.

Chapter 7

Some work undertaken by Dr James Evans and Dr Emanuela Carlotti was added to the phenotypes chapter of this thesis. I had assisted in tissue collection and staining but the purpose of adding the work was for contextual reasons as detailed in the text.

Endoscopic Biopsies

Dr James Evans commenced the lab's fresh frozen and fixed formalin paraffin embedded endoscopic tissue collection in February 2015, samples taken and initially processed by him have been used in this thesis. Endoscopic tissue collection duty was passed to me at my commencement in September 2017. Historical fresh frozen biopsies from University College Hospital taken by Professor Laurence Lovat and his team were relocated to the Queen Mary University London Cryostorage Facility under a Material Transfer Agreement in November 2018, some of these samples were used in this thesis.

Histopathology

All endoscopic biopsies were processed and sectioned by BCI Pathology Services personnel including Dr George Elia, Ms. Nadia Rahman and Ms. Irineja Cubela.

Sequencing

All libraries submitted for sequencing were handled by the Genome Centre at Bart's Cancer Institute and latterly, following their relocation, The Blizzard Institute.

This thesis was proofread by Dr Stuart McDonald

Work arising from this thesis to date has contributed to the following original research publication:

Evans J A, Carlotti E, Lin M L, **Hackett R J**, Haughey M J, Passman A M, Dunn L, Elia G, Porter R J, McLean M H, Hughes F, ChinAleong J, Woodland P, Preston S L, Griffin S M, Lovat L, Rodriguez-Justo M, Huang W, Wright N A, Jansen M, McDonald S. A. C. Clonal Transitions and Phenotypic Evolution in Barrett's Esophagus. *Gastroenterology*. 2022;162:1197-1209. d*oi: 10.1053/j.gastro.2021.12.271*

Abstract

Barrett's oesophagus (BO), a metaplasia affecting the distal oesophagus, is the only known precursor for oesophageal adenocarcinoma (OAC) which, despite advances in healthcare, continues to carry a dire prognosis. The emergence of BO has been shown to have polyclonal origins forming a mosaic of distinct geno-phenotypic clones across the space. However, the precise transformative cellular population, rate of clonal expansion and subsequent neighbouring clonal dynamics that promote a benign or malignant course remain unknown.

Recent work has demonstrated the ability to use the sequence of epigenetic methylation marks, which are somatically inherited at mitosis, as a marker of clonal ancestry, mutational ordering in progression and to determine mitotic age in colorectal adenomas and cancer. Characterisation of the dynamics that underpin stem cell behaviour, which ultimately are the precursor cells for the cancer phenotype, can also be inferred by mapping methylation patterns at high resolution across epithelium. These techniques are transferrable to BO, another epithelial, glandular and clonal disease, and form the primary modus operandi of this thesis.

This thesis adds to the debate regarding whether there is a particular dwell time to cancer; whether there is a mitotic age that predicts it; where and how the BO lesion expands at inception and varies over its natural history; the turnover of glandular phenotypes and whether they follow a linear or direct evolution to cancer; and how molecular diversity can be a proxy marker for cancer risk.

I have designed a novel targeted allele specific methylation sequencing (ASM-Seq) array that utilises modern next generation sequencing technology to significantly enhance resolution and coverage over previous studies. Furthermore, my protocol is the first of its type in a cytosine deaminated DNA template that incorporates unique molecular identifiers (UMIs) in efforts to reduce confounders in sequencing data.

Targeted ASM-Seq has the potential to reveal the intricate tissue dynamics not just of BO but any disease characterised by a clonal organisation and ancestry. Ultimately, this understanding will assist in better targeting of surveillance, clinical resources and therapies to patients deemed at risk of OAC.

Acknowledgements

To all those with whom I have crossed paths with on this journey in academia I give a heartfelt and sincere thank you. There are far too many to name who have all played some part in the formulation of this piece of work. Without your guidance, kindness, teaching and laughter this thesis would not have come to fruition.

A special thanks must go to Stuart who has been there all along to listen to my trials and tribulations and always managing to point me back on path. He is an excellent researcher and teacher but more so a true friend through this experience.

Most of all my family of which, over the duration of this evolutionary work has clonally expanded itself with the arrival of my most beautiful children Vivian and Russell who made any difficult day melt away with a single smile.

And finally, my darling wife, Ashley, who has provided more to me than she will ever know and I will spend my lifetime giving back to her all that she deserves.

Table of Contents

Table of Figures

Supplemental Figure

Table of Tables

Supplemental Tables

Abbreviations

5caC	5-carboxycytosine
5mC	5-methylcytosine
ABCB1	ATP binding cassette subfamily B member 1
Abr-US	Abridged-adapter universal sequence
AFAP1-AS1	Actin filament associated protein 1 antisense RNA 1
ANKRD2	Ankyrin repeat domain 2
APC	APC regulator of WNT signalling pathway
APOBEC	Apolopioprotein B mRNA Editing Catalytic Polypetide-like
ARID1A	AT-rich interaction domain 1A
ASM-Seq	Allele specific methylation sequencing
AT	Adenine-thymine
bDNA	Bisulfite converted deoxyribonucleic acid
BFB	Breakage fusion bridge
BGN	Biglycan
BM	Basement membrane
BMP	Bone morphogenic protein
BO	Barrett's oesophagus
CAMK2B	Calcium/calmodulin dependent protein kinase 2 beta
Car4+	Carbonic anhydrase 4
CCO	Cytochrome *c* oxidase
cDNA	Copy deoxyribonucleic acid
CDKN2A	Cyclin dependent kinase inhibitor 2A
CDX2	Caudal type homeobox 2
CIMP	CpG island methylator phenotype
CIN	Chromosomal instability
CMG	Compact mucous gland
CNA	Copy number alteration
CNTNAP5	Contactin associated protein family member 5
CpG	Cytosine-phosphate-guanine dinucleotide
CRBP1	Retinol binding protein 1
CSRP3	Cysteine and glycine rich protein 3
DBCAT	Database of CpG islands and analytical tools
deUIDs	Dual ended unique molecular identifiers

DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTPs	Deoxynucleotide triphosphates
dsDNA	Double-stranded deoxyribonucleic acid
DTT	Dithiothreitol
dUTP	Deoxyuridine triphosphate
Dys	Dysplasia
eDNA	Enzymatically converted deoxyribonucleic acid
EET	Endoscopic eradication therapies
EM	Enzymatic Methyl-seq
EMR	Endoscopic mucosal resection
ERBB2	Erb-B2 receptor tyrosine kinase
FFPE	Fixed formalin paraffin embedded
FHIT	Fragile histidine triad
GC	Guanine-cytosine
gDNA	Genomic deoxyribonucleic acid
GI	Gastrointestinal
GOJ	Gastro-oesophageal junction
GORD	Gastro-oesophageal reflux disease
GTEx	Genotype Tissue Expression project
H&E	Haematoxylin & eosin
HGD	High grade dysplasia
Hh	Hedgehog
HP-US	Hairpin universal sequence
HPA	Human Protein Atlas
HPP1	Hyperpigmentation, progressive 1
HRP	Horse radish peroxidase
i5	3rd round primers containing an Illumina® index from the 500 series
i7	3rd round primers containing an Illumina® index from the 700 series
IBL	Interpapillary basal layer
IFD	Indefinite for dysplasia
IHC	Immunohistochemistry
IM	Intestinal metaplasia
IMC	Intramucosal cancer

ITH	Intratumoral heterogeneity
KPI	Key performance indicators
Krt	Human keratin
LA-US	Linear adapter universal sequence
LCM	Laser capture microdissection
LEFTY1	Left right determination factor 1
LGD	Low grade dysplasia
LGR5	Leucine-rich repeat-containing G-protein coupled receptor 5
LOH	Loss of heterozygosity
LOS	Lower oesophageal sphincter
L-sUS	Linear shared universal sequence
MGMT	O-6- methylguanine DNA methyltransferase
MIP	Methylation-indifferent primers
MLE	Multi-layered epithelium
mRNA	Messenger ribonucleic acid
MSP	Methylation-specific primers
mtDNA	Mitochondrial deoxyribonucleic acid
MTA	Material transfer agreement
MUC	Mucin
MUC2	Mucin 2
MUC5AC	Mucin 5AC
MUC6	Mucin 6
MYLK3	Myosin light chain kinase 3
MYO18B	Myosin XVIII B
MYOD1	Myogenic differentiation 1
NDBO	Non-dysplastic Barrett's oesophagus
NGS	Next generation sequencing
NKX2-5	NK2 homeobox 5
NPPB	Natriuretic peptide B
NRT	No reverse transcription control
NTC	No template control
OAC	Oesophageal adenocarcinoma
OCT	Optimal cutting temperature
OSCC	Oesophageal squamous cell carcinoma
PBL	Papillary basal layer

PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PE	Paired end
Ptch1	Patched1
PWD	Pairwise distance
PXDNL	Peroxidasin like
RFA	Radio-frequency ablation
RLH	The Royal London Hospital, London, UK
RNA	Ribonucleic acid
RT	Reverse transcription
RT-PCR	Reverse transcription polymerase chain reaction
SBK2	SH3 domain binding kinase family member 2
SBK3	SH3 domain binding kinase family member 3
SBS	Sequencing by synthesis
SCA	Somatic chromosomal alteration
SCJ	Squamo-columnar junction
SCN5A	Sodium voltage-gated channel alpha subunit 5
SDH	Succinate dehydrogenase
SEER	Surveillance, epidemiology and end results
seUMI	Single ended unique molecular identifier
SMAD4	SMAD family member 4
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4
SNP	Single nucleotide polymorphism
SOX9	SRY-box transcription factor 9
SPTB	Spectrin beta, erythrocytic
ssbDNA	Single-stranded bisulfite converted deoxyribonucleic acid
ssDNA	Single-stranded deoxyribonucleic acid
SWGS	Shallow whole genome sequencing
TBC	Transitional basal cells
TCGA	The cancer genome atlas
TET2	Tet methylcytosine dioxygenase 2
TF	Transcription factor
TFF	Trefoil factor
TGFβ	Transforming growth factor β
TNNI3	Troponin I, cardiac type

TNNT2	Troponin T2, cardiac type
TP	Timepoint
TP53	Tumour protein 53
TPM	Transcripts per million
TSG	Tumour suppressor gene
UCH	University College Hospital, London, UK
UMI	Unique molecular identifiers
US	Universal Sequence
UTR	Unstranslated region
VAF	Variant allele frequency
VEGF	Vascular endothelial growth factor A
WES	Whole exome sequencing
WGBS	Whole genome bisulfite sequencing
WGS	Whole genome sequencing
WWOX	WW Domain-containing oxidoreductase

1 Introduction

1.1 The normal and metaplastic human oesophageal epithelium

1.1.1 The macroscopic oesophagus

The oesophagus is a flattened muscular tube connecting the hypopharynx to the cardia of the stomach. It serves little to no secretory function aside from mucus lubrication of food as this passes along its length under active peristalsis. It's length ranges from 18-26cm in adults from the upper oesophageal sphincter to the gastro-oesophageal junction (GOJ) usually located 1cm inferior to the diaphragmatic hiatus through which the oesophagus passes[1]. A region of high pressure, approximately 4 cm in length, exists at this transition point called the lower oesophageal sphincter (LOS), this is encircled by and anchored to the crural aspect of the diaphragm via the phrenoesophageal ligament[2] (**Fig. 1.1**). Of importance, the LOS in normal physiology works to reduce reflux of caustic gastric contents into the distal oesophagus which is less adapted to withstand such an insult. The LOS competence relies on the resting tone of the intrinsic circular oesophageal muscle, augmented by a flap-valve effect of sling fibres of the stomach under fundal pressure and extrinsic oesophageal compression by the crural diaphragm during respiration[2].

The wall of the oesophagus comprises the mucosa at the luminal surface, submucosa, muscularis propria and adventitia as the outermost layer (**Fig. 1.1**). The oesophagus is unique in the gastrointestinal (GI) tract with a transitional gradient proximally to distally of striated to smooth muscle fibres.

Macroscopically at endoscopy the oesophageal mucosal lining appears pale, smooth and non-glandular which contrasts well against the darker pink glandular gastric epithelium. The transition between the oesophageal and gastric mucosa is known as the Z-line or squamo-columnar junction (SCJ). This irregular line coincides with the GOJ, which is defined as the proximal margin of the gastric folds and/or end of the oesophageal mucosal palisade vessels. (**Fig. 1.2)**

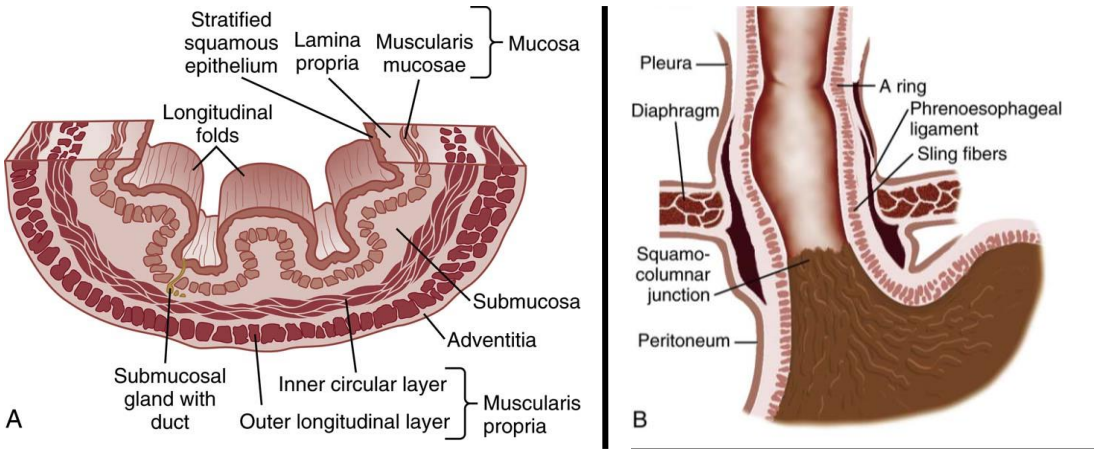

Figure 1.1: *Normal macroscopic structure of the human oesophagus. (A) Cross section of the anatomical layers forming the wall of the oesophagus. (B) Anatomy of the gastro-oesophageal junction where the lower oesophageal sphincter is located and its spatial orientation to the diaphragm, phrenoesophageal ligament and squamo-columnar junction. Taken from Part V, Esophagus of Sleisenger and Fordtran's Gastrointestinal and Liver Disease: Pathophysiology, Diagnosis, Management* [1].

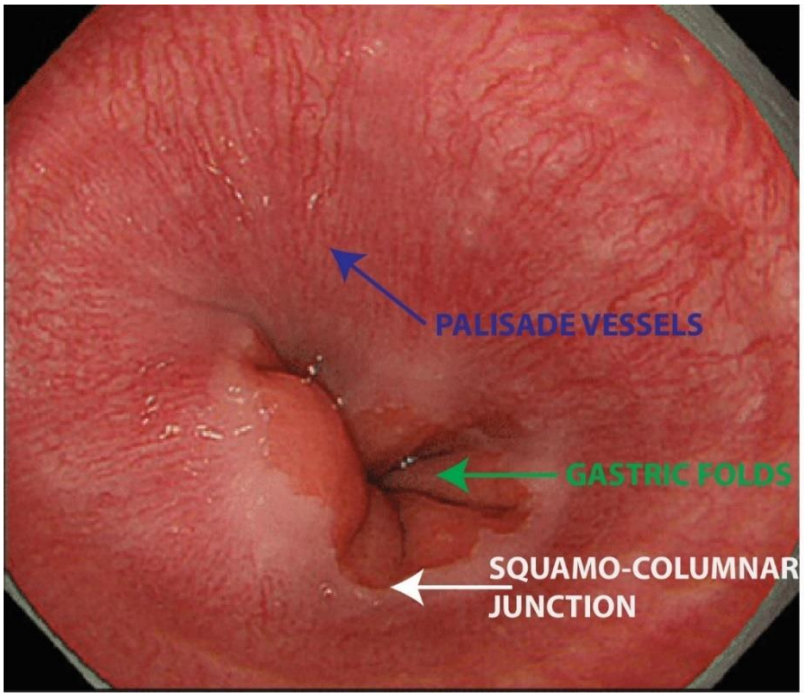

Figure 1.2: *Normal endoscopic appearances of the gastro-oesophageal junction demonstrating the transition from squamous epithelium of the anatomical oesophagus to glandular (columnar) epithelium of the stomach. The junction is dynamic with dilatation and contraction under peristalsis during endoscopy. Its true anatomical location is identified at the top of the gastric folds and/or end of the distal palisade vessels of the oesophagus which are both not impacted by proximal migration of the SCJ or the presence of a hiatus hernia that can lead to erroneous reporting and diagnosis of oesophageal pathology*[3].

1.1.2 The microscopic oesophagus

The normal oesophageal mucosa is a thick, non-keratinised flat stratified squamous epithelium. It comprises of 10-20 layers of organised cells divided into three regional zones: i) stratum basale; ii) stratum spinosum; and iii) stratum granusolum[4, 5]. The basal cells are cuboidal and act as the proliferative units that replenish the layers above them. As they migrate and differentiate towards the superficial layer they become progressively flatter and disc-like lying parallel to the surface and are eventually sloughed off[6]. The lack of an overlying keratinised stratum corneum, as seen in some animals such as mice, makes the human oesophagus more prone to insult from noxious stimuli[7, 8]. The basal layer is further divided by invaginations of the basement membrane forming connective tissue structures known as papillae[9]. Based on analysis of mitotic figures in the epithelial cells, the interpapillary basal cells divide at a slower rate compared to basal cells residing towards the apex of the papillae[10]. The infrequent division of interpapillary basal cells yields one daughter cell that remains adjacent to the basement membrane and one cell that enters the proliferative region of the epibasal layers. This suggests the interpapillary compartment comprises the site of the oesophageal epithelial stem cell zone[10] (**Fig. 1.3**).

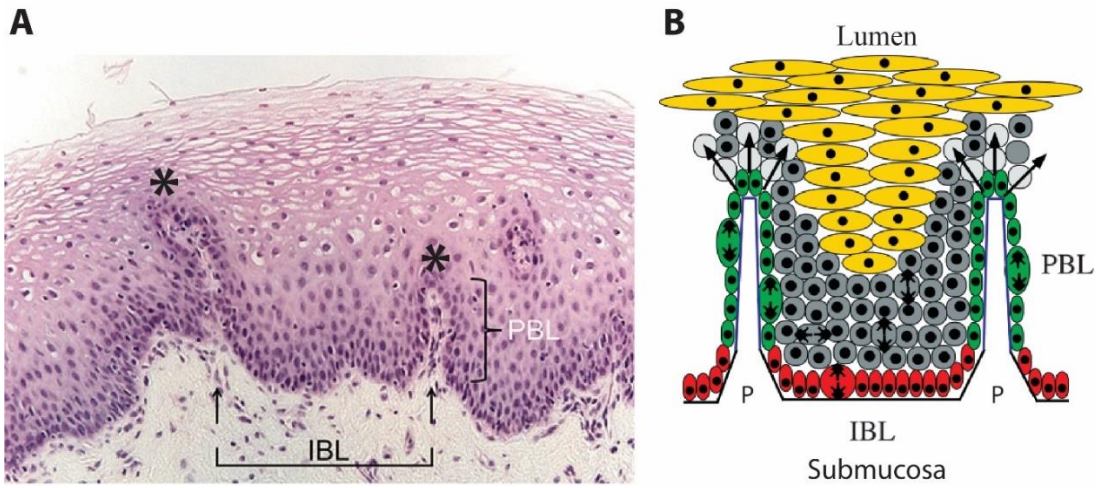

Figure 1.3: *Normal microscopic structure of human oesophageal epithelium. (A) Haematoxylin & Eosin (H&E) section at 20x[11] of the non-keratinised stratified squamous epithelium. Cells become progressively flatter towards the luminal surface (top of image). Asterisks denote the apices of two papillae in this section. Papillae are invaginations of the basement membrane (BM) and submucosa, which define the boundaries of the interpapillary basal layer (IBL) and the papillary basal layer (PBL). (B) Schematic representation for a model of cellular organisation[9]. The IBL cells (red) constitute the stem cell compartment proliferating infrequently at right angles to the BM. PBL cells (green) proliferate more readily heading towards the papilla (P) apex to migrate out of the basal layer.*

The epithelium is supported underneath by the lamina propria (**Fig 1.1a**), a layer of loose connective tissue that also acts as an important barrier to the submucosa in health and disease, especially epithelial cancer. It provides nutrients to the epithelium via a network of small interlacing blood vessels and a first line defence of host immunity being rich in macrophages and lymphocytes. Beneath this, the muscularis mucosae constitutes the final layer of the mucosa, smooth muscle arranged longitudinally[12].

The submucosa, another layer of connective supporting tissue, contains the oesophageal submucosal glands (**Fig 1.1a**). They have a tubulo-acinus structure resembling labial salivary glands in the mouth and secrete mucin for lubrication of the oesophageal lumen[13]. The acinar component consists of a predominantly columnar epithelium[5] which has implications when discussing the origins of Barrett's oesophagus (**section 1.3.2**). The ductal portion remains stratified similar to the

general oesophageal epithelial mucosa discussed above. Beyond this, the submucosa contains Meissner's plexus, which provides parasympathetic nervous input to the superficial layers, and a network of lymphatic drainage, vasculature and immune function properties[14].

1.1.3 Overview of Barrett's oesophagus

Barrett's Oesophagus (BO) is the metaplastic replacement of the normal stratified squamous epithelium with a columnar-lined epithelium presumably as a result of chronic exposure to reflux of bile and acid (Gastro-oesophageal reflux disease, GORD)[15, 16]. Metaplasia is defined as the "transformation from one tissue type to another" occurring at the level of the tissue specific stem cell that re-commits itself to an alternative spectrum of differentiated cells that make up the tissue[17]. BO is the only known precursor lesion to oesophageal adenocarcinoma (OAC) and follows a metaplasia-dysplasia-carcinoma sequence of progression[18] (**Fig. 1.4**). Because of this, all patients diagnosed with BO are enrolled into national endoscopic surveillance programmes in efforts to detect sinister changes early which are more likely to be amenable to curative therapies[3, 19-23]. Persistent GORD is believed to be the primary environmental driver for its emergence and subsequent progression[16, 24, 25].

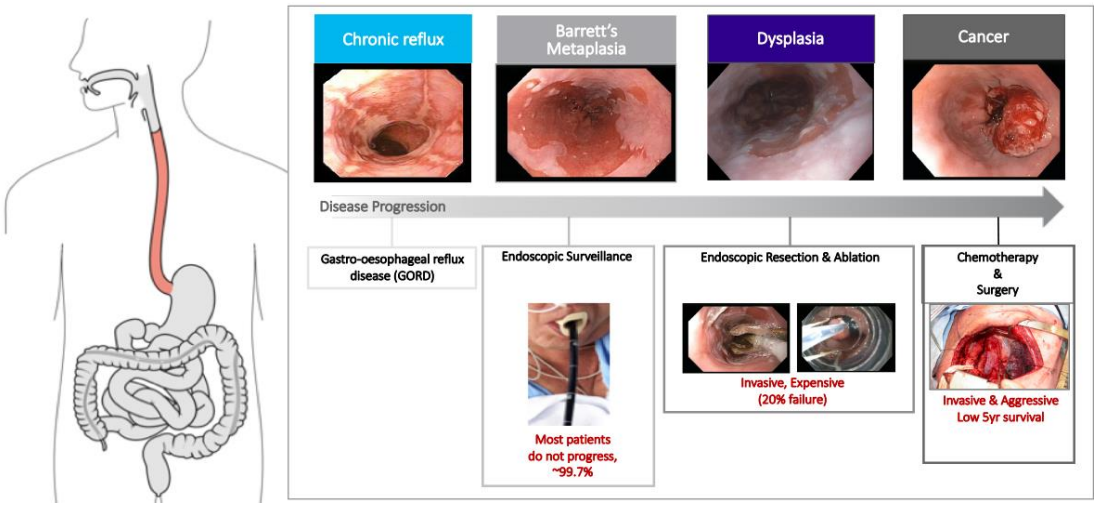

Figure 1.4: *The progression of chronic GORD to adenocarcinoma in the human oesophagus. Gross oesophageal anatomy is shown to the left with endoscopic images to the right. Reflux induces ulceration at the GOJ and SCJ resulting in the metaplastic replacement of squamous epithelium to columnar-lined epithelium, seen here as a "Salmon-Pink" colour. Progression proceeds in sequence through dysplasia to cancer. The potential clinical intervention and survival rates are shown below each stage.*

A BO diagnosis relies on the detection of columnar-lined epithelium in the anatomical oesophagus through upper gastrointestinal endoscopy and subsequent histopathological assessment of biopsies[16, 18]. Macroscopically, BO is classically described appearing as salmon-pink mucosa seen in the distal oesophagus due to the proximal migration of the squamo-columnar junction (SCJ or Z-line) which normally resides at the anatomical gastro-oesophageal junction (GOJ)[26]. The proximal extent of the BO lesion varies dramatically between patients with a mean length of 3.5cm, but interestingly, does not vary over the lifetime of the individual patient[27].

In reporting the presence of BO, a scoring system called the Prague Criteria is used which documents the circumferential (C) component and maximal (M) extent of BO in centimetres from the anatomical GOJ to the Z-line[28]. A greater risk of progression to OAC exists with longer segments of BO[29-32], crudely and somewhat arbitrarily, the clinical cut-off for enhanced risk is defined as segments ≥3cm in length[3]. While the Prague Criteria serves as a reasonable tool of language in diagnosis and clinical discussion, the measurements have variable reproducibility between endoscopists, are subjective and open to bias[28, 33]. This variability clearly has implications when attempting to define risk.

The current standard of surveillance requires quadrantic biopsies from the BO segment every 1-2 centimetres (*Seattle Protocol*[34]) to be sent to the lab to confirm the diagnosis and screen for areas of dysplasia or intramucosal cancer (IMC) that may or may not be seen macroscopically. The Seattle Protocol is not without its flaws. Especially when considering only ~2-5% of the BO surface area is actually sampled thereby exposing a significant false negative risk through endoscopic failure to identify and biopsy a dysplastic region[35-37]. Furthermore, up to 40% of IMC can be missed despite this extensive and laborious sampling[38]. The findings at light-microscopy are defined by the Vienna classification[39] (**Table 1.1**), this, coupled with BO segment length, form the basis of risk stratification and timing of endoscopic interval follow-up[3, 20].

Category 1	Negative for neoplasia/dysplasia
Category 2	Indefinite for neoplasia/dysplasia
Category 3	Non-invasive low grade neoplasia (low grade adenoma/dysplasia)
Category 4	Non-invasive high grade neoplasia
	4.1 High grade adenoma/dysplasia
	4.2 Non-invasive carcinoma (carcinoma in situ)*
	4.3 Suspicion of invasive carcinoma
Category 5	Invasive neoplasia
	5.1 Intramucosal carcinoma†
	5.2 Submucosal carcinoma or beyond

*Non-invasive indicates absence of evident invasion.
†Intramucosal indicates invasion into the lamina propria or muscularis mucosae.

Table 1.1: *Vienna classification of gastrointestinal (GI) neoplasia. Ensure global standards of reporting epithelial changes in the GI tract. Adapted from Schlemper et al[39].*

1.2 The clinical challenges

1.2.1 The burden of oesophageal adenocarcinoma

A diagnosis of oesophageal cancer remains one of the most devastating. Despite advances in clinical care, 5-year survival remains poor at <12%[40]. This is partly due to the late presentation of the disease which has usually spread beyond the local confines of the epithelium but also the aggressive nature of this cancer[41]. Globally, oesophageal cancer is the 8th most common and can be histologically subdivided into oesophageal squamous cell carcinoma (OSCC) and Oesophageal Adenocarcinoma (OAC)[42]. OAC is the more common subtype in the UK and western Europe[43]. The incidence of OAC has also been dramatically rising in incidence over the past four decades partially driven by the mounting burden of GORD and obesity[44-47]. There is a striking male to female predominance of 4:1 in OAC and a strong correlation of incidence to chronological age, rising steeply after 50 years[42, 48]. In the UK, there are roughly 4,500 diagnoses of OAC each year, the vast proportion of these (>90%) present as *de-novo* patients to clinical teams outside of preventative clinical surveillance programmes[49, 50]. Clinical strategies are needed to identify patients at an earlier time-point to enable delivery of more effective curative therapies.

1.2.2 The epidemiology and risk of progression to cancer in Barrett's oesophagus

It is estimated that the prevalence of BO is 0.5-2% in an unselected population[51], approximately 300,000 – 1.3 million UK nationals. For patients with symptoms of GORD the estimated prevalence is between 2-20%[52-54]. Half of BO patients suffer no symptoms and are unaware of its existence[55, 56]. In fact, the majority of BO is diagnosed as an incidental finding at endoscopy for an alternative indication.

Here in lies one of the fundamental unknowns regarding BO, at the time of index diagnosis it is currently impossible to know whether that patient has had the lesion for weeks, months or years and whether this time factor confers a risk disadvantage. An early study gave an estimated mean age of onset of BO at 40 years old (mean

diagnosis was 63 years old), with a mean age of incident OAC at 64 years old suggestive of a 2-3 decade dwell time[57]. This is supported by a more recent study of the Rotterdam cohort of Barrett's patients[58] implying protracted lead times are necessary to confer an increased risk of OAC. It is also noted OAC incidence is highest within the first year after diagnosis of BO, however, this is confounded by a failure to screen for and detect BO at an earlier timepoint and the greater likelihood of neoplasia at index endoscopy, rather than explained by a rapid progression rate[58-60]. Indeed, OAC incidence rate gradually increases over long-term follow-up in previously persistent non-dysplastic BO patients (NDBO)[60]. Moreover, high grade dysplasia (HGD) and OAC development is positively correlated with age[61], further lending strength to dwell time being a risk factor for progression.

The second issue with BO is, despite such a high burden of potential sufferers, the annual incidence of OAC remains comparatively low, yet large swathes of this population who are unlikely to ever develop cancer continue to undergo invasive endoscopy through our inability to confidently identify those at risk.

In patients who have NDBO, the risk of progressing to HGD or OAC is between 0.22 – 0.56% per year[32, 51, 59, 62, 63]. This gives an estimated 1 in 8 to 1 in 14 lifetime risk[64], which clearly depends on the current chronological age of the patient. For a diagnosis of low grade dysplasia (LGD) it proves more difficult to ascertain true progression rates, with studies ranging from 0.6% to 13.4% per year[65-69]. There is a propensity for over-calling LGD by community histopathologists such that when expert histopathological review is arranged, the majority (73%) of LGD is downgraded to no dysplasia or indefinite for dysplasia (IFD) with the remaining LGD (27%) affording a more accurately defined 9.1% yearly risk of progression to HGD or OAC[66]. In this case series NDBO and IFD had a conversion rate of 0.6% and 0.9% respectively.

With a diagnosis of HGD it does not always follow that OAC is inevitable. Isolated HGD itself can be quiescent for long periods or even regress. However, this diagnosis prompts endoscopic therapy and/or surgery as the yearly progression is around 19% and frequently OAC can co-exist with HGD but may have been missed in the biopsy protocol[38, 70-72].

1.2.3 The problem with the surveillance programmes

The majority of patients with BO die from other causes such as cardiovascular, pulmonary and other oncological disease, with OAC accounting for only 7% of total deaths in one such large meta-analysis[62]. Interval endoscopic surveillance of BO has been shown to pick up earlier stage OAC compared to those not receiving surveillance[73]. However, although surveillance is widely implemented, this benefit does not always translate to reduced mortality from OAC despite the significant outlay of exorbitant time and money cost, invasive testing and risk of complications[62, 74-76]. Nevertheless, a mortality benefit has been demonstrated with surveillance conferring a relative mortality risk of 0.386 (95% CI: 0.242-0.617) in one meta-analysis[77] and reduction in 2-year (unadjusted hazard ratio 0.4 [95% CI: 0.32-0.50]) and 5-year mortality in a separate nationwide population-based cohort study of OAC patients[78]. In this latter study though, crucially, the mortality benefit was lost in 20% of BO patients who received inadequate surveillance delivered beyond the appropriate timing interval for the histopathological grade. Follow-up meta-analysis was reconfirmed a lower OAC-related and all-cause mortality in those in surveillance[79]. Attrition and non-compliance of patient participation in surveillance endoscopies over time is well recognised[73]. In addition, there can be variability in the quality, competence and experience of the endoscopist to reliably identify lesions and maintain compliance with the Seattle biopsy protocol potentially rendering that particular surveillance encounter redundant[80, 81]. These factors can erode the intended benefit of a surveillance programme in reducing mortality.

Another issue is that >90% of OAC is diagnosed in the absence of a prior history of BO or surveillance. Part of this will be due to mis-classification of gastric cancers as OAC[82] and due to asymptomatic BO[55]. To reconcile this de novo presentation, one group has suggested an alternative pathway to OAC independent of BO that occurs in 50.4-55.3% of these cases and carries greater mortality[83]. Computational modelling of BO and OAC cases on the SEER (Surveillance, epidemiology and end results) registry in the USA however finds that >90% of OAC arise from BO, this model also provided a BO prevalence rate of 1.9%-2.4% in men and 0.4%-0.5% in women across the general population aged 45-55 consistent with other estimates[84]. The

output of this model is backed up by RNA-sequencing expression data that indicated [85]OAC arises from an undifferentiated BO progenitor cell regardless of whether BO is clinically detectable or not. Research efforts to validate use of a community screening tool are ongoing to try and address this issue and pick up at-risk patients for surveillance[52, 86, 87].

On the converse, 25% of BO patients who develop OAC have non-dysplastic mucosa or maximally LGD in the year prior to its progression[88]. Thus, despite enrolment in surveillance either a focus of HGD / IMC was missed during macroscopic endoscopy assessment or in the randomness of Seattle protocol biopsies. Alternatively, sudden and rapid evolution has occurred for which current endoscopic guideline intervals are inadequate to mitigate against (**section 1.4**). It is recognised that endoscopic detection and biopsy of early dysplasia even with the use of chromoendoscopic and enhanced imaging techniques remains a challenge owing to its subtlety and focal nature within the wider benign metaplastic landscape that dominates the BO segment[89]. Given the low rate of annual progression and majority of surveilled patients who never progress, informally, there can be a degree of apathy from endoscopists conducting such surveillance. This is exacerbated by a perception of limited evidence base in its effectiveness to reduce mortality, limited value to the patient and the laborious nature of fulfilling a Seattle protocol biopsy series in a patient who may be poorly tolerating the procedure and time pressure of a busy endoscopy list[90]. Some of this apathy may be a hangover from an historic lack of any effective management for patients found to have dysplasia such that its detection would not ultimately result in changing the clinical course. However, newer endoscopic eradication therapies (EET) such as radio-frequency ablation (RFA) have now revolutionised the space and clinical outcomes of arresting progression and maintaining remission[91, 92]. This has finally led to the development of key performance indicators (KPI) being defined in Barrett's surveillance to ensure high quality care is delivered for patients. Colonoscopy has long had auditable KPIs defined with subsequent improvement in procedural standards and quality that is now hoped for from an upper gastrointestinal perspective[81, 93-95].

Finally, histopathological grading is prone to interobserver disagreement[96]. The Vienna classification assists in defining grade by assessment of characteristics such as glandular architectural change, budding, surface maturation, crowding and cytological changes of nuclei variabilities, mitotic indices and loss of polarity[39]. Even so, LGD agreement is subject to significant interobserver variability (κ = 0.31) with IFD scoring worse (κ = 0.15)[97, 98]. A diagnosis of HGD/OAC and non-dysplastic BO fared better (κ = 0.65 and 0.58 respectively). Given this report, that is open to wide interpretation, forms a large basis for subsequent clinical management there is an unmet need to identify more robust markers of risk.

1.3 The phenotypic and molecular determinants of progression to cancer

1.3.1 The structure and phenotypes of the Barrett's gland

The canonical Barrett's gland is an admixture of cells from both gastric and intestinalised differentiated epithelial cell lineages forming the so-called *specialised* epithelium[24]. Morphologically these glands consist of an architecture of gastric foveolar cells interspersed with multiple mucinous goblets cells. The glands are generally tortuous and branch deep into the mucosa. At their mid portion is a stem cell zone evidenced by LGR5 and Ki67 expression which maintains the clonal population of cells within the gland through a bidirectional flow of differentiation[99, 100]. In the gland's acinar base reside mucous secreting cells whose function, along with the goblet cells, is to secrete bicarbonate, sialomucins and sulfomucins towards the luminal surface presumably to form a protective layer against further refluxate insult to the oesophagus[24]. **Figure 1.5** shows the stereotypical organisation of BO glands in relation to anatomical space.

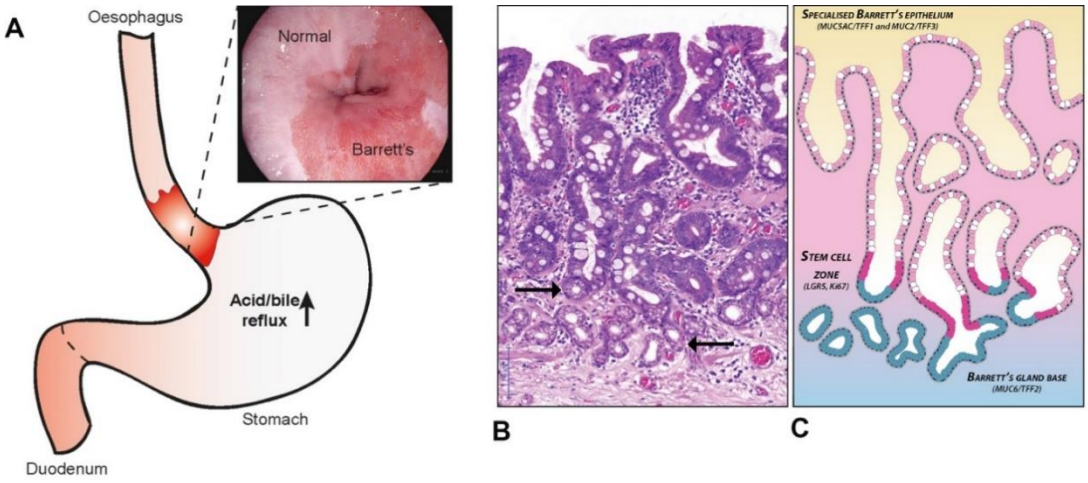

Figure 1.5: *(A) The anatomical location of Barrett's oesophagus adjacent to the GOJ. (B) H&E of the stereotypical Barrett's gland (arrows indicate the stem cell zone at the neck region). (C) Cartoon of (B) showing the compartmentalisation of Barrett's glands with goblet cells and foveolar cells superior to the stem cell zone and mucous secreting cells inferior to this zone. Taken from McDonald et al.[99]*

The finding of goblet cells and hence "intestinal metaplasia" (IM) at microscopy remains a much debated controversy regarding BO diagnosis. In particular, this is an absolute necessity in American guidelines whereas in the UK an absence of goblet cells (so called columnar-lined oesophagus or cardiac-type epithelium) does not preclude a BO diagnosis and subsequent surveillance[3, 19]. This discrepancy exists due to the perceived enhanced risk of progression in intestinalised epithelium over and above the relative safety of gastric metaplasia alone. However, a growing body of evidence refutes this argument which includes OAC being shown to evolve from metaplastic columnar epithelium without goblets cells[101]. Here, shared mitochondrial DNA mutations identified the clonal relationship between cardia-type glands and OAC with this common clonal origin subsequently validated through whole-exome sequencing (WES) alongside the presence of mechanistic oncogenic mutations (for example affecting *TP53*) within the non-goblet containing glands. Through phylogenetic analysis, the WES also demonstrated separate branching of IM within the same BO segment that diverged prior to dysplastic progression of the cardia-type metaplasia and thus the IM did not contribute to the subsequent neoplastic lesion[101]. Previous spatial work whereby histological and immunohistochemical examination of mucosa adjacent to early OAC revealed >70% were associated with primarily cardia-type glands with complete absence of IM in 56.6% of specimens, this finding was preserved regardless of OAC location within the length of BO[102]. Additionally, an increasing concentration of goblet cells within a gland has been shown to be inversely proportional to cancer risk[103], when we consider that goblet cells are fully differentiated, it is unlikely they are the cancer origin cell in BO. Furthermore, comparable risk of progression to OAC has been noted between IM containing and non-IM containing mucosa in a retrospective study examining biopsy specimens from BO surveillance patients between 1980-1994[104], this finding was corroborated by a similar study published a year later[105], although these are at odds with a separate Irish study that did confer a more significant risk of HGD or OAC progression over time in IM containing BO (0.38% per year vs 0.07% per year; hazard ratio = 3.54, 95% CI = 2.09 to 6.00)[63]. However, it is important to note despite absence of IM on biopsies at index endoscopy this is generally not the case during serial follow-up at the 5 year and 10 year interval with IM subsequently

accrued over time in 54.8% and 90.8% of such cases respectively[105]. What is unclear in these studies is whether the original absence of IM is a reflection of sampling error (for example, in long segment BO, if eight biopsies are taken IM is detected in 67.9% of cases versus 34.7% of cases when only four biopsies are taken[106]), or if there has been a new emergence of phenotypic evolution and glandular diversification across the BO segment over time. Nevertheless, these data lend strength to the UK recommendation of continued surveillance of patients with pure cardia-type (non-goblet cell) glandular epithelium at index endoscopy[3].

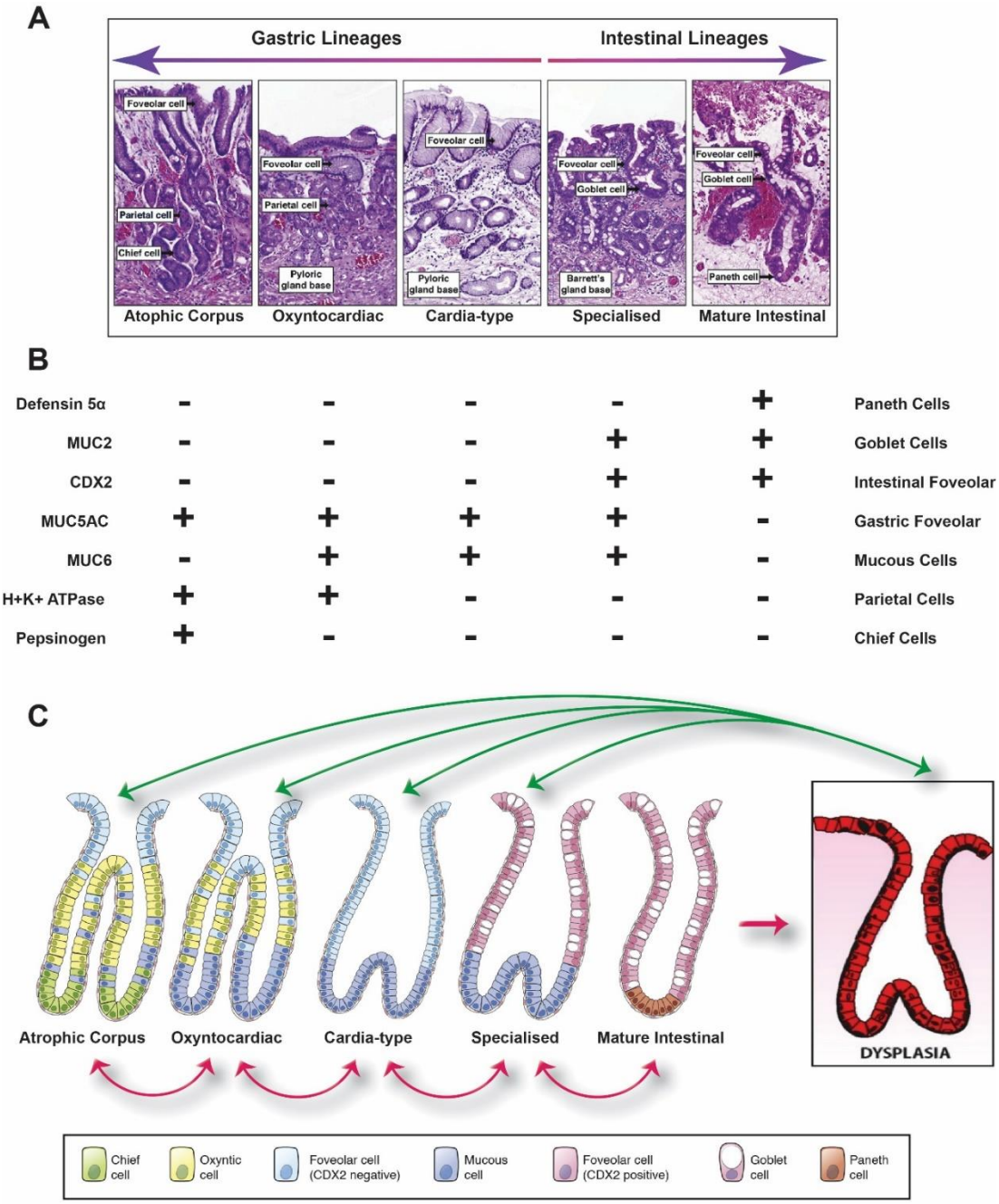

Figure 1.6: *The glandular phenotypes and evolutionary theories. (A) H&E demonstrating the five recognisable glandular phenotypes that form as a spectrum of gastric and intestinal lineages. All phenotypes can co-exist in a single patient. The distinct cell-types that define each phenotype are detailed on each H&E. (B) Lineage markers (left) used in immunohistochemistry analysis to identify the particular cell-types (right) within each phenotype. (C) The phenotypic evolutionary theories of change between phenotypes and progression to dysplasia. It is unknown whether phenotype changes occur in linear stepwise sequential fashion bidirectionally along the spectrum (pink arrows) or if each individual phenotype is capable of evolving to dysplasia directly (green arrows). Adapted from Quante et al.*[24]

The landscape of the BO epithelium is complex and includes at least five different intertwined glandular phenotypes, each distinguishable under light microscopy and immunohistochemistry, that can sometimes form a mosaic pattern across the segment[99, 107, 108]. These are summarised in **Figure 1.6a** existing as a spectrum from gastric to intestinal lineages. At the most basic level is the gastric cardia-type gland, often called the pioneer gland as it is seen interspersed throughout the segment and may be the putative origin gland for BO[24, 108]. Oxynto-cardiac and fundic-type glands become respectively more differentiated towards a gastric phenotype with the presence of both parietal cells (also known as oxyntic cells) and chief cells respectively. In contrast, some glands accrue absorptive enterocytes and Paneth cells resulting in a complete (mature) intestinal phenotype[109]. Zonal mapping demonstrates goblet cell containing glands throughout the lesion but a greater likelihood of gastric phenotypes closer to the GOJ creating a proximal to distal reducing goblet cell gradient[110]. This is perhaps driven by low pH exerting an environmental selection pressure in this region of these more adapted gastric lineage glands.

This Barrett's mosaic is curious in the fact that it's novel assorted glandular architecture and epithelial organisation is not found elsewhere in the gastrointestinal tract but, on a glandular unit basis, striking resemblances to other areas do exist. Thus, each glandular phenotype can be identified by immunohistochemistry (IHC) directed against different, well characterised expression markers specific to the cell types within a gland (**Fig 1.6b**). Mucin (MUC) glycoproteins have a role in the gastrointestinal tract in mucosal protection from toxins, environmental irritants and pathogens in conjunction with cellular signalling activity and immune regulation with over 20 mucin genes now identified[111]. MUC2 expression is the principal marker of intestinal goblet cell and is important to confidently highlight specialised glands which also express gastric lineage differentiation[100, 112]. MUC5AC representing the superficial gastric foveolar cells and MUC6 are usually primarily expressed in the stomach, the latter mucin arising in the deep mucous base of the gland where bicarbonate is also secreted[24, 112, 113]. These three gel-forming mucins provide a viscoelastic protective mucus layer to the Barrett's epithelium that would not be

present in the native squamous epithelium[111, 114]. Combined MUC5AC+/MUC6+ form the basis of the cardia-type gland. Further gastric differentiation with expression of H+K+ATPase (parietal cells) or Pepsinogen (chief cells) result in the oxyntocardiac or fundic-type respectively[24, 99]. Mature intestinal differentiation with presence of Paneth cells is seen uncommonly but detectable by defensin 5α expression at IHC[99].

CDX2 (Caudal homeobox2) is a transcription factor that activates gene expression involved in initial intestinal proliferation, differentiation and maintenance[115], its presence therefore usually precedes the appearance of MUC2+ goblet cells[116]. Indeed, CDX2 expression is found in 30-43% of non-goblet containing BO epithelium suggesting it has a key role in driving intestinal phenotypic evolution[116-118]. Further, other markers of early intestinal differentiation including DAS-1 and villin have a similar prevalence (30% and 17% respectively[116]). Taken together, these markers likely define an early phase of cardia-type glandular intestinalisation that would not be detectable under standard clinical Haematoxylin & Eosin (H&E) assessment and supports the notion of neoplastic potential of such glands.

However, the glandular dynamics and mechanisms that bring about the spatial mosaicism remain to be fully elucidated. Whether cardia-type glands are the true founders of BO and undergo subsequent selective pressures towards increasingly gastric or intestinalised phenotypes or whether all phenotypes arise in unison is a topic of continued debate. We have already seen how both gastric and intestinal lineages share a common clonal ancestry through Lavery et al.'s mitochondrial and WES lineage tracing work[101] proving individual phenotypes are not mutually exclusive genotypic entities but instead related. Evidence lacks however on the exact clonal ordering and mutability between these phenotypes, whether transitions are reversible and the subsequent progression pathway. Perhaps progression arises as a sequential march along the spectrum from gastric to intestinal to dysplastic phenotype, alternatively each phenotype may harbour the ability to become dysplastic directly (**Fig. 1.6c**). Equally, a loss of intestinalisation may be the precipitant factor for neoplastic evolution that could explain Takubo et al.'s study[102] into the predominant cardia-type mucosal appearances adjacent to early OAC. Indeed, decreasing expression of intestinal markers including MUC2 (goblet cell) is

reported in progression from HGD to OAC[119]. The speed and spatial composition of such a bidirectional phenotypic flux, if it exists, along with defining the particular malignant potential of each phenotype would have implications on the clinical diagnosis of BO itself and augment current histopathologically directed surveillance intervals. Finally, studies on (epi)genotypic diversity throughout BO have already given a clear signal of progression risk (discussed in **section 1.4**) but whether phenotypic diversity as a measure is commensurate with these findings is not known, nor whether there is an interplay with the degree of (epi)genotypic diversity breeding phenotypic diversity or vice versa.

1.3.2 Origin theories of Barrett's oesophagus

At diagnosis, the BO lesion is fully established and changes very little, in terms of lesion size and shape, if at all over the duration of surveillance[27, 57]. The ancestral cell of origin is not known however. Understanding the histogenesis of BO to cancer would have implications in surveillance of the metaplasia, where primary focus should be directed during endoscopy and how to deliver targeted therapies to eradicate the condition if and when dysplasia develops to prevent recurrence. We already witness a recurrence of BO between 20-33% after radiofrequency ablation (RFA), including recurrent dysplasia in 5.9% at 8 years and even an OAC relapse rate of 4.1% at 10 years[91, 120-122]. The columnar epithelium that replaces squamous epithelium must arise from either a native cell residing within the oesophagus or a cell migrating into the anatomical space outside the confines of the true oesophagus. Herein, such origin theories are discussed.

1.3.2.1 Gastric cardia-type glandular migration

BO is traditionally thought of as a metaplasia of the normal squamous epithelium and this has permeated scientific research. Many attempts to demonstrate the cell of origin as the squamous stem cell have failed[123, 124]. Our laboratory works on the fact that both specialised and cardia-type phenotypic glands contain differentiated gastric cell lineages[100], ergo an original gastric precursor to BO inception is most likely. Recently, Odze et al. sought to define the normal histological transition across the GOJ from freshly fixed oesophagectomy heart-beating deceased organ donors without history of prior gastric or oesophageal disease[125]. Here, they found a short span of cardia-type mucosa (defined by MUC5+/MUC6+/MUC2-/CDX2- glands) averaging 5.7mm (range 1.4-11.0mm) in length situated between the oesophageal squamous mucosa proximally and the gastric fundic-type (oxyntic) mucosa distally. In addition, populations of densely lobulated glands were also located within the lamina propria beneath the oesophageal squamous epithelium at the SCJ, and in greater density, beneath and across the resected gastric fundic epithelium. The authors termed these collections "compact mucous glands" (CMG). The CMGs

appear morphologically and phenotypically (mucin glycoprotein expression) indistinct from the basal aspect of BO cardia-type glands and so called pseudopyloric metaplasia that are thought to be the basic reparative gland of the gastrointestinal tract[126] including in the ileum[127, 128] (Crohn's disease) and stomach[129, 130] (as a response to oxyntic atrophy and parietal cell loss in *Helicobacter pylori* infection). Thus, CMGs of the underlying lamina propria conceivably expand, proliferate and repopulate areas of squamous epithelium denuded by acid reflux forming the pioneer cardiac-type columnar epithelium characteristic of early BO as discussed in **section 1.3.1** above. It is an elegant and parsimonious hypothesis; Barrett's glands arising from an analogous "normal" structure found intimately related to the squamous epithelium it replaces. A subtly alternative theory might be that the reparative epithelium is derived from migration of the adjacent gastric cardia-type mucosa itself under the process of *fission*[131, 132] (glandular division) driven by natural selection being more appropriately suited to harsh acidic environment. The proximal sited squamous cell progenitors are inhibited by the persistent GORD[24]. Detailed manometric and pH studies of asymptomatic obese patients or patients with hiatus hernia reveal this proximal migration of the SCJ with lengthening of the cardiac-type mucosa in response to acid exposure[133, 134]. Further support for this model comes from mouse models where LGR5 labelled progenitor cells located in the gastric cardia are found to migrate to regions of inflamed squamous epithelium to establish a Barrett's-like metaplasia[135]. Although, how translatable this model is to the human oesophagus is contentious given there are significant anatomical differences of the murine oesophagus and forestomach[85].

1.3.2.2 Residual embryonic and other progenitor cells residing at the gastro-oesophageal junction

Another putative cell of origin is an embryonic-like population of cells at the GOJ that undergo columnar metaplasia as a response to GORD[136]. p63-deficient transgenic mice develop a columnar lined metaplasia of the forestomach owing to inability to maintain stratified epithelium though loss of squamous stem cell self-renewal[137]. Wang et al.[136] identified a monolayer of Car4+ (carbonic anhydrase 4) cells during

embryonic development that is liberated to evolve into columnar epithelium by the absence of p63+ cells that would usually displace them. In the adult p63-wildtype mouse, these Car4+ cells, which also expressed Krt7 (human keratin 7), had largely disappeared aside from a small population of 30 or so cells remaining in situ at the SCJ. Subsequent insult to the squamous epithelium resulted in expansion of these residual embryonic stem cells into the void to form a BO-like metaplasia[136].

Similarly, Jiang et al. describe a population of p63+ cells they termed "Transitional basal cells" (TBC) found at the SCJ as part of a multi-layered epithelium (MLE) that express both squamous (Krt5) and columnar marker (Krt8+ and Krt7+)[138] keratins, again in a murine model of BO. The adjacent cardia mucosa lacked such differentiation markers and thus these cells are not present here. They demonstrated that ectopic CDX2 expression resulted in expansion of the MLE and differentiation of the TBCs to the intestinal phenotype including presence of goblet cells. Comparable cellular populations were subsequently identified in human SCJ samples in particular p63+/Krt7+ cells in efforts to bring credibility to the theory. Analogous to these findings, work in the 1990s had previously defined an MLE present at the SCJ in patients with BO with combined squamous and columnar epithelium potential[139]. A distinctive putative cell sharing these phenotypic features was also identified under electron microscopy as the potential intermediate origin cell to BO[140]. The MLE was present in 41% of patients with columnar lined oesophagus, with a predominance in shorter segments and all MLE was associated with goblets cells leading the investigators to conclude that MLE may represent the early transitional stage to BO with intestinal metaplasia that is then subsequently lost once the inflammatory stimulus subsides[141]. Whilst intriguing, this theory's fatal flaw is that when the GOJ is removed, patients experience uncontrollable acid and bile reflux and Barrett's can often return in the absence of the GOJ[142].

1.3.2.3 Transdifferentiation of squamous cells

Another mechanism is the idea of *transdifferentiation* whereby fully differentiated squamous cells, under environmental stress, change or *transdifferentiate* into

columnar cells[143]. This differs from *metaplasia* which is the transformation of undifferentiated stem or progenitor cells within a tissue into a new repertoire of cell types typically forming a new tissue type[17]. The theory of transdifferentiation involves an initial process of de-differentiation and re-entry into the cell cycle (termed *paligenosis*[144]) at the mature cellular level with reversion to a plastic multipotent phenotype capable of gastric and intestinal lineages to reconcile the cellular architecture seen in BO[123, 145, 146]. However, there are no strong data that validates this hypothesis either *in vivo* or indeed *in vitro* with continuing failure to demonstrate phenotypic conversion of mature squamous cells. Furthermore, the theory necessitates an environmental trigger such as GORD to drive it, presumably transdifferentiation should therefore occur in reverse (columnar to squamous) when this trigger is removed either medically or surgically, but this is not the case with persistence of Barrett's observed. On the converse, when BO columnar mucosa is eradicated by RFA therapy, neo-squamous re-epithelialisation commonly occurs[91] thus suggestive of an alternative epithelial progenitor cell within or migrating to the subsequent tissue defect. Lastly, as we have discussed above, BO consists of a diverse population of cell-types creating the different glandular phenotypes, while the theory proposes that squamous cells transdifferentiate to all of these, it is unlikely and yet to be proven for even a single BO cell-type. For these separate cells to then coalesce to form 3D glandular structures that maintain a single clonal origin both within and between neighbouring glands[132] (through fission) renders the hypothesis implausible. It is important to note that no lineage tracing experiments in human Barrett's tissue has demonstrated a squamous source.

1.3.2.4 Transcommitment of oesophageal progenitor cells

Perhaps more plausible, transcommitment describes the capability of immature progenitor cells to form different tissue types dependent on the environmental selection pressure during differentiation[11]. In this case, the exposure of acid reflux would drive native squamous epithelial progenitors to differentiate into columnar lined oesophagus following mucosal injury. The theories of cardia-type glandular migration and gastro-oesophageal junction progenitor cells described fail to

reconcile findings of so-called neo-BO of the oesophageal remnant following oesophagectomy as surgery usually removes the SCJ and gastric cardia[142, 147].

Molecular reprogramming of the squamous basal cells such that they commit towards a columnar metaplasia is the most studied. Reflux oesophagitis induced in a rat model demonstrates reduction in expression of the basal cell squamous marker SOX2 with upregulated expression of columnar progenitor cell marker SOX9. Mechanistically, a recent *in vivo* transgenic multi-omic murine study reported that activation of the Hedgehog (Hh) signalling pathway (important in embryo- and tumourigenesis[148]) by chronic reflux leads to a heterogenous conversion of Krt5+ basal cells to columnar phenotype expressing SOX9, possibly requiring an initial step of de-differentiation to a more plastic state prior to transition[149]. Notably, many basal cells in this study were unable to convert resulting in and providing a feasible route to how an intermediate MLE of variably mixed squamous and columnar expression and cells may arise. These findings add to similar work by Wang et al. in 2010 whereby downstream targets of the Hh pathway including Patched1 (Ptch1) and bone morphogenic protein (BMP) 4 are not only present in the stromal compartment surrounding Barrett's glands in frozen tissue samples but also shown to induce SOX9 expression in human oesophageal squamous epithelial cells (HET-1A cell line)[150]. Hh signalling is important during embryological formation of the foregut where the oesophagus is initially lined by simple SOX9 expressing columnar epithelium prior to its subsequent transition into nonkeratinized squamous stratification under the influence of progressive Noggin expression that antagonises BMP activity[136, 150-152]. The reactivation of Hh that can be induced by acid and bile reflux[150] acts as the proposed first step of oesophageal progenitor cell transcommitment on a journey towards Barrett's by reverting to the embryonic columnar state[149]. Work in an oesophageal squamous cell line attempts to define how the continuation of transcommitment occurs whereby markers of columnar and intestinal differentiation (CDX2; SOX9; MUC2; Villin) are found to serially increase with persistent acid and bile exposure over time resulting in morphological cellular change, although this is subtle and a far-cry from true Barrett's metaplasia in this model[123].

1.3.2.5 Oesophageal submucosal glands

There is evidence that oesophageal submucosal glands can develop into BO glands and clonal evidence for this has been presented that gives strong credence to this theory[132, 153, 154]. Submucosal glands are lined by simple columnar epithelium in the acinar base that transitions from cuboidal cells in the basal portion of the duct to squamous epithelium opening onto the luminal surface[5]. Various studies have associated them with the formation of squamous islands in treated or ablated BO, the MLE and intestinal metaplasia making them a target of interest in supplying the elusive BO progenitor cell either from the columnar lined acinus or their basal duct[153-156]. Examination of the ducts underlying BO can reveal either a gradual transition from ductal cuboidal epithelium merging into intestinal metaplasia or a more abrupt morphological change[154]. Though, this study of static images cannot resolve whether this transition is truly because of "inside-out" (duct-to-lumen) migration or the reverse with BO arising from another source and extending down the ducts of the submucosal glands in an "outside-in" fashion. Additionally a significantly greater concentration of submucosal glands is found at the transition zones between squamous islands and columnar mucosa within a BO segment suggestive of dual commitment potential of an intrinsic progenitor cell[156]. Further support of this potential is found through expression of both p63 (squamous progenitor marker) and SOX9 (columnar progenitor marker) at the junction between the acinus and basal portion of the main duct[157]. Leedham et al. were also able to identify a shared silent p16 (*CDKN2A*) point mutation in the squamous duct, the submucosal gland acini and the overlying metaplastic BO epithelium confirming a clonal origin and common ancestor of all cell types[153]. In conjunction, a p16 wild-type squamous island was shown to arise from the adjacent wild-type squamous duct into a field of mutated Barrett's demonstrating the capacity for the submucosal glands to be the source of a neo-squamous regenerative epithelium. These findings are comparable to prior and subsequent work that identifies a common precursor capable of generating BO and neo-squamous epithelia through p16 mutation analysis and mitochondrial DNA lineage tracing[132, 158]. A single-cell RNA-sequencing study also identified ~70% of BO is enriched with *LEFTY1*-expressing cells that, in the main, clustered alongside

transcriptional gene expression profiles of normal native oesophageal submucosal glands, although the analysis was from bulk samples rather than separately micro dissected structures[159].

Recently, isolation and separation of submucosal glands into their cellular component parts identified a p63+Krt5+Krt7+ cell population dead-ringer to the TBC described by Jiang et al.[138] found within the MLE[85]. Subsequent single-cell RNA-sequencing (scRNA-seq) of normal SCJ samples identified the TBCs (p63+Krt5+Krt7+), residual embryonic cells (Krt7+MUC4+) and a MUC5B+Krt7+ cell type with maximal homology to expression profiles from submucosal glands suggestive of their common origin rather than arising from adjacent normal squamous or gastric cardia. The presence of Krt7+ across all these cell-types points towards their relatedness but just at different stages of differentiation. However, when this analysis was performed on BO-SCJ and BO samples the expression profiles matched closer to gastric cardia cellular origins (with MUC5AC+Krt20+ foveolar and MUC2+TFF3+ goblet cells) and there was an absence of differentiated or intermediate cell populations expressing Krt7+ in the BO-SCJ samples. The finding of Krt7+ cells in the normal non-BO human tissue with absence of these in BO-SCJ tissue refutes the prior murine models' extrapolation stating their presence and expansion was important and indicative of human Barrett's evolution[136, 138]. The authors used methylation and open chromatin profiles to strengthen their origin assumptions with BO continuing to resemble gastric cardia without any significant overlap of submucosal glands or normal oesophagus profiles. Furthermore, whole genome sequencing (WGS) validated with directed Sanger sequencing identified clonal mutations between BO and gastric cardia in 4 out of 5 patients examined[85]. Taken together, these data would point away from submucosal glands as harbouring the origin cell and provide stronger argument towards a gastric cardia progenitor although expression similarities alone do not prove causality.

1.3.2.6 Circulating bone marrow cells

A final hypothesis involves the haematological delivery of ectopic bone marrow progenitor cells to the inflamed oesophagus[160, 161]. The evidence of this originating from mouse and rat work. In both cases, identifiable bone marrow cells (beta-galactosidase expressing bone marrow cells transplanted into a wild-type mouse[160]; XY male bone marrow cells transplanted into a female rat[161]) were found to contribute to the regeneration of ulcerative oesophagitis to the BO phenotype. Furthermore, in Hutchinson et al.'s paper[160], a case study is presented of a mixed oesophageal adeno- and squamous cell carcinoma in a male human subject who had undergone prior allogenic bone marrow transplant received from his donor sister demonstrated that overall the tumour consisted of at least 6.1% of donor derived bone marrow cells. While these limited studies suggest that bone marrow cells may contribute to distant tissue homeostasis, perhaps in a reparative capacity, they do not prove that BO is originated from them.

It is prudent to note the theories presented here all share commonalities with evidence presented for each, thus they are not mutually exclusive. Indeed, multiple origins could exist for the generation of BO from intrinsic oesophageal progenitor cells to extrinsic progenitors migrating into the commonly shared necessity of corrosive squamous epithelial denudement. **Figure 1.7** outlines the principles theories on the origins of BO.

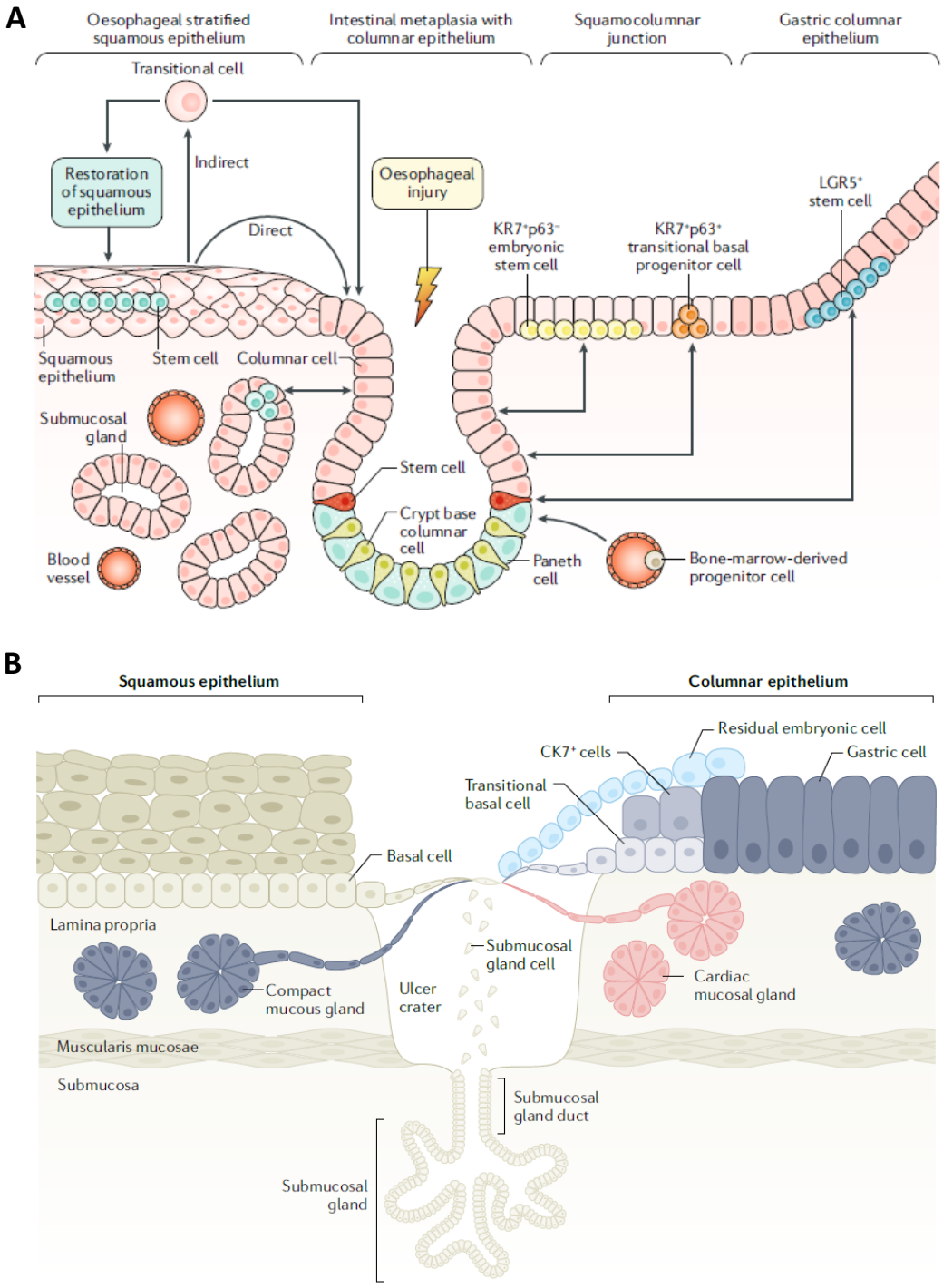

Figure 1.7: *Theories of cellular origin and re-epithelialisation to Barrett's oesophagus. (A) Schematic from Peters et al.[16] represents the anatomical location of the putative cells implicated in Barrett's oesophagus histogenesis acting through expansion, distant migration, transdifferentiation or transcommitment as described in the main text. These cells include squamous epithelial differentiated or progenitor (stem) cells; cells arising from the submucosal glands; residual embryonic stem cells; transitional basal cells; LGR5 labelled gastric stem cells or circulating bone marrow cells. (B) Following oesophageal injury, an ulcer crater exists into which any of the progenitor cells could promote re-epithelialisation from their anatomical locations. This schematic[162] also depicts the location of the newly identified dense cardiac mucosal glands[125] and an additional cellular component of the multi-layered epithelium[138] (CK7⁺ cells, a cytokeratin marker analogous to KRT7⁺).*

1.3.3 The genetic and epigenetic landscape of Barrett's and oesophageal adenocarcinoma

1.3.3.1 The genetic landscape

The progression of BO to OAC provides an archetypal model to study carcinogenesis from a genetic basis. Unlocking this process has clear clinical benefit in risk stratification, prognosis and treatment algorithms. Recent pan-cancer genomic analyses demonstrates that significant driver mutations in tumourigenesis commonly precedes the clinical diagnosis of cancer and can be detected many years prior to presentation[163]. In OAC, there is a single-base mutation (mut) burden in the realm of 8-10 mut Mb^{-1} that is unparalleled by many cancers and surpassed only by melanoma, lung and bladder cancer[164-167]. Perhaps more surprisingly non-dysplastic BO in resection specimens that contains OAC demonstrates a high mutational burden ranging from 1.2-5.4 mut Mb^{-1},[168] this is comparable to non-dysplastic BO from non-progressors with a median of 1.28 mut Mb^{-1} (range 0.12-9.10)[169]. Indeed 79-83% of the mutations found in OAC can be found in non-dysplastic BO[170]. A high rate of progression would therefore be expected, yet of course this does not transpire meaning the majority of such mutations have no functional consequence.

When performing whole-exome sequencing it appears that single nucleotide variations (SNVs) accumulate over time to the same degree in both benign, dysplastic BO and invasive OAC[168] (**Fig. 1.8**). A long tail of mutated genes exists at low frequencies both within a Barrett's segment and between patients, and except for *TP53* (see below) and *SMAD4* (a tumour suppressor gene who's eponymous protein mediates TGFβ signalling[171]) none are discriminatory for histopathological grade[166]. This holds even in many mutated canonical cancer-associated genes such as *ARID1A*, *SMARCA4*, *CNTNAP5*, *ABCB1* previously identified as being at high frequency in OAC[166, 167, 172]. The low frequency ubiquity across histopathological grade renders these mutated genes unlikely to be causal in progression to cancer. Although *SMAD4* could discern a genetic difference between OAC and HGD, disappointingly it was only present in 13% of OAC, thus it is a late event and would therefore be of less value

clinically to identify a point of intervention. Furthermore, Ross-Innes et al. demonstrate that the mutational spectrum poorly overlaps between paired Barrett's samples and OAC[165]. This suggests either significant additional mutations along the pathway to cancer have occurred or that the cancer evolved out of a Barrett's clone within the segment that either has been overgrown by the OAC or was not directly sampled in this analysis[165]. The clinical implications of this marked heterogeneity are twofold: biopsies can easily miss the most dangerous regions harbouring the correct balance of mutations to drive cancer; and secondly, any therapy must target and completely eradicate the entire segment.

Early findings demonstrated key losses of two tumour suppressor genes (TSG) *CDKN2A* and *TP53*[173-175]. *CDKN2A,* located at 9p21, encodes the protein p16 which regulates a cell cycle checkpoint blocking G1 to S phase progression[176]. Hemi- and biallelic losses or inactivation of this gene are found in >85% of Barrett's patients at all grades from metaplasia to high grade dysplasia[174, 177]. The mechanism involves a combination of loss of heterozygosity (LOH), promoter hypermethylation and/or somatic mutations to knock out both alleles[178, 179]. Moreover, separate biopsies with identical *CDKN2A* aberrations are found at multiple levels throughout the BO segment suggesting a period of clonal expansion early in BO development[173, 174, 177], however, it alone does not explain the progression to cancer as many non-dysplastic samples destined for a benign course also exhibit these findings[180].

The deletion by copy number alterations (CNAs) or LOH of chromosomal "fragile sites" occur at high rate in early Barrett's probably soon after *CDKN2A* mutation and prior to loss of *TP53*[181]. Fragile sites exist as areas of chromosomal vulnerability during metaphase prone to breakage or gap formation that may contribute to carcinogenesis by promoting instability with fragile histidine triad (*FHIT*) and WW Domain-containing oxidoreductase (*WWOX*) being the most keenly studied[182]. However, *FHIT* expression is altered in 86% of BO and 93% of OAC[183], is observed in BO of both progressors and non-progressors but typically remains static over time signifying importance at inception of BO but not subsequent malignant transformation[181, 184, 185]. Structural rearrangements to *WWOX* are also noted to be an early and common finding in BO.

The emergent picture of BO progression is one of significant genomic instability with broad chromosomal rearrangements, CNAs, genome doubling, breakage-fusion-bridge (BFB) cycles and aneuploidy that is triggered by loss of *TP53*[168, 169, 186-189]. Located at 17p13, abnormalities of this gene are almost ubiquitous in cancer[190-192]. In normal function, p53 protein promotes G1/S cell cycle arrest, activates DNA repair mechanisms and can trigger apoptosis if the cellular stressors are not rectified[193]. LOH at 17p is associated with genomic doubling to a tetraploidy state and genomic instability[168]. Loss of *TP53* is present in non-dysplastic samples prior to progression to OAC and confers a 13.8 fold increased risk of progression[180]. There remains however a small percentage (2.5-5%) of non-dysplastic patients who never progress who exhibit *TP53* mutation suggesting that a mutation here does not always lead to OAC[166, 180]. This highlights the redundancy in the molecular pathways in that a single gene aberration is likely insufficient to cause cancer in this context[191], rather, a complement of widespread aberrations is required and indeed seen, especially in dysplastic BO and OAC.

Paulson et al.[194] have conducted a large-scale case-control WGS study of multi-regional longitudinal (two timepoints) biopsy samples from 80 patients with non-dysplastic BO (40 who progressed to OAC and 40 non-progressors) followed for a median of 17.47 years (range 4.46 – 29.63 years). Here, a "two-hit" bi-allelic inactivation of *TP53* (–/–) was seen in 75% (30/40) of progressors versus 2.5% (1/40) of the non-progressors, with at least "one-hit" homozygous *TP53* deletion (+/–) observed in 90% and 22.5% case-controls respectively. Among the cases that progressed to OAC, the specific *TP53* aberration detected was present at a higher variable allele frequency (VAF) within biopsies, had expanded across multiple biopsies within a BO segment and was seen at both timepoints. Non-progressors, despite exhibiting the pathogenic *TP53*, the genotypic result was of a more localised, one-hit subclonal population that failed to expand or even persist over time. This latter finding parallels Martincorena et al.'s work where *TP53* aberrant clones are seen to even cover 5-10% of normal oesophageal squamous epithelium increasing up to 35% coverage in advancing age without causing cancer[195]. Nevertheless, the importance of biallelic *TP53* loss is further reflected in clinical p53 IHC where

abnormal staining within non-dysplastic BO, IFD or LGD stratifies patients destined to progress to OAC with a sensitivity of 50.8%, 90.0%, 94.2% and specificity of 98.3%, 84.6%, 54.6% respectively[196]. In the more challenging clinical entity of IFD, use of p53 IHC helped expert histopathologists to re-classify over half of cases to non-dysplastic BO and 5.6-7.4% of cases to definite dysplasia[197].

1.3.3.2 *TP53* mutation drives genomic instability.

Just over two decades ago 17p (p53) LOH was noted to be prevalent and a strong predictor in patients who progressed to OAC[198]. Moreover, there was a strong association with the development of 4N tetraploidy or aneuploidy identified by flow cytometry in patients with 17p LOH (~47%) compared to those without (~9%). More recent whole-exome sequencing and statistical probability corroborates this finding where 90% of *TP53* mutations occur prior to whole genome doubling events that is seen in 62.5% of OACs[168]. All genome doubling events in Paulson et al.'s study occurred after TP53 alteration, in either *TP53* (–/–) or *TP53* (+/–) biopsies[194]. In addition, non-dysplastic BO surveillance samples taken from a cohort of 24 progressors (14 HGD; 10 OAC) with a mean years of follow-up prior to progression of 3.3 (range 1.4-9.0 years) demonstrated mutated *TP53* in 46% (11/24) versus 5% (4/73) of the non-progressor controls but without any significant difference in ploidy or CNAs observed between the two groups[180]. This is consistent with *TP53* mutations arising early before large-scale somatic chromosomal alterations (SCAs) that occur later usually within 2-4 years of OAC detection[185]. Lack of p53 may also be responsible for chromothripsis (shattering of chromosomes with gross genomic rearrangements[199]) and BFB cycling events that amplify oncogenes such as *MYC* and *KRAS* resulting in OAC[187]. Such catastrophic events can be detected in a third of OAC cases (8/22 in the WGS aspect of the study; 40/123 when SNP-array data was added). A follow-up WGS study of 129 OACs was concordant with the finding of dominant CNAs of genes involved in transcription, cell signalling and communication and large-scale genomic events including chromothripsis (30%) and complex rearrangement events (32%) that included BFB cycles resulting in a heterogenous intratumoral landscape[189]. Of note, BFB arises exclusively in the presence of altered *TP53*[194]. A putative mechanism includes the shortening of telomere sequences that has been correlated both with TP53 aberrations and development of chromosomal instability[200]. Loss of the telomeres exposes chromatid ends that become vulnerable to BFB cycles, amplification, recombination, regional chromosomal gains and losses[201]. Shorter telomeres in the context of BFB cycles, *TP53* and histological dysplasia was also found in the study by Newell et al[169]. Finally, The Cancer Genome

Atlas (TCGA) study of 72 OACs defined recurring amplifying or deleting CNAs in key genes including *VEGFA* (angiogenesis), *ERBB2* (receptor tyrosine kinase oncogene), *GATA4* and *GATA6* (transcription factors), and *SMAD4*, in addition to confirming the prominence of *CDKN2A* and *TP53* aberrancy[82]. The genomic profile closely matches chromosomal instability (CIN) gastric tumours even when tumours arising at the GOJ are excluded suggestive of a common tumourigenesis and indirect support for proximal migration of gastric cells in BO origins[202]. Although OAC exhibits greater propensity for SMARCA4 mutation and deletion of RUNX3 (a TSG) compared to their gastric CIN counterparts.

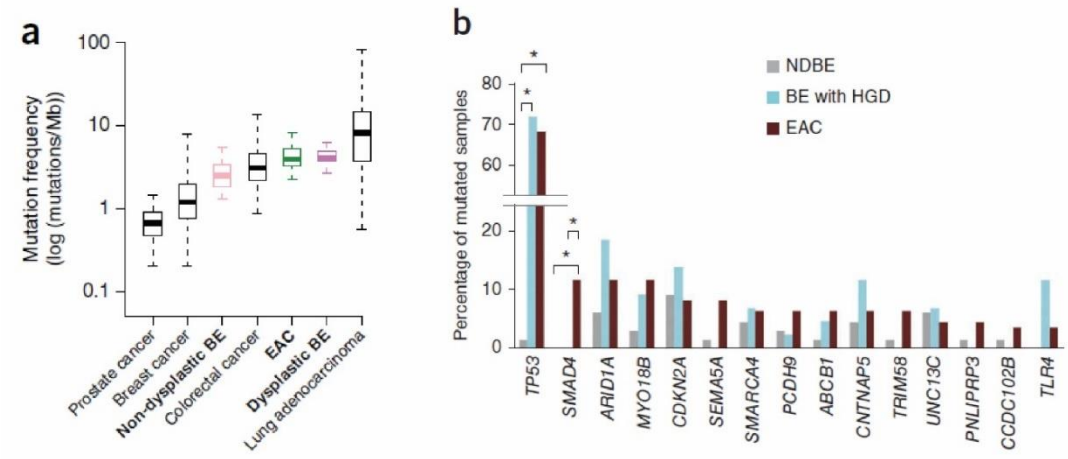

***Figure 1.8**: The genetic landscape of BO. (A) mutation density of Non-dysplastic BO (BE), dysplastic BO (BE) and OAC (EAC) compared to other malignancies. Note that non-dysplastic BO shares comparable levels of genetic aberrancy. (B) Percentage of non-dysplastic BO (NDBE), HGD and OAC with recurrently mutated genes, note that only TP53 and SMAD4 are discriminatory for stage of progression and a long tail of low frequency variants in common cancer driver genes is observed. Adapted from Stachler et al.[168] and Weaver et al.[166] respectively.*

1.3.3.3 The epigenetic landscape

In addition to somatic DNA abnormalities, epigenetic alterations have also been noted in BO and OAC. We have discussed hypermethylation of the *CDKN2A* gene promoter as one particular mechanism for this gene's silencing which is found in 85% of OACs and ~30% of premalignant BO with or without accompanying LOH[178, 179]. A similar picture is observed with the *APC* gene, a TSG coding the adenomatous polyposis coli protein important for cell adhesion and the Wnt signalling pathway, with a contiguous pattern of hypermethylation across the non-dysplastic Barrett's of patients with or without OAC[203-206]. Taken together, the degree of hypermethylation of these two genes at index biopsy of non-dysplastic Barrett's strongly predicts progression to HGD or OAC in one study conducted over a mean follow up of 4.1 years[207]. Promoter hypermethylation of p16 (*CDKN2A*), *RUNX3* and *HPP1* is greater in OAC compared to BE and in the case of the TSG *RUNX3* is seen to reduce mRNA expression levels in OAC as the functional outcome[208]. Additionally, the hypermethylation of these three genes appeared to occur between the BO and LGD interface and were independently predictive of a march towards HGD or OAC. In combination with patient age, segment length and hypermethylation of three other targets (*APC*, *TIMP3* and *CRBP1*) a cox proportional hazards model could identify patient samples destined to progress to malignancy within 2 years of onset[208]. More recently, promoter hypermethylation of these particular gene targets (*APC, CDNK2A, TIMP3* and also *MGMT*) were confirmed to be early events in the metaplasia-dysplasia[209].

Genome wide microarray methods show greater density of methylation at CpG island regions in BO and OAC with hypomethylation away from these sites when compared to normal oesophageal tissue[210, 211]. This indicates differential methylation occurs early in the histogenesis of BO. The hypomethylation occurring in intragenic and non-coding regions potentially promotes expression of non-coding RNAs, such as *AFAP1-AS1*, which has been shown to have cancer driving properties[212]. Complicating this genome wide picture further, not all methylation densities of BO and OAC samples are similar, with heterogeneity identifying a high methylation epigenotype subgroup (similar to so called CpG Island Methylator Phenotypes, CIMP, seen in other

malignancies) and a low methylation subgroup[211, 213]. The significance of which shows a trend towards poorer patient survival in CIMP tumours[211]. This phenotype can also be induced by obesity and smoking[214]. Subsequent in-depth analysis of BO (43% dysplastic samples) and OAC methylation patterns, coupled with functional outputs such as RNA-sequencing and clinical data, has now defined four distinct OAC subtypes with variable prognosis and response to available oncological therapies[215]. The CIMP-like phenotype's (classed subtype 1) poor prognosis is surpassed by an immune cell infiltrative phenotype (classed subtype 3) that harbours little to no significant alteration of methylation pattern over normal tissue controls (oesophagus, stomach, duodenum)[215]. The hypomethylation phenotype (classed subtype 4) was characterised by excessive copy number of the oncogene ERBB2 amongst other high frequency SCAs suggesting reducing methylation levels also promote genomic instability. Subtype 2 clustered the majority of BE cases together with OAC only representing 17% of the subtype samples. The BE cases here have regions of genomic hypomethylation important in maintaining the BE phenotype with a correlative methylation pattern found in the normal gastric tissue control, these regions are variably methylated in the OAC samples of this subtype[215]. This is indicative of progressive hypermethylation prompting a transition to the malignant phenotype. Aligned with this, in a separate study of non-dysplastic BO, a pattern towards epigenomic hypermethylation can distinguish histologically identical samples into progressors vs non-progressors[216]. Aligned with this, TCGA data revealed a proximal to distal DNA hypermethylation gradient from OACs to CIN gastric cancers (70% vs 30% cancers with hypermethylation)[82].

Clinical risk stratification tools based on differential methylation at all stages of oesophageal histology from normal to OAC under current research generally report the same or similar discriminatory genes but as yet have failed to materialise as clinically translatable entities[86, 87, 203, 205, 208, 217-221]. Though recently, a methylation panel has been shown to be useful in defining successful eradication of BO following clinical RFA treatment[222]. Here, the panel differentiated between recurrent or residual IM (whether macroscopically visible or not) with a 35.9% methylation level versus remission with a normal GOJ and 1.8% methylation level, this directed

subsequent RFA retreatment to the residual IM resulting in a further 7.6 fold reduction in methylation levels.

1.4 Evolutionary dynamics and the clonal mosaic of Barrett's oesophagus

As discussed, the Barrett's segment can exhibit a broad range of gland phenotypes with variable mutational genotypes. Of importance, the cellular architecture of a gland is maintained by the stem cell zone found at the mid portion, therefore the most basic clonal unit under environmental selection is the individual gland[132, 223]. All cells within a gland share their ancestry from this small pool of stem cells (see section 1.5.4). Furthermore, through mitochondrial lineage tracing, large patches comprising several Barrett's glands are clonally related suggesting that glands expand across the segment by fission (gland division)[131, 132]. Similar findings of clonally related glandular patches are present in the stomach[224]. Fission is proposed to initiate from the stem cell zone with bifurcation and longitudinal separation to form two daughter glands[99]. This process is vital in the normal post-natal physiological growth of the intestinal tract with mean percentage of bifid crypts peaking between age 6-12 months at 18% compared to just 1.7% seen in adults[225]. Despite this apparent quiescence in adulthood, inflammatory processes such as ulcerative colitis, Crohn's disease and polyposis syndromes have also been shown to drive fission, including the finding of specific clonal genetic mutations at multiple locations along the colon demonstrating the proficiency of expansion[226-228]. Also in the colon, a process of crypt fusion has been described that balances the effect of crypt fission in a homeostatic manner resulting in an estimated rate of 0.011 divisions/crypt/year (range 0.002-0.024)[229].

It was previously believed BO arose as a monoclonal lesion, that is, from a single transformative progenitor cell, forming glands that spread across the entire segment[173, 174, 230, 231]. This was based on the finding of apparently identical *CDKN2A* and *TP53* aberrations (mutations or pattern of LOH) at multiple levels filling up to 85% of the Barrett's segment[173, 177, 231, 232]. The aberration would confer a phenotypic selective advantage to the clone allowing a sweep to fixation throughout the lesion at the expense of potentially other competing clonal populations that went extinct,

for example squamous epithelium progenitors[233]. In this monoclonal model, a proportion of distinct aberrations would always be present no matter what region of BO was sampled. However, this is not what is found. With the advent of higher resolution techniques such as tissue microdissection and next generation sequencing, the BO is revealed to be polyclonal[153, 165]. Here, multiple spatially separated progenitor cells evolve into the Barrett's phenotype driven by their newly acquired genotype and environmental selection to form clonal patches that compete for the space. As discussed in section 1.3.2, the key progenitor cell of origin remains to be fully elucidated and may indeed involve a combination of many, if not all, theories. Whatever the answer, the outcome is a BO lesion formed as a patchwork mosaic of both genotypically and phenotypically distinct clones, and ubiquitous monoclonal sweeps are not found[24, 223, 234] (**Fig. 1.9**).

It is important to note that the two theories are not mutually exclusive merely the timepoint of reference is shifted. Indeed, all cells within an organism are ultimately descendants of a single zygote, however, between conception and BO histogenesis, different populations of cells within the GI tract have accrued different genotypic histories which may be advantageous, neutral or deleterious and are thus now polyclonal. It is these alterations alongside new BO driver changes (such as p16 [*CDKN2A*] variants) measured across a segment that define the polyclonal origin of the disease. The monoclonal model recognises the BO segment is heterogenous but suggests polyclonality arises through subsequent branched evolution[235] after the initial monoclonal expansion. Evidence of both models is available. In a recent phylogenetic reconstruction study that deconvoluted subclonal populations of bulk biopsy samples across the BO segments of progressors and non-progressors, 55% were determined to have a monoclonal common ancestor with 45% arising from polyclonal origins which could be from up to 4 (or potentially more) founder sources[236]. Each model conferred the same chance of progression.

Following inception, the BO lesion appears to enter a prolonged period of evolutionary stasis where a *dynamic equilibrium* is established[223] (**Fig 1.9**). Longitudinal sampling finds that while there can be small expansions of clones, balanced by the contraction of another clone, no single clone comes to dominate the

landscape[237]. Furthermore, clonal populations can become extinct and new clones can emerge, however the net effect during this dynamism seemingly bears no significance to phenotypic progression as positive evolutionary selection is not seen[234, 237]. Over time, the global diversity of the clonal mosaic therefore remains constant (**Fig. 1.9a**). This persistence is seen in both patients with non-dysplastic BO who progress to cancer, and those who do not. Though crucially, the former demonstrates a significantly higher baseline genetic diversity[238, 239]. The use of diversity indices such as Shannon index (a measure of frequency and relative abundance of clones[240]) and genetic divergence (a measurement of the accumulation of differences between two ancestrally related clones) has thus become a promising marker for risk stratification[184, 238] (**Fig. 1.10**). By taking a single sample at any timepoint and establishing the make-up of clonal populations an assessment of risk that predicts progression can be made[238, 241]. The theory goes that a diverse segment is "born to be bad" carrying the necessary complement of geno-phenotypic variation with a protective redundancy to tolerate extrinsic environmental changes and an intrinsic propensity to a more rapid mutation rate that results in increasing genomic instability over time and selection of a dysplastic phenotype[184, 237, 242].

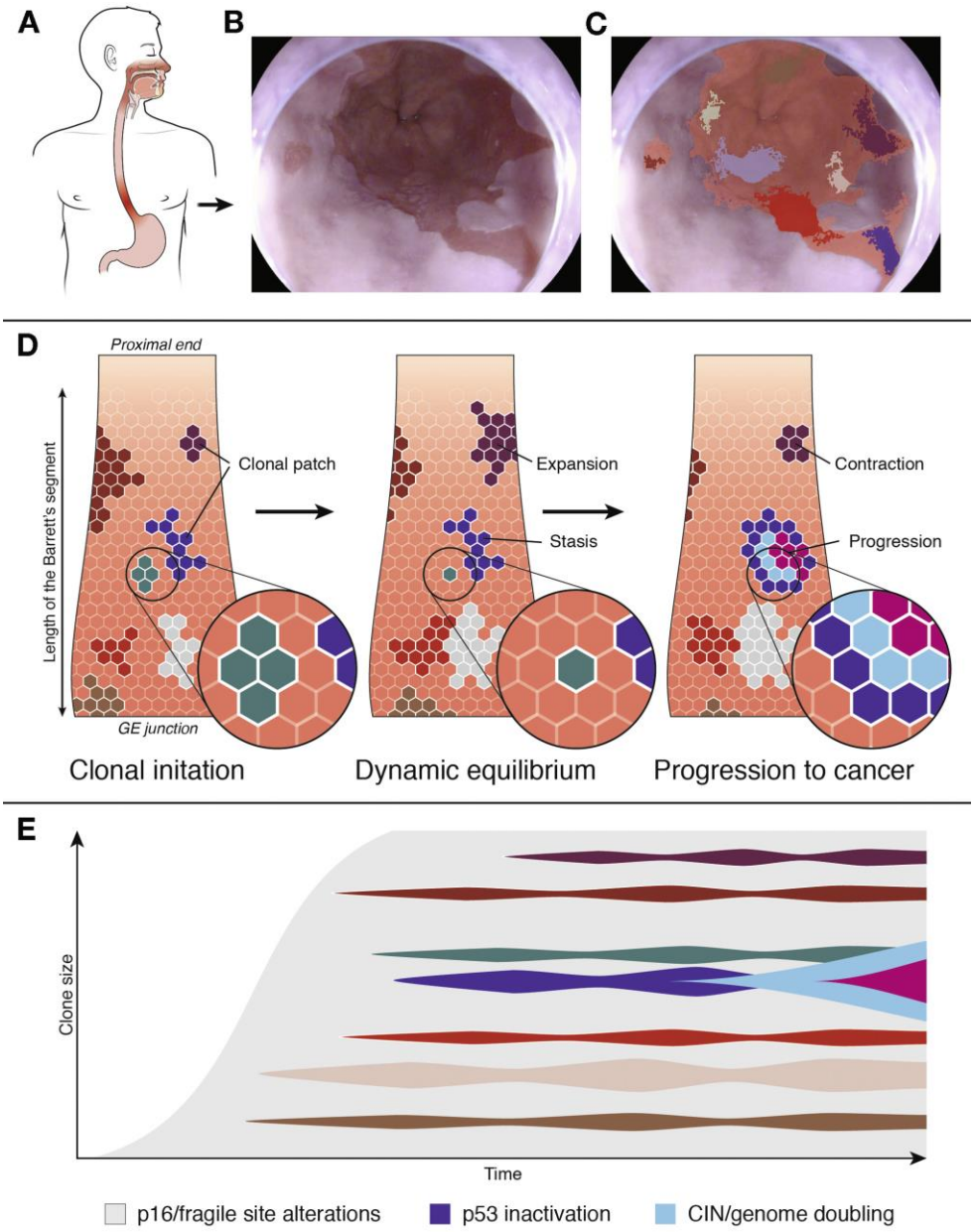

***Figure 1.9**: The clonal mosaic of Barrett's oesophagus. (A) The anatomical location of BO. (B) Endoscopic view of BO in the distal oesophagus. (C) Schematic colour overlay of representative individual clones within a BO segment. (D) Schematic of the unfolded oesophagus containing Barrett's. The lesion is polyclonal forming a patchwork mosaic across the space (left) with each gland maintained by a small population of stem cells. In dynamic equilibrium (middle) each gland expands and contract over time but there is no significant net change in the proportion of clones and diversity. To progress to cancer (right), clonal selection occurs with expansion of the dysplastic phenotype, the trigger of which is potentially the development of genomic instability. This can be multi-focal from individually distinct clonal populations. (E) Clonal frequency and abundance over time. The y-axis demonstrates the size of a clone within the BO segment, the x-axis is time. Polyclonal expansion of mutated p16 (CDKN2A) clones occur early from which subclones emerge with progressively divergent genotypes that compete with each other. Aberration of p53 (TP53) is a key event that promotes genome instability and widespread somatic chromosomal alterations resulting in an evolutionary advantageous dysplastic phenotype that expands to form OAC. Taken from Quante et al.[24]*

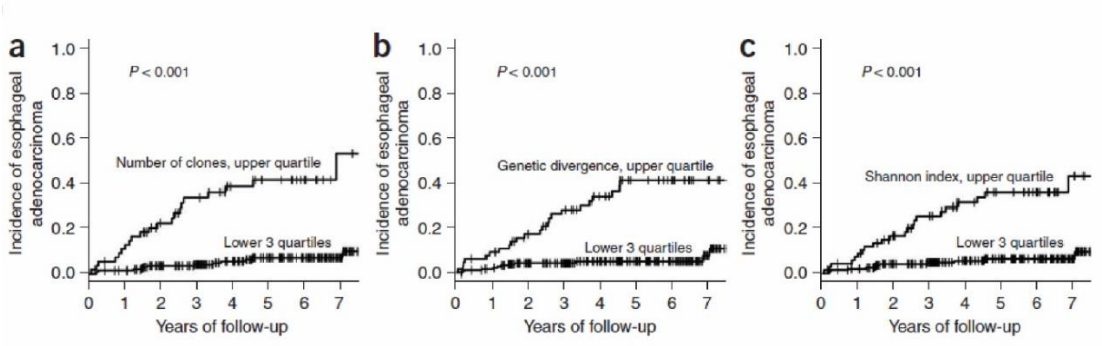

***Figure 1.10**: Clonal diversity of the BO segment predicts progression to OAC. Kaplan-Meier incidence curves split all data into two groups, those that form the upper quartile of values and those that form the remaining lower three quartiles with incidence of OAC over time. All plots show an increased probability of cancer the greater the index. (A) Number of clones, (B) mean pairwise genetic divergence, (C) Shannon diversity index. Taken from Maley et al.[238]*

It is now recognised that while diversity over time appears to remain stable (**Fig. 1.11a**), there is a measurable transition to chromosomal instability that breeds an exponential increase in diversity and catastrophic genome doubling causing dysplasia and OAC[24, 185]. While *TP53* mutation likely precedes this, it is unclear what exactly triggers this transition in the non-dysplastic Barrett's but large scale SCAs are seen to occur up to 4 years before the clinical phenotype and presentation of OAC[185] (**Fig. 1.11b-c**). Killcoyne et al. suggest the detection window for high risk segments may be even longer at ≥8 years before clinical progression[188]. Here, they performed shallow whole genome sequencing (SWGS) on a retrospective cohort of non-dysplastic BO surveillance samples where half of patients progressed to HGD or IMC. Patients were binned into a low, moderate or high risk categories based purely on the complexity of the CNA profile observed, with greater complexity conferring greater risk. CNA profiles were again stable over time but 50%, 78% and 85% of high risk patients could be identified in the ≥8th, 2nd and last year respectively prior to HGD/IMC evolution[188]. A significant stepwise increase in complex SCA burden is seen from non-progressors to progressors to OAC patients[194]. As the structural variant burden gradually increases over time there is high chance of catastrophic events including chromothripsis and BFB cycles that triggers a dysplastic clonal expansion[243]. These studies, in conjunction with detailed temporospatial mapping that shows phenotypic dysplasia arising in a polyclonal and multifocal fashion, lends strength to

the theory of *field cancerization*[244] and a BO segment that is inevitably primed for progression[165, 242]. Furthermore, when in state of pre-progression indolence, large clonal expansions are not observed, instead clones exist in the order of square millimetres and there is a comparable degree of genetic diversity in single glands that is observed at the whole biopsy scale[184]. The value of a single "snapshot" biopsy to inform risk is clinically enticing. Coupled with an understanding of how the SCA or CNA profiling changes (or does not change) over time provides a window of opportunity to intervene clinically or shorten surveillance intervals. For example, for a case of high baseline genomic complexity or if a sudden change to a rising trajectory of genomic diversity occurs.

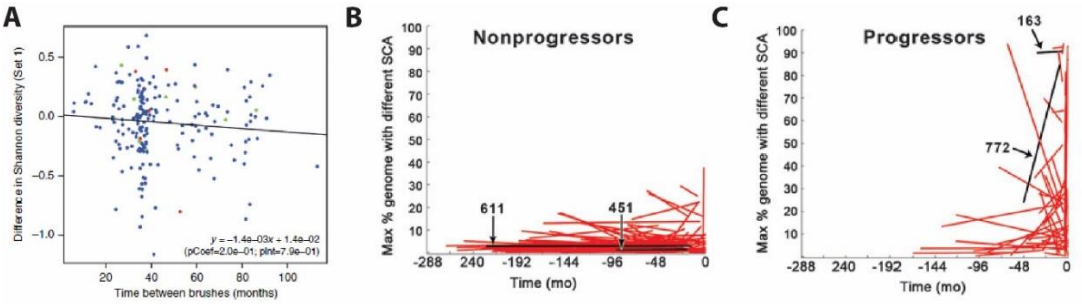

***Figure 1.11**: Change in clonal diversity over time. (A) Plot of Shannon diversity index at two timepoints from BO brushing samples which demonstrate net diversity is stable over time in keeping with a dynamic equilibrium. (B) and (C) demonstrates that diversity (development of somatic chromosomal abnormalities [SCA]) increases in patients who progress to cancer vs non-progressors with a rapid accrual of aberrancies seen up to 48 months before presentation with OAC. The non-progressor cohort remains stable, in keeping with (A). Plots taken from Martinez et al[237] and Li et al[185] respectively.*

1.5 Exploiting the epigenome to infer pathogenesis, natural history and progression risk of epithelial neoplasms including Barrett's.

1.5.1 Overview of epigenetics

Epigenetics refers to the study of gene expression brought about by heritable changes that do not affect or involve the DNA sequence[245]. The principle effect of modifications is to alter transcription brought about by variations to the chromatin (DNA packaging) structure[246]. In particular, such mechanisms include posttranslational modification to histones usually occurring on the histone tail, nucleosome ordering, variations in higher-ordered folding of chromatin and methylation of cytosine bases[247]. The result is multifaceted whereby gene expression can be affected by modifications not just directly but also by adjacent genes and changes at sites very distant to them, all determined by the protean 3D structure of chromatin[248]. Furthermore, epigenetic modifications are reversible and more readily impacted by environmental factors[247].

DNA methylation is the most well studied modification and refers to the addition of a methyl group by DNA methyltransferases (DNMT) to the C-5 of cytosine bases within the somatic DNA generating 5-methylcytosine (5-mC)[249]. Almost all methylation occurs at Cytosine-phosphate-Guanine (CpG) loci where cytosine is located immediately 5' to guanine in the DNA sequence. Dense clusters of CpG loci are termed CpG islands[250] and are often found associated with promoter regions, first exons and the 3' end of genes providing a means of variable methylation to interfere with transcriptional machinery impacting the expression of that gene[251].

The distinct pattern of methylation is inherited during mitosis to both daughter cells (**Fig. 1.12**), DNMT1 is the maintenance methylase that copies across the sequence on the hemi-methylated DNA to the new strand[252, 253]. The replicative fidelity results in reasonably stable patterns, however, stochastic errors are estimated to occur at a rate of $2x10^{-4}$-10^{-5} per cell division[254-256]. This equates to ~500-5000 alterations in the pattern observed across the genome which can be either hyper or hypomethylation events. The error rate may even be as high as 5% per CpG per

division in some studies[253, 257-259]. Furthermore, *de novo* methylases DNMT3a and 3b can also erroneously methylate CpG loci, whether this happens in conjunction with DNMT1 or as a separate process is not clear[260]. For reference, the median estimate somatic DNA mutation rate is 0.38-2.8 x 10^{-9} per bp/mitosis or 1.14-8.4 mutations/division[261, 262].

Of importance, at fertilisation to form a zygote there is extensive epigenetic reprogramming, in particular involving global active demethylation of the paternal pronucleus (from the sperm) and less pronounced passive demethylation of the maternal pronucleus (oocyte)[263]. The exact mechanisms underpinning this remain to be fully explained but the net result is that by the 8-16 cell morula stage return to a totipotent state has been achieved by significant reduction in methylation levels[264]. In females, who carry an XX karyotype, there is a further step of X-inactivation through hypermethylation of promoters to silence the extra (usually paternal) X chromosome and prevent "double-dosing" gene expression[265]. These embryological events permit new differential methylation and gene expression that generates a phenotypic plasticity between offspring that ultimately share the same underlying somatic genetic code[266]. This renewed relative hypomethylated state is relevant when we consider the process of epigenetic drift during chronological aging.

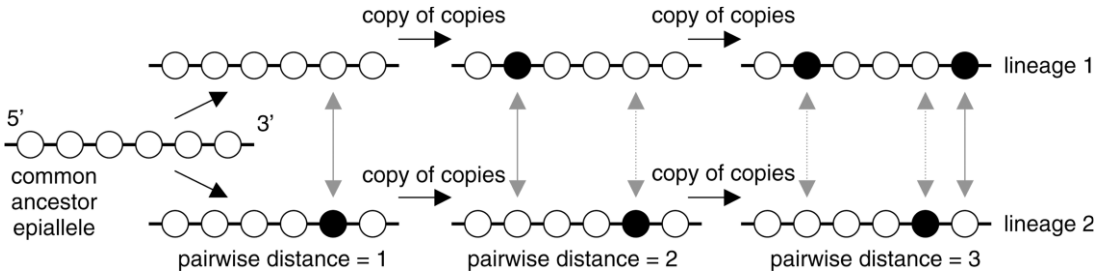

***Figure 1.12**: The emergence of stochastic methylation replication errors. During mitosis the common epiallele methylation sequence should be copied to the two daughter cells, however the fidelity is poor and random errors are introduced. Through successive rounds of mitosis further errors occur with the epialleles diverging from the common ancestral origin. Calculating the pairwise difference between two populations infers their mitotic age. Also, increasing methylation from an unmethylated starting epiallele is a function of age and can be measured as average methylation density. Each CpG locus can either be unmethylated or methylated generating a binary string code of 0 or 1 useful for bioinformatics analysis. Taken from Shibata et al.[267]*

1.5.2 Epigenetic drift and the epigenetic clock

It has long been recognised that the methylation state of the 28 million CpG loci across the genome are seen to change with chronological age[255]. The pattern of change is complex but in general it appears that CpG islands associated with promoter regions, which overall have low levels of methylation, become progressively hypermethylated[267], whereas non-island CpGs found in intragenic regions become hypomethylated[268]. The drivers of this balance and subsequent biological contribution to the ageing process are yet to be fully elucidated. The random errors that occur at mitosis or those induced by environmental factors are such mechanisms that generate the variable methylation pattern over time. When this is seen as a discordance across populations of cells, tissue types and individuals it is termed "epigenetic drift" and introduces epigenomic mosaicism. This is in contrast to certain CpG loci across the genome which reliably undergo a consistent methylation change between tissue types and individuals proportional to chronological age, so called "clock CpGs". This latter phenomenon has allowed the development of commercially available "epigenetic clock" algorithms that can predict biological age and even the risk of age related illnesses and mortality[269, 270].

Epigenetic drift is seen as one of the potential hallmarks of aging that also includes genomic instability, telomere shortening, stem cell exhaustion and mitochondrial dysfunction[271]. It presents as a heterogenous entity with bidirectional hypo- or hypermethylation at varying frequency between age-matched individuals and within different chronologically age-matched tissues types[266]. The determinants that set the pace of drift are multifactorial including environmental exposures, nutritional factors, obesity, smoking, inflammation, and presence of neoplasia (for example drift is 3-4x faster in the neoplastic colonic mucosa[272]) that leads to variable time of onset of age-associated disease[246, 255]. In inflammation, there is stimulation of stem cells to divide and repair tissue, the consequence being chronic or repeated inflammatory cycles promote accelerated epigenetic drift, potential for gene silencing and constriction of stem cell plasticity with age[255].

Copying errors that cause drift induced by DNMT1 are more commonly observed in highly proliferative tissues[256]. This defined error rate could thus calibrate a mitotic

clock to age tissues provided the affected gene was not prone to evolutionary selection and the new methylation state was neutral in this respect[273]. Methylation patterns in a 5' to 3' direction can be converted to a binary string of 0 ("unmethylated") or 1 ("methylated") depending on the presence or not of a methyl group when the epigenetics is interrogated through various laboratory methods (**Figs. 1.12 and 1.13**). Each binary string represents a single allele-specific methylation "tag". Comparing the binary string between two cells or populations of cells within a tissue-type reveals their clonal relationship[273]. Because CpG islands are hypomethylated at birth, replication errors in these regions lead to more progressive methylation density over time[274]. Each newly methylated site represents a heritable error produced during mitosis. Where the pattern is similar or the same this suggests a recent common ancestor in comparison to widely divergent patterns which indicate more substantial drift from the common ancestral cell and hence a mitotically older population or tissue. The difference between the 5' to 3' sequence of two epiallele tags is measured as a pairwise distance (PWD) such that within a population of cells (and hence multiple distinct epialleles) the average PWD can be calculated where higher values indicate greater diversity and age[273] (**Fig. 1.13**). Thus, the study of methylation tags from invariably evolutionary neutral CpG loci altered randomly by drift during mitosis without biological consequence forms an elegant lineage tracing technique *in vivo*.

In the normal colon, methylation tags reveal the somatic phylogenesis from the single cell zygote to the colon crypts. Closely related crypts, for example those that have recently undergone fission, will have similar methylation patterns, so called low "intercrypt distance". Whereas greater intercrypt distance through epigenetic drift is indicative of more disparate crypts over time since the common ancestor. Findings demonstrate that intercrypt distances are no different between adjacent crypts and crypts at least 15cm away consistent with long lived crypts and minimal subsequent fission events in the normal colon[275]. This fits with fission peaking during infancy and childhood with subsequent stasis thereafter[225].

1.5.3 Epigenetic drift in pre-malignancy and cancer

The stochastic epigenetic drift of CpG loci during mitosis has permitted the study of clonal dynamics and rates of expansion particularly in conditions such as colorectal adenomas and cancer[272, 276-279].

With this concept in mind, colorectal adenomas have been shown to establish their entire crypt population through a process called "punctuated" evolution where long periods of stasis of little to no growth is observed with intermittent periods of fast expansion[276, 279-281]. This is evidenced by a diverse methylation pattern seen between crypts within bulk adenomas and that the pattern of adjacent crypts is no more similar than distant crypts suggesting their evolution occurred at the same time[276, 279, 282]. Moreover, the patterns demonstrate that an intratumoral *APC/KRAS* subclone likely simultaneously arose with the general background bulk *APC* clone as they had similar methylation patterns. Other intratumoural clones, defined by mitochondrial DNA (mtDNA) mutations showed a more homogenous methylation pattern with less diversity than the bulk adenoma suggesting a more recent expansion but, like the *APC/KRAS* clone, a clonal sweep through the tumour had not occurred. Rather intratumoural heterogeneity (ITH) exists where all subclones appear to occupy a distinct region in spatial competition whose life histories of branched evolution are revealed by the epigenetic mitotic clock they exhibit[235, 276, 283]. These findings contradict the gradualism model of step-wise sequential accumulation of cancer gene driver mutations postulated by Vogelstein[284, 285] and adds to our understanding of mutational ordering and evolutionary dynamics in colorectal carcinogenesis that is instead marked by punctuation[281]. Furthermore, high pattern diversity across each adenoma implies a mitotically old population to account for the pattern variability change since the founder ancestral crypt[276].

Further studies on colorectal cancer provide similar findings that challenges the dogma of sequential clonal sweeps over time that would result in a globally homogenous methylation pattern. In particular, methylation tag patterns are complex and heterogenous within and between different fragments of a cancer but share similar diversity of patterns as measured by PWD[282]. This suggests that there is a rapid "flat" clonal expansion at inception of the cancer from its common ancestor

that subsequently abates to a more indolent state. If a new favourable somatic mutation arose sequentially, for example the ability to metastasise, the average PWD would fall as the clone expanded across space as a "younger" mitotic population. However, when the investigators compared superficial cancer crypts with invasive ones there was no significant difference in average PWD suggesting each phenotype arose at the same time[282]. This is aligned with the "born-to-be-bad" theory[280] of cancer evolution and reconciles the finding that occasionally small primary tumours have already metastasized[286]. Cancers from different patients also exhibited different mitotic ages (average PWD) consistent with varying time to clinical presentation since biological onset[282]. Although, mitotically older cancers have been correlated with advancing chronological age where the cancer arises from a population of already mitotically older stem cells, measured by methylation density[287]. These examples confirm that analysis of pattern variability is suitable to provide information about relative age of both benign metaplasia and malignant tumourigenesis.

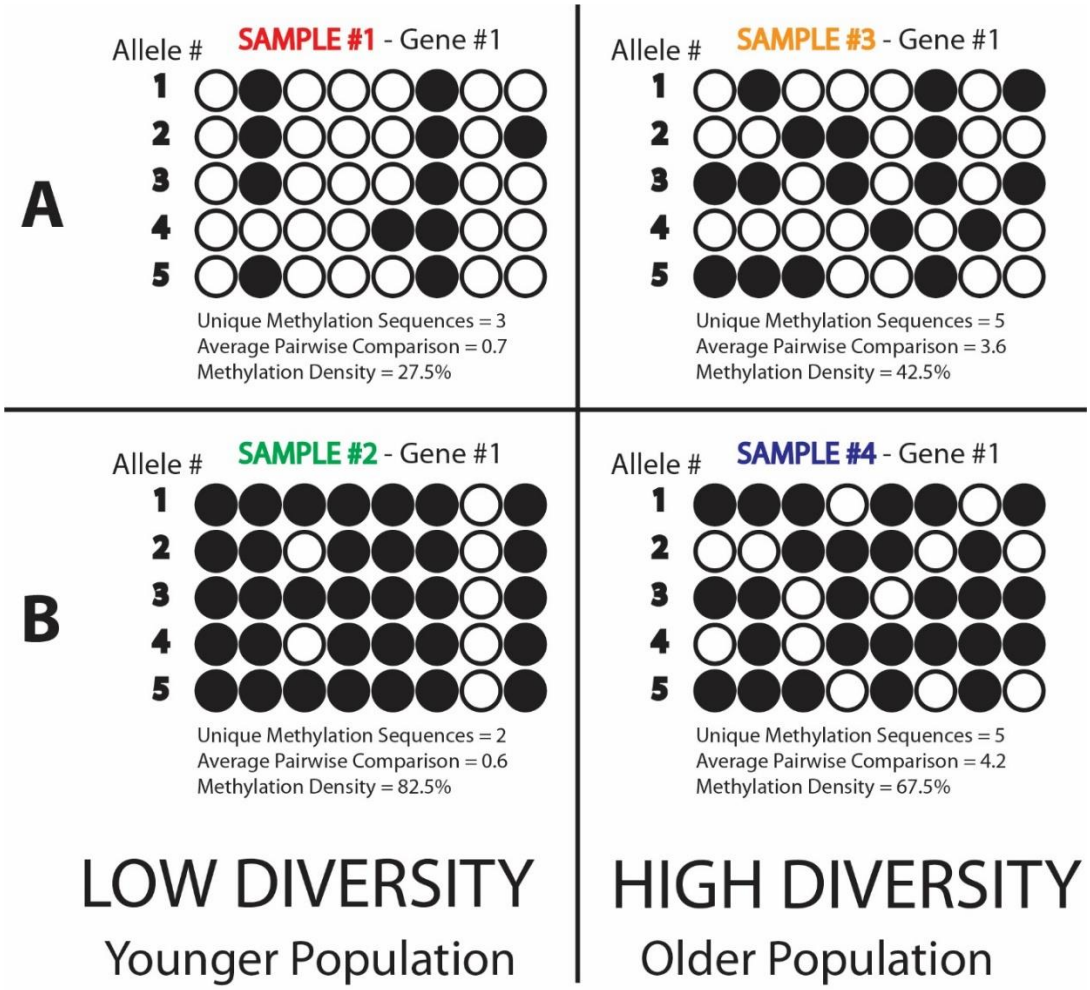

***Figure 1.13**: "Lollipop Plot" examples of methylation sequencing output. Each plot shows five methylation sequence tags (rows) for eight CpG loci (columns) for a particular gene target. An open circle is unmethylated locus, closed circle is methylated. A single tag represents a single allele from a sample, all alleles are clonally related by virtue of coming from the same glandular unit. For each plot, the unique methylation sequence number, average pairwise distance (PWD) comparison between the five alleles and methylation density can be calculated. The left column represents methylation patterns which are more homogenous, less diverse and subsequently have lower PWD and thus identify a mitotically "younger" population of cells. The right sided patterns are diverse with more unique sequences and greater PWD, therefore far more mitosis has had to have occurred to reconcile these findings and the population is "older". Rows "A" and "B" represent patterns with lower or greater methylation density respectively and are a function of chronological age as methylation generally increases with time through drift.*

1.5.4 Measuring stem cell dynamics using methylation patterns

Study of methylation pattern data within individual crypts reveals the stem cell dynamics and numbers that maintain the colon crypt. Long lived stem cells are the ancestors of all the differentiated cells within the crypt[288]. They reside in a *niche* (situated at the base of the colon crypt and in the neck of the Barrett's gland) where the stem cell phenotype and behaviour is reliant on the influences of the surrounding microenvironment from differentiating epithelial cells and paracrine signalling from mesenchymal cells[288, 289]. This is in contrast to stem cells autonomously maintaining their "stemness" through gene expression regardless of their spatial orientation in the crypt/gland[290]. The small pool of stem cells is equipotent and follows a stochastic model of cell division whereby there can be loss of individual clones from the niche which is compensated by expansion of an alternative clone to fill the space left[291]. Stem cells that migrate out of the niche's influences during symmetrical division commit to cellular and tissue differentiation with their clonal ancestry subsequently being lost as they are shed into the lumen[288]. The loss of individual stem cell clones by chance is termed *neutral drift* and over time results in monoclonal conversion of crypts/glands whereby a single original stem cell clone populates the entire niche and gives rise to all progenies. This evolutionary bottleneck is termed *niche succession* and, through comparison of methylation pattern analysis against a computational model, has been estimated to occur every 8.2 years (95% CI 2.7 – 19 years) in the colon crypt[254, 278]. The clonal evolution of a single stem cell that then provides the entire crypt progeny results in homogenisation of the intracrypt methylation pattern with an epi(genomic) diversity reset to re-accumulate over time until the next succession[292].

This theory contrasts with the *immortalised* stem cell theory where asymmetrical division of stem cells is mandated[293]. If this were to be true neutral drift (loss of a particular stem cell clone) would never occur with multi-lineage stem cell persistence. With this, intracrypt methylation pattern diversity would increase over time through successive stem cell generations[292]. In addition, any neutral somatic mutations within individual stem cells would never be lost leading to significant intracrypt somatic heterogeneity. Eventually a particular mutation combination may

arise precipitating a new selection advantage incorporating prior neutral mutations and transition to the malignant phenotype. In the *niche succession* model where one stem cell comes to dominate a crypt, progression to malignancy is therefore more protracted through chance due to the attrition of potentially risky stem cell clones and limits to crypt diversity[294]. Even though immortalisation is now refuted, in the same manner, longer times to achieve niche succession maintains crypt (epi)genomic diversity and progression risk[254, 295].

Because the life span of the differentiated crypt/gland cells is limited (5-7 days) they cannot collect sufficient (epi)genetic aberrations to promote tumorigenesis before being shed into the lumen, cancer therefore arises from the stem cell niche[288]. "Intracrypt distance" analysis of the epiallele sequences gives a reflection of stem cell division since the most recent common ancestor cell, which will be the last niche succession event[274]. When comparing methylation pattern data from colorectal cancers an estimation of 4-1,024 stem cells per crypt niche has been determined combined with an estimated mitotic age of cancer ranging from 250-1,130 divisions since transformation[282]. Gabbutt et al. refined this stem cell number in normal colon to 5.8 ± 1.7 stem cells per crypt with a mean fixation (succession) time of 8.3 ± 5.5 years[296]. Intracrypt individual CpG loci patterns synchronise to either 0%, 50% or 100% methylation density across the cellular population following monoclonal expansion events (e.g. niche succession). To determine these values, the investigators compared the real-world intracrypt methylation data distribution to computational models of variable stem cell parameters (number; replacement rate; (de)methylation rate per allele)[296]. More traditional estimations of stem cell numbers is dependent on the proportions of unique sequences (and hence relative diversity of patterns) at the intracrypt level compared against intercrypt analysis: low intracrypt diversity with high intercrypt diversity would indicate few stem cells maintaining crypts compared to situations of more uniform diversity between the two sites of comparison which indicates a larger niche population[282].

An understanding of stem cell dynamics, ancestry and niche architecture is important, as the drivers of cancer appear at the level of the stem cell and

consequently confer the phenotypic crypt/gland that progresses under environmental selection[223].

1.5.5 An epigenetic drift model for Barrett's

Given Barrett's epithelium consists of a similar glandular type structure to colon epithelium, it follows that utilisation of the epigenetic drift model to infer mitotic aging, stem cell dynamics and clonal relationships is potentially transferrable. Indeed, preliminary data from our lab employing methods akin to Humphries et al.[276] on Barrett's biopsies is also suggestive of a punctuated evolution in the histogenesis of the lesion (unpublished). Not only would this help in our understanding of the biology of Barrett's discussed above but also could have a role in developing a suitable risk stratification test that thus far has remained elusive. A recent study by Curtius et al. lends further precedent in potential utilisation of epigenetics in risk stratification[297]. Here, a distinct set of 67 CpGs were identified to undergo differential age-related epigenetic drift that discriminated between BO and normal squamous epithelium. Furthermore, by combining the data of each CpG and patient demographics into a computational algorithm, much like Horvath's[269] or Hannum's[270] clock mentioned in section 1.5.2, an estimation of individual patient's BO dwell time was possible. The median age of onset was 33.6 years (range 2-59) in a cohort 30 BO patients aged 21-88 years (mean 63.4). There was wide inter-patient heterogeneity of the BO segment age, but this model stopped short of attributing risk as comparisons with dysplastic BO or OAC were not made. It therefore remains an unknown whether harbouring the BO segment for longer is associated with progression or not.

1.6 Summary of introduction

This introduction has demonstrated BO to be a poly-phenotypic, polyclonal, heterogenous and pre-malignant metaplasia at all grades of histology. There is vast divergence of the underlying (epi)genomic milieu both within and between patients such that ascribing risk based on traditional identification of mutated canonical cancer driver genes is obsolete. The search for alternative novel risk stratification indices has pushed research towards in-depth next generation sequencing and multi-omic analyses with an explosion of such studies over the past 2-3 decades.

The Fitzgerald group, based out of Cambridge, UK, have just published a recent multi-omic cross-sectional study which nicely confirms prior work, encapsulates and summarises the landscape of Barrett's[243]. Here, indolent non-dysplastic BO demonstrates high SNV mutation burden, at similar frequency to the latter dysplastic phenotype, that is important at inception and maintenance but does not have a bearing on progression. Crucially, large scale structural variants, chromosomal rearrangements and genomic catastrophe are distributed as a continuum along the IM-Dysplasia-OAC sequence and drive its progression. In contribution, there is correlative increase of abnormal methylation and transcription along this sequence with downregulation of key metabolic pathways, and upregulation of genes affecting cell cycle checkpoints, DNA repair and chromosomal stability. The determinant of histological phenotype is late and with variable manifestation that belies the underlying (epi)genomic aberrations. In some cases, a loss of intestinal metaplasia is seen at the point of transition, in another subtype there is upscaling of inflammatory markers within the tumour microenvironment. Key tumourigenesis events are documented including CDKN2A loss at inception, fragile site (FHIT / WWOX) structural variations, *TP53* mutation promoting genomic instability, excessive ERBB2 copy number and whole genome doubling seen from dysplasia onwards. A new finding of LINE1 retrotransposon activity that adds to the genomic chaos in erroneously affecting gene structure and expression has also been described[243].

Conclusively though, no one single gene aberration in the clonal mosaic of BO defines the stepwise phenotypic or pathological grading that is seen clinically.

Adapting and translating indices of ecological diversity carries the promise however of providing this discrimination. In conjunction with the measurability and greater (epi)genomic understanding afforded by next generation sequencing, lesions that are "born-to-be-bad", that suffer catastrophic events or gradually accumulate diverse aberrancy can be detected confidently[298]. Conceptually it is straightforward to appreciate that the arrival of CIN, for example, will often cause cancer. But this is the end point of OAC development, therefore our window to intervene clinically is much reduced. It is reassuring other diversity measures at earlier timepoints could act as a proxy, although there are no reported prospective studies (yet) that demonstrate a reduction in mortality by knowing this. Given the low chance of progression to OAC and relative high prevalence of BO in the community that will only increase if screening is approved[52], it is prudent to shift focus to fully understand the early evolution of the lesion.

We do not fully understand the cellular origins or how BO expands to fill the inflamed oesophagus at inception with a resulting mosaic of geno-phenotypic clones. Crucially for this thesis, we do not know the true onset and timing of the condition and whether a longer dwell time and mitotic age is a factor in progression to cancer. We have seen how the clonal diversity and dynamics of colorectal adenomas and cancer can be elucidated through methylation sequencing. Developing an understanding of the physiological and pathological factors and the influence of intrinsic and extrinsic evolutionary drivers to cancer is necessary to ultimately enhance clinical algorithms in the surveillance, management and treatment of BO and OAC.

2 Hypotheses

That a mitotic model of tissue aging in Barrett's demonstrates that patients who progress to cancer have an "older" and more diverse methylation pattern that predicts their progression and that the origin of Barrett's is polyclonal with evolution characterised by punctuation.

That the process of clonal evolution can be predicted using a mitotic clock approach and that altered stem cell dynamics occurs in patients that will eventually progress to dysplasia.

3 Aims

Establish a longitudinal prospective tissue bank of fresh frozen biopsy samples from patients with Barrett's oesophagus who progress and who do not progress to cancer.

Design and optimise a novel sequencing protocol for high resolution, high coverage, targeted methylome analysis to reveal the ancestral histories of cellular populations and establish a mitotic clock.

Use the mitotic clock model to time the histogenesis of Barrett's oesophagus in patients and whether dwell time is a risk factor for progression to cancer.

Add to the debate regarding the polyclonal origins of Barrett's and use methylation analysis to infer the cellular origins of the lesion both from a timing and somatic inheritance perspective.

Characterise the evolutionary dynamics, ordering and rate of expansion of individual glandular phenotypes both on an intra and intergland perspective.

4 Materials and Methods

4.1 Tissue acquisition

4.1.1 Prospective fresh frozen tissue collection

Biopsy material was obtained prospectively from Royal London Hospital (RLH), Whitechapel, London, UK from patients who are undergoing routine Barrett's surveillance. Research biopsy samples were taken distally to proximally, commencing at the gastric cardia, through the Barrett's segment to the squamous epithelium. 4-6 individual samples are taken in total depending on the segment length with at least 2 samples at different levels from the BO itself. The distance of biopsies as measured from the patient's incisor teeth were recorded. Gastric cardia biopsies are taken 1-2cm distal to the anatomical GOJ, squamous epithelium biopsies are taken 1-2cm proximal to the SCJ. Biopsies were snap frozen with Cellpath^{TM} Cryospray (Fisher Scientific, Loughborough, UK), placed on dry ice for transportation, indexed and subsequently stored at -80^{o}C for later processing. A further distal biopsy was also taken and stored immediately in formalin for an additional project within our lab, some prior formalin samples we however to be clear, work that formulates this thesis was undertaken in fresh frozen tissue.

Patient age, sex, Prague criteria, coincidental endoscopic findings and clinical histopathological grading both current and historic were taken. These latter details consign the patient into the cohorts of non-progressors or progressors (to HGC/IMC/OAC) and their individualised duration of such designation, this includes retrospective endoscopy and histopathological data that precedes our initial date of tissue collection commencement. Where the patient presents with IFD or LGD this is also recorded but for the purposes of this thesis, if this is the patient's highest (worst) level of grading over the duration of surveillance, then they are classed as non-progressors. This is in-line with many published studies in the literature.

Nevertheless, some breakdown of these patients is presented later in thesis in efforts to try and understand this tricky histological half-way house dynamic[3].

All patients provided informed written consent and appropriate ethical approval has been obtained from the Research Ethics Committee (REC) for the collection of this material and data (REC numbers: 11/LO/1613 & 15/LO/2127, Stanmore REC and East London REC respectively).

4.1.2 Archival prospective fresh frozen tissue collection, obtained for this thesis retrospectively

As progression in BO is uncommon[32], to enrich the cohort of patients with dysplasia and/or OAC, a collaboration has been established with Professor Laurence Lovat of University College Hospital (UCH), London, UK. Typically, 3-5 research biopsies were taken that included at least one sample from each of the distal and proximal oesophagus and the squamous oesophagus. At index timepoint, samples were taken prior to the first treatment. Samples were snap frozen in liquid nitrogen with continued storage in liquid nitrogen Dewars at UCH. These samples were transferred to our stewardship in November 2018 under a Material Transfer Agreement (MTA) in November 2018 stored in the Cryostorage Facility at Bart's Cancer Institute. Demographic data of age, sex, subsequent or prior treatment modality and Prague criteria was also kept alongside a record of the biopsy site location details.

4.2 Tissue Processing

4.2.1 Preparation onto microscopy slides

Tissue processing onto microscopy slides was undertaken by the BCI Pathology service on my behalf. In brief, each biopsy tissue sample is embedded first into optimal cutting temperature (OCT) compound prior to sectioning on the microtome-cryostat. A mixture of laser capture membrane and regular microscope slides were used. PALM Laser Capture Membrane Slides 1.0 PEN (Carl Zeiss GmbH, Göttingen, Germany) are treated by exposure to UV light at 254nm wavelength for 30 minutes. All serial sections of each embedded sample were taken at 10 µm thickness cut onto PALM slides and then charged glass slides (Colorfrost^{TM} microscopy slides, Fisher Scientific, UK) for immunohistochemistry (IHC). Prior to the first and after the third PALM slide, two sections each were taken for H&E staining for reference purposes as macroscopic glandular appearances are well demonstrated with this technique, especially in occasionally degraded fresh frozen sections which tend not to stain as readily with cytochrome c oxidase (CCO), described in section 4.2.2. Each PALM slide can fit 4 individual sections, thus 12 were taken in series across three slides. Two sections were taken for each IHC slide. All PALM and IHC sections were immediately stored at -80^{o}C. H&E reference slides are stored at room temperature. **Figure 4.1** demonstrates this sectioning protocol schematically as a Z-stack of the ten total slides per fresh frozen biopsy.

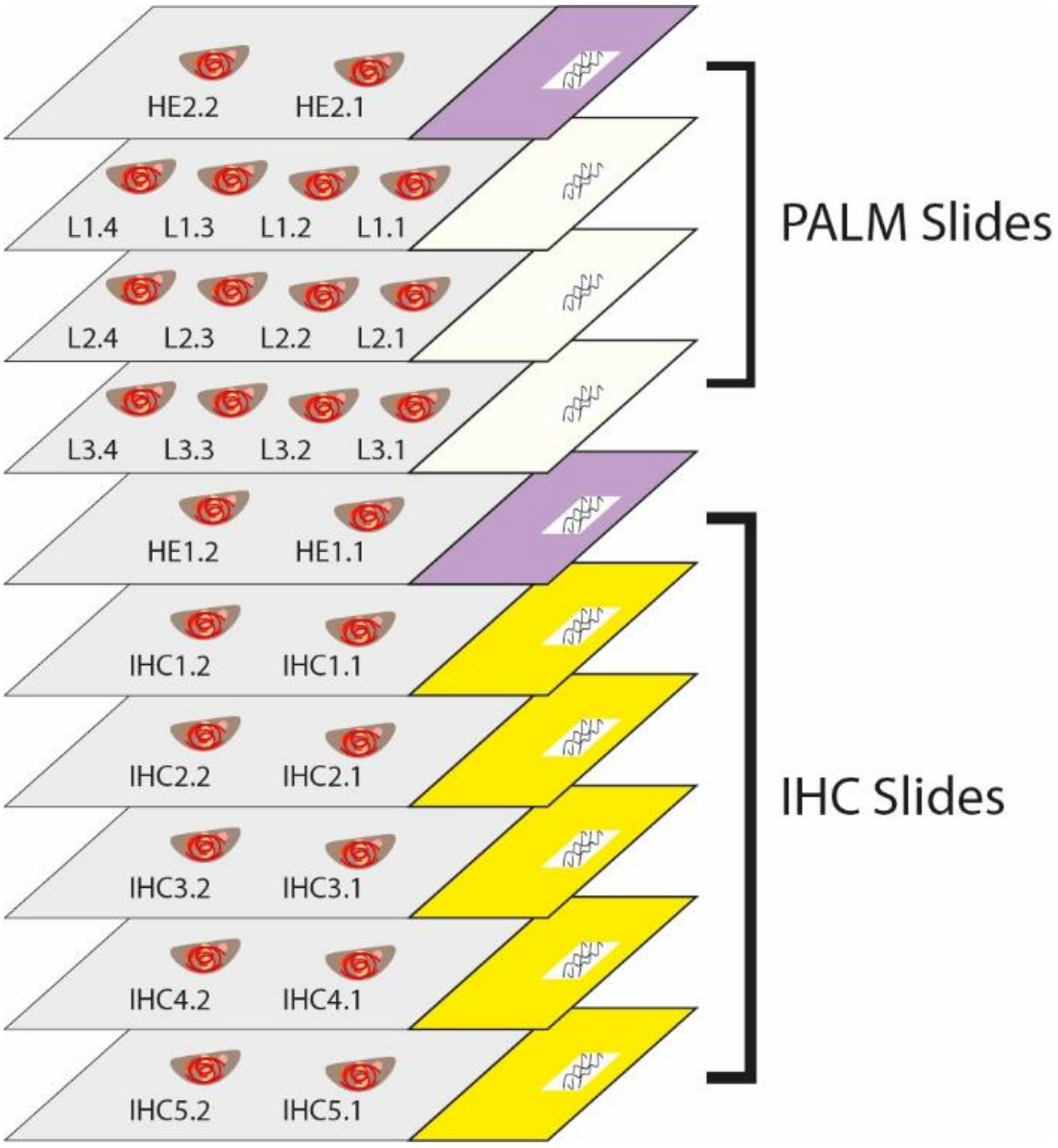

***Figure 4.1**: Representative Z-Stack of protocol for sectioning fresh frozen biopsies. Purple headed slides represent Haematoxylin & Eosin (H&E) stained slides; Cream headers are laser capture microdissection (PALM) slides; yellow slides are for immunohistochemistry (IHC). In total, 26 sections at 10µM thickness are taken in series from each frozen biopsy. Each section is coded with the slide type (HE, L, IHC), slide number and location on slide. H&E slides are stored for reference. PALM and IHC slides are immediately stored at -80°C to preserve tissue integrity.*

4.2.2 Dual cytochrome *c* oxidase (CCO) / succinate dehydrogenase (SDH) histochemistry

To identify clonal units and to provide adequate differentiation of the glandular architecture of Barrett's glands against the background stroma, LCM sections were subjected to dual CCO/SDH histochemistry. CCO is encoded by the *CCO1* gene in mtDNA, mutations to this gene can cause deficiency of this protein however this confers no selective advantage or disadvantage to the cell[299]. Mutations in mtDNA can been used as a marker of cellular clonality in lineage tracing studies[300], therefore the added benefit of this protocol alongside gland visualisation is to identify potential clonal populations of Barrett's glands. When differential staining patterns occur, the mtDNA is extracted, amplified with polymerase chain reaction (PCR) and the mitochondrial genome can be interrogated through Sanger sequencing[299]. When available, this will add further depth to the clonality assessments made through methylation sequencing.

Sections on PALM slides were air-dried for 1 hour at room temperature. CCO medium was prepared containing 100 mmol/l cytochrome *c*, 20 µg/ml catalase and 4mmol/l diaminobenzidine tetrahydrochloride in 0.2 mol/l phosphate buffer, pH 7.0 (all products sourced from Sigma Aldrich, Gillingham, UK). 50-200 µl medium was pipetted over each section and incubated at 37^{o}C for up to 30 minutes. When the desired level of staining was obtained, sections were then washed in phosphate buffered saline (PBS), pH 7.4, three times for 3 minutes.

Glands or cell populations with normal CCO activity stain brown. In order to identify CCO-deficient glands/cells all slides are subjected to SDH histochemistry after the PBS washing step above. SDH medium was prepared containing 130 mmol/l sodium succinate, 200 mmol/l phenazine methosulfate, 1 mmol/l sodium azide and 1.5 mmol/l nitroblue tetrazolium in 0.2 mol/l phosphate buffer, pH 7.0). Sections were incubated in 50-200 µl SDH medium at 37^{o}C for up to 45 minutes, ensuring that the entire section is covered with the medium. A separate spare tissue sample that skips the CCO medium incubation step is used as a positive control to confirm staining activity and depth of colour for the SDH medium. All sections are washed again in PBS three times for 3 minutes and dehydrated in a graded ethanol series (70%, 95%,

100%, 100%) and allowed to air-dry for 1 hour prior to cutting on the LCM microscope. Any sections/slides not for immediate use are stored back at -80°C.

With CCO staining there is an underlying biological threshold effect that determines whether the gland will appear brown, blue or perhaps even a combination which is driven by the degree of cellular mitochondrial *heteroplasmy*[301]. Heteroplasmy is where there is a mixture of mutated and wild-type mitochondrial genomes within the cell. Each mitochondria contains its own copy of the 16.5kb mtDNA genome[302]. Through genetic drift a particular mtDNA mutation can come to dominate the proportion of mitochondria and lead to a *homoplasmic* conversion where all copies of the genome within the cell are the same. However, the threshold of brown stain to blue stain is achieved in the heteroplasmic state, estimated to be when ~80% of mtDNA genomes are deficient[303]. This is correlated with chronological age that permits approximate time to accumulate such mutations, in fact in colonic studies, very few CCO-deficient crypts are seen before the age of 40 years[131]. These mutations are inherited during cell division to both daughter cells. The *CCO1* gene mutation is important to allow IHC differential detection between clonally related populations but the real interest resides in examining the remaining genome through Sanger sequencing (described in **section 4.3.7**) to reveal shared or discrepant mutations that can define the ancestry, lineage and dynamics of that population[300, 304].

4.2.3 Double immunohistochemistry

Frozen sections were cut from Barrett's biopsy specimens as described previously and stored at -80°C. Slides were thawed at room temperature in a chamber for 15 minutes. Sections were then fixed in 4% paraformaldehyde (Sigma-Aldrich, UK) for five minutes at room temperature then rinsed in PBS with 0.1% Tween 20 (Sigma-Aldrich, UK). For any fixed formalin paraffin embedded (FFPE) samples an alternative preparation is required with initial de-waxing with three washes of xylene for five minutes each and rehydration through a graded ethanol series to water. Antigen retrieval for FFPE is performed by adding the slides to a boiling 0.01M solution of

sodium citrate buffer (Sigma-Aldrich, UK) (pH 6.0) and microwaving for ten minutes. Sections were then permeabilised through incubation in PBS containing 0.1% Tween 20 (Sigma-Aldrich, UK) for ten minutes then blocked for endogenous peroxidase activity in 3% hydrogen peroxide (VWR International, Leicestershire, UK) for 15 minutes. Sections were then rinsed three times in PBS with 0.1% Tween 20 (Sigma-Aldrich, UK) for three minutes each. Sections were then antigen blocked in serum free protein blocking solution (DAKO, Cambridge, UK) for an hour with a subsequent incubation with horse serum blocking solution (Sigma-Aldrich, UK) for a further hour, sections were not washed between the two blockings.

All antibodies were diluted in PBS with 5% donkey serum (Invitrogen, Paisley, UK) and a negative control where no primary antibody was applied was used in all experiments. Primary antibodies were applied to each section on the slides (**Fig. 4.2**). To differentiate cardiac from specialized lineage glands, mouse anti-human MUC5AC (dilution 1:500 for frozen sections) for gastric foveolar cells and mouse anti-human MUC2 (dilution 1:1000 for frozen sections) for intestinal goblet cells (Novacasta, UK), antibodies were used respectively. For FFPE antibody dilutions see **table 4.1**. Sections are incubated overnight at 4^{o}C in a cold room or fridge.

Sections were then washed for 3x5 minutes in PBS with 0.1% Tween 20 (Sigma-Aldrich, UK) followed by 15 minutes incubation with a polyclonal rabbit anti-mouse IgG secondary antibody conjugated to biotin (DAKO, Cambridge, UK) and streptavidin conjugated horse radish peroxidase (HRP) (DAKO, UK). A brown reaction product was developed for all sections using a solution containing 0.6 mg/ml 3,3'diamino-benzidine (DAB) with 0.03% (v/v) H_2O_2 (Dako, UK) and sections were counterstained with Gill's haematoxylin (Pioneer research Chemicals, UK) before dehydration through increasing concentrations of ethanol (70%, 90%, 100% and 100%, one minute each), and mounting with Vector Hardset™ mounting media (Vector laboratories, CA, USA).

On review of the sections and their respective MUC5AC or MUC2 staining patterns a call can be made on whether a particular gland represents a gastric or intestinal phenotype[305]. IHC staining was corroborated on assessment by an expert

gastrointestinal pathologist (Dr. Marnix Jansen, Dr Joanne Chin-Aleong or Professor Sir Nicholas Wright).

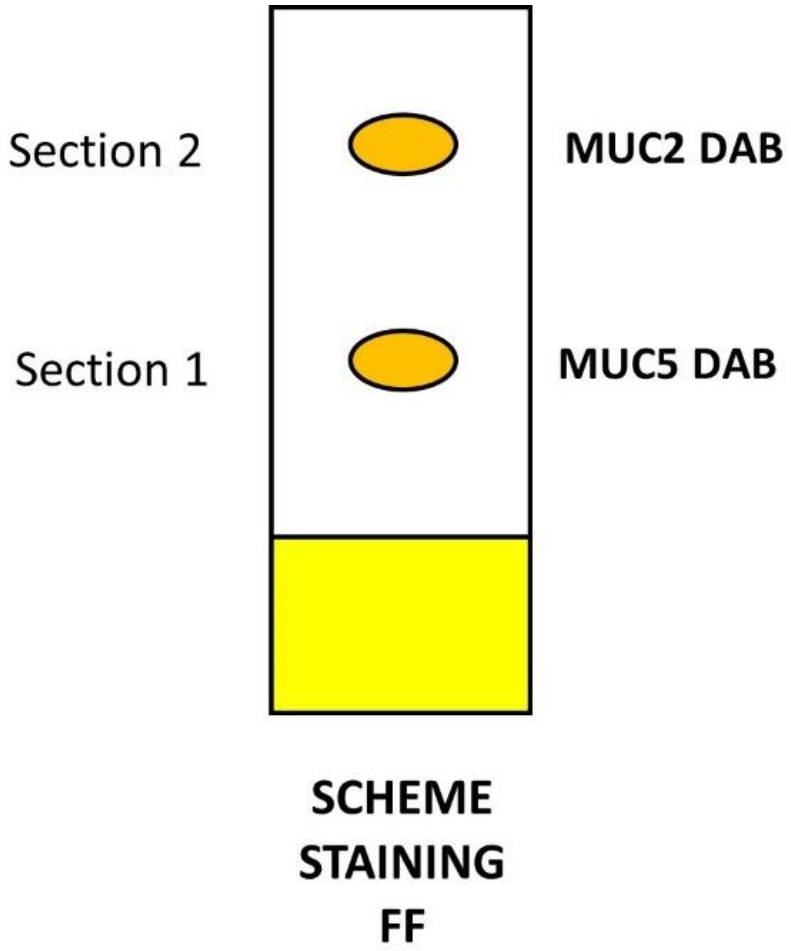

***Figure 4.2**: Schematic of double immunohistochemistry. The two side-by-side serial sections are subjected to either Mucin 2 (MUC2) DAB (3,3' diaminobenzidine tetrahydrochloride) or MUC5AC DAB IHC. MUC2 identifies intestinal goblet cells, MUC5AC gastric foveolar cells. The benefit of double IHC is in the down-stream work-flow of having essentially identical adjacent serial sections to easily trace individual glands when calling their phenotype.*

Antibody	Host species and details	Dilution for FFPE	Usual antigen retrieval for FFPE samples	Source
MUC 2	Mouse polyclonal	1:100	10 minute microwave in 0.01M sodium citrate	Abcam, Cambridge, UK
MUC 5AC	Mouse monoclonal NCL-MUC-5AC, clone 2b4, IgG1	1:50	10 minute microwave in 0.01M sodium citrate	Novacastra, Newcastle UK
MUC 6	Mouse monoclonal NCL-MUC-6 Clone CLH5, IgG1	1:25	10 minute microwave in 0.01M sodium citrate	Novacastra, Newcastle UK
H^+/K^+ ATPase	Mouse monoclonal clone 2G11	1:200	10 minute microwave in 0.01M sodium citrate	Invitrogen, Paisley, UK
α 5 Defensin	Mouse monoclonal [8c8] to alpha defensin	1:1000	10 minute microwave in 0.01M sodium citrate	Abcam, Cambridge, UK
Pepsinogen	Mouse monoclonal [7G3] to pepsinogen I	1:50	10 minute microwave in 0.01M sodium citrate	Abcam, Cambridge, UK
Secondary antibody IgG biotin complex	Polyclonal rabbit anti-mouse	1:300	Applied as secondary layer	DAKO, Cambridge, UK
Streptaviadin conjugated HRP	Goat anti rabbit IgG	1:500	Applied as tertiary layer	DAKO, Cambridge, UK

Table 4.1: *Details the common antibodies used in our lab for immunohistochemistry with focus on the Barrett's glandular phenotype as discussed in* ***section 1.3.1****. The dilutions listed are for fixed formalin paraffin embedded (FFPE) samples that usually require more concentrated solutions and the need for an additional antigen retrieval step due to greater antigen degradation over their fresh frozen counterparts. MUC – Mucin; IgG – immunoglobulin G; HRP – Horse radish peroxidase; H+/K+ ATPase – Hydrogen/Potassium adenosine triphosphatase.*

4.2.4 Imaging

All slides (H&E, PALM, IHC) were scanned using the NanoZoomer Digital Slide Scanner S60 or S210 model (Hamamatsu Photonics, Welwyn Garden City, UK) creating a digital image for storage, reference, planning and to allow accurate measurements for spatial orientation of glands where necessary using the NDP.view2 viewer software package (Hamamatsu Photonics, UK).

4.2.5 Laser Capture Microdissection (LCM)

All tissue sections were assessed for Barrett's metaplasia and presence of dysplasia and confirmed by an expert gastrointestinal histopathologist appropriately blinded to the clinical information. Gland and tissue morphology was examined on H&E reference slides to identify suitable glands and/or regions (for example whole epithelium or stroma) for LCM. The same individual glands and/or regions were identified and traced across the serial sections on the PALM membrane slides. Using the PALM Laser Microdissection System, an area was defined, the tissue is cut and catapulted into the cap of 500 µl AdhesiveCap Opaque tubes (Carl Zeiss, Germany) which have been prior treated in a UV hood to reduce occult DNA contamination. Pre and post cut images of each section and adhesive cap were taken to later reference and proof of capture. 15ul of Buffer ATL (QIAamp® DNA Micro kit, Qiagen, Manchester, UK) was carefully pipetted onto the adhesive cap to re-suspend the micro-dissected tissues prior to closing the tube. The tube was placed on ice to await DNA extraction while the further LCM was ongoing for additional gland cutting, usually in batches of 12-24 glands during a session.

4.2.6 Total DNA extraction of single glands or cells

For the microdissected single glands in 15ul Buffer ATL, these were first centrifuged at 8000 rpm for one minute. The adhesive cap was then re-inspected under a magnifying lamp to ensure the sample had detached into the tube solution. Where there was remnant tissue on the cap (seen as miniscule brown dots) a small aliquot,

5 µl, of the tube solution was pipetted and used to release the tissue. This is important as the subsequent targeted methylation work-flow (see results **chapter 5**) is highly toxic and degrading to DNA. Thus, DNA yield at this stage is paramount especially from microdissected samples.

Commercially available DNA kits were then used to extract the DNA. For the LCM Barrett's glandular tissues the QIAamp® DNA Micro kit (Qiagen, Manchester, UK) was used. In brief, 10 µl of proteinase K was added to the 15 µl sample solution and pulse vortexed for 15s. These were incubated in a thermomixer at 56°C at 500 rpm for at least 4 hours or overnight sealed with Parafilm™ (Fisher Scientific, UK). Centrifuge at 8000 rpm for one minute. 25 µl Buffer ATL and 50 µl Buffer AL is added and mixed to 15s. 50 µl 100% ethanol is added and mixed followed by incubation for 5 minutes at room temperature. Centrifuge briefly. The sample is transferred to QIAamp MinElute columns (Qiagen, Manchester, UK) and centrifuged for 1 minute at 8000 rpm. Two washing steps, first with Buffer AW1 and second with Buffer AW2 are undertaken with centrifugation between at 8000 rpm for 1 minute and discarding of the follow-through. The column membrane is dried with a final centrifugation at 14,000 rpm for 3 minutes. 25 µl Buffer AE is applied directly to the column membrane and incubated at room temperature for 5 minutes. Centrifuge at 14,000 rpm for 1 minute to elute samples. A second elution step was undertaken with the same 25 µl passing through the column again to increase DNA yield.

Where Sanger sequencing was to be performed on single cells or glands, total DNA extraction was carried out using the following method with all reagents mentioned below being sourced from Arcturus® Picopure® DNA extraction kit (Thermo-Fisher, UK). Extraction solution was composed by the addition of 155 µl Picopure® buffer to a pre-prepared vial from Picopure® kit containing Proteinase K and kept on ice. After tissue had been dissected into an LCM cap, 12 µl of the Picopure® extraction solution was immediately added to added and then inverted to ensure mixing. Samples were then immediately placed on ice, then briefly spun down at 2000rpm for fifteen seconds. The samples were then incubated at 65°C for three hours, then centrifuged at 1000rpm for one minute. The sample was then warmed on a heat block at 95°C

for ten minutes to inactivate the Proteinase K then allowed to cool to room temperature. No further purification steps were necessary.

4.3 Nucleotide Analysis

4.3.1 Bisulfite conversion of genomic DNA (gDNA)

It is not possible to directly determine the consecutive sequence of methylated CpG loci without first undergoing bisulfite treatment or enzymatic conversion. Bisulfite converts all unmethylated cytosine within the genome to uracil through deamination. Methylated cytosine remains unchanged. Subsequent downstream PCR substitutes uracil for thymine thereby ascertaining single-nucleotide resolution of differential methylation of CpG loci.

All bisulfite conversion was carried out as per manufacturers protocol using the EpiTect® Fast DNA Bisulfite kit (Qiagen, UK). In brief, 85 µl bisulfite solution and 15 µl DNA protect buffer was combined with the 40 µl DNA sample, vortexed and subjected to 2 cycles of 95^{o}C for 5 minutes, 60^{o}C for 10 minutes for bisulfite conversion in a Tetrad 2 thermocycler (Bio-Rad, Hercules, CA) then cooled to 20^{o}C.

Samples were briefly centrifuged prior to adding carrier RNA at a concentration of 10 µg/ml. 310 µl of freshly prepared buffer is added to the post bisulfite-converted DNA (bDNA) followed by 250 µl 100% ethanol with vortexing and centrifugation between each reagent addition. The mixture was transferred to a MinElute DNA Spin column, centrifuged at maximum speed for 1 minute then washed and desulphonated with 500 µl each of Buffer BW and Buffer BD respectively with additional centrifugation between. Two further washes and centrifugation with 500 µl Buffer BW followed by 250 µl ethanol were then applied. The DNA sample was eluted in 15 µl nuclease-free-water of which this was passed through twice to increase yield without unnecessarily diluting the sample.

4.3.2 Enzymatic conversion of genomic DNA

In July 2019, New England Biolabs® Incorporated (Massachusetts, USA) released a new product available to UK purchasers that uses enzymatic conversion of DNA rather than bisulfite modification. This product, called NEBNext® Enzymatic Methyl-

seq (EM) Conversion Module, was subsequently trialled and then utilised thereafter for the sequencing technique set-up. This included comparing PCR product outputs between enzymatic and bisulfite converted gDNA templates (see results chapter 5).

Enzymatic conversion of gDNA (herein for this thesis called eDNA) is a two-step process that detects methylated cytosine. Step one involves protection from down-stream deamination of 5-methylcytosine (5mC) through its oxidation catalysed by Tet methylcytosine dioxygenase 2 (TET2). 5mC is oxidised to through sequential steps ultimately to 5-carboxycytosine (5caC) which prevents its deamination. Alternatively, 5mC bases are also protected from deamination through glycosylation by the incorporation of their proprietary *Oxidation enhancer* reagent. Unmethylated cytosine bases are not oxidated or glycosylated in this fashion. The second step involves enzymatic deamination of unmethylated cytosine to uracil catalysed by Apolopioprotein B mRNA Editing Catalytic Polypetide-like (APOBEC) enzyme, which subsequently becomes thymine following PCR processing much like the bisulfite conversion.

The laboratory protocol is as follows. The TET2 Reaction Buffer Supplement is re-formulated from a powder by vortexing in 400 µl of TET2 Reaction Buffer, this is stored. gDNA samples are normalised to a volume of 28 µl in either water or 10 µM Tris (pH 8.0) in a 200 µl PCR tube, to each, a master mix of 10 µl reconstituted TET2 Reaction Buffer and Supplement (above), 1 µl Oxidation supplement, 1 µl dithiothreitol (DTT), 1 ul Oxidation enhancer and 4 µl of TET2 enzyme is added per sample. Samples are vortexed and centrifuged briefly. 5 µl of 400 nM Fe^{2+} is added to the 45 µl samples, this co-factor initiates the oxidation reaction. Samples are incubated at 37^{O}C for 1 hour on the Tetrad 2 Thermocycler (Bio-Rad, CA). Samples of transferred to ice where 1 µl of Stop Reagent is added to each followed by a further 30 minutes on the Tetrad 2 Thermocycler (Bio-Rad, CA) at 37^{O}C. This completes the first TET2 step with 5mC now oxidised to 5caC or glycosylated and thus protected from deamination. The samples are purified using an AMPure XP bead (Beckman Coulter, Brea, CA) based clean-up protocol, described in **section 4.3.5**. Specifically for the EM conversion technique, 90 µl of beads are mixed with the 51 µl samples with the final elution performed using 17 µl of elution buffer. 16 µl of the

supernatant is transferred to a new 200 µl PCR tube with 4 ul of Formamide (Sigma-Aldrich, UK), vortexed and incubated at 85^{o}C for 10 minutes on the thermocycler. They are immediately placed on ice and a master mix of 68 µl nuclease-free-water, 10 µl APOBEC Reaction Buffer, 1 µl BSA and 1 µl APOBEC enzyme per sample is added, vortexed, centrifuged briefly and incubated at 37^{o}C for 4 hours on the Tetrad 2 Thermocycler (Bio-Rad, CA) to deaminate the oxidised/glycosylated DNA. Samples are then cleaned-up again using AMPure XP beads (Beckman Coulter, Brea, CA), specifically, 100 µl are added to the 100 µl post-deamination reaction sample. Samples are eluted in 21 µl of elution buffer with 20 µl subsequently transferred to individually labelled storage Eppendorf vials. eDNA samples were stored at -20^{o}C until their use in downstream processing.

4.3.3 Sample quantification

All quantification steps within this thesis were performed using a Qubit 4 Fluorometer (Invitrogen, by Thermofisher Scientific, UK) and a Nanodrop One (Thermofisher Scientific, UK) spectrophotometer was used to calculate the 260nm:280nm ($A_{260/280}$) ratio as a marker of sample purity aiming for values >1.8. Two readings are taken for each gDNA sample with the average then forming the genome copy number within the sample used throughout the subsequent bisulfite or enzymatic conversion process, library preparation, sequencing, bioinformatics and statistical analysis. Prior to next generation sequencing (NGS) all prepared libraries were subjected to quality control analysis and quantification using High Sensitivity (HS) D1000 Screentape® on the 4200 TapeStation System according to the manufacturers protocol (both sourced from Agilent Technologies, Santa Clara, CA). The TapeStation Analysis software package (available from **https://www.agilent.com/**) was used for this purpose and to visualise the resultant PCR products and distribution across the electropherogram. A desired target amplicon size region between the bounds of 450-900 bp was set to optimise the calculated library concentration in pg/µl necessary for equimolar pooling and sequencing read assignment. The software package also gave an indication of the proportion of product that fell within these bounds and thus reflected the degree of

off target amplification or failure that may or may not have occurred of particular samples and whether the particular gland/sample was suitable to be included for onward sequencing.

4.3.4 Polymerase Chain Reaction (PCR)

Various PCR protocols were utilised during this thesis including the set up and optimisation of a novel PCR protocol forming the basis of the first results chapter. Typically, PCR reactions varied in volume of 20-50 µl but primarily involved the use of a commercially available master mix containing polymerase, $MgCl_2$ and deoxynucleotide triphosphates (dNTPs) within a proprietary buffer, this was combined with sequence specific primers from a 10 µM stock concentration, DNA template and nuclease-free-water. Additions of further $MgCl_2$ solution, dimethyl sulfoxide (DMSO) and Q Solution (Qiagen, UK) were trialled and optimised where necessary. Each reaction was subjected to a specific thermocycling protocol which is detailed in the relevant sections in results chapter 5. All reactions were set up on ice, mixed by vortexing and centrifuged prior to commencement on the Tetrad 2 thermocycler (Bio-rad, CA). PCR products were run through a 1.5% agarose gel (Bioline, Nottingham, UK) containing Gel Red nucleic acid stain (Biotium, Fremont, CA) and demarcated using 100bp or 1kb HyperLadder™ (Bioline, UK). See **section 4.3.6** below for full agarose gel details. Bands were visualised on an Amersham Imager 600 (GE Healthcare, Chicago, IL) and photographed.

4.3.5 Bead-based clean-up of PCR products

Purification of post-PCR products was performed to remove contaminants such as dNTPs, salts, primers and primer dimers prior to library preparation and sequencing. Bead-based purification was performed using AMPure XP beads (Beckman Coulter, Brea, CA). The beads were removed from the fridge at least 30 minutes prior to use to ensure normalisation to room temperature. The ratio of bead volume to PCR volume determines the size selection of amplicon product that will bind to the beads

and this is stated in the relevant section in results chapter 5. Therefore, a proportional volume of beads was added to completed PCR reactions. This was mixed by pipette and briefly centrifuged. The reaction was incubated at room temperature for 10 minutes to allow binding of DNA fragments to the beads. The tube was placed on a magnetic stand for 5 minutes to separate the beads from the supernatant which was then removed and discarded. Freshly prepared 200 µl 80% ethanol is added and incubated for 30 seconds at room temperature to wash the beads before being removed and discarded again. This step is repeated. Excess remaining fluid is carefully pipetted away and the beads are left to dry for maximum of 3 minutes or until they become dull. A minimum of 10 µl nuclease-free-water is added and the tube is removed off the magnetic stand, vortexed and briefly centrifuged (<1 second) to resuspend the beads. Incubation for 10 minutes at room temperature allows the DNA to be eluted off the beads. The beads are then separated from the supernatant by placing back on the stand for 5 minutes. The supernatant contains the purified target PCR product(s) and is removed and stored at -20^{o}C for later processing.

4.3.6 Mitochondrial DNA polymerase chain reaction for Sanger sequencing

Full primer sequences (Sigma-Aldrich, UK) spanning the entire mitochondrial genome are listed in the **supplementary table S1**. PCR of mtDNA was performed utilising a nested protocol where nine individual first-round PCR reactions were performed to generate overlapping 2kb regions spanning the entire mitochondrial genome. A subsequent second-round PCR was then performed which split the first-round amplicons into four 500bp amplicons.

Tissue was dissected and DNA extracted as per above. All reagents (except those containing DNA or polymerase) were subjected to UV irradiation for forty-five minutes and the entire 1st round PCR was performed in a UV cabinet. In detail; 1µl of template DNA was added to 49µl of first round PCR reaction mix containing; 10mM Tris-HCl PCR buffer (pH 8.3) (Thermo-Fisher, UK), 0.2mM dNTPs (Roche, UK), 1.5mM Mg^{2+} (Thermo-Fisher, UK), forward and reverse primers (0.6µM each) and 0.35U

AmpliTaq gold (Thermo-Fisher, UK), and subjected to the conditions shown in **Table 4.2** in a G-Storm thermocycler (G-Storm, Catcombe, UK).

Number of Cycles	Step	Temperature(°C)	Time
1	Denaturation	95	5 mins
38	Denaturation	94	45 secs
	Primer annealing	58	45 secs
	Extension	72	2 mins
1	Final extension	72	8 mins

Table 4.2: *PCR conditions for 1st round amplification of mtDNA*

Each second round PCR reaction was prepared on ice and utilised 36 M13-tailed primer pairs to amplify overlapping regions of the 1st round products, generating amplicons of approximately 500-600bp spanning the entire mitochondrial genome. M13 is a universal sequence that facilitates Sanger sequencing and avoids the use of multiple sequencing primers. 1µl of 1st round PCR product was added to 24µl of second round PCR reaction mix containing; 10mM Tris-HCl PCR buffer (pH 8.3) (Thermo-Fisher, UK), 0.2mM dNTPs (Roche, UK), 1.5mM Mg^{2+} (Thermo-Fisher, UK), forward and reverse primers (0.6µM each) and 0.35U AmpliTaq gold (Thermo-Fisher, UK). The mixture was then subjected to thermocycling under conditions outlined in **Table 4.3** in a G-Storm thermocycler (G-Storm, Catcombe, UK):

Number of Cycles	Step	Temperature(°C)	Time
1	Denaturation	95	10 mins
30	Denaturation	94	45 secs
	Primer annealing	58	45 secs
	Extension	72	1 min
1	Final extension	72	8 mins

Table 4.3: *PCR conditions for 2nd round amplification of mtDNA for Sanger sequencing*

Second round PCR products were electrophoresed through a 1.5% agarose gel (Bioline, Nottingham, UK) to ensure a successful amplification. Agarose gels were prepared by addition of 1g agarose (Sigma-Aldrick, UK) to 125ml of 1XTris/Acetic Acid/EDTA (TAE) solution (Severn Biotech, Worcester, UK). This was microwaved for 90 seconds until melted and clear. This was cooled until starting to become viscous at which point 1.5µl of Gel red fluorescent nucleic acid dye (Biotum, Freemont, CA) was added. Once set, the gel was placed in the electrophoresis tank (TaKaRa Bio, Kyoto, Japan) and 2.0 µl of 100kb Hyperladder™ (Bioline, London, UK) was added to provide molecular weight markers. To the rest of the wells, a mix of 2µl of loading buffer (0.25% w/v bromophenol blue, 0.25% w/v xylene cyanol, 30% v/v glycerol) with 5µl DNA was added to each well. Samples were run for 35-45 minutes at 135V, visualized using the Amersham Imager 600 (wavelength 294nm) and photographed.

Only reactions generating bands where the negative controls were blank on the gel were put forward for purification and sequencing. PCR products were then cleaned-up using ExoSaP-IT® (GE Healthcare, UK) clean up kit for the removal of unused primers and nucleotides. ExoSaP-IT® utilises exonuclease I which digests excess primers, and shrimp alkaline phosphatase which removes unincorporated dNTPs. 2µl ExoSaP-IT® was added to 5µl of second round PCR product on ice and subjected to 37°C for 15 minutes on a thermocycler. The final stage was performed by the Barts Genome Centre. Treated DNA was subjected to a sequencing reaction using BigDye 3.1 Terminator cycle sequencing (Thermo-Fisher, UK), and run on an ABI Prism 3100 genetic analyser (Thermo-Fisher, UK),

Digital sequence files were viewed using 4Peaks software (available at **https://nucleobytes.com/4peaks/**) and compared to the 2nd revised Cambridge reference sequence (rCRS) [306], using the pairwise sequence alignment software provided by European Bioinformatics Institute (**https://www.ebi.ac.uk**). Previously reported polymorphisms were identified and eliminated by comparing their sequence against that held in the Ensembl database (**https://www.ensembl.org**) and the Mitomap database (**www.mitomap.org**) of reported polymorphisms and also by comparing the genotype of non-epithelial tissue within each sample. The sequence from each region of interest was then compared for unique variants.

Identified mutations were confirmed in both the forward and reverse mtDNA sequences by re-sequencing from the original lysate.

4.3.7 Additional Sanger sequencing during technical set-up

Additional Sanger sequencing outside of the mitochondrial work was performed. Sanger sequencing is a cheaper and faster alternative to next generation sequencing (NGS) during the set-up phase of this project when testing certain experimental conditions. An aliquot of 5 µl of purified PCR products were combined with 5 µl of a single sequence specific primer (forward or reverse) at a concentration of 5 µM into a 1.5ml Eppendorf tube. The tubes were sealed with Parafilm (Fisher Scientific, UK). The bisulfite specific primer sequences to facilitate the Sanger sequencing are available in results chapter 5. The samples were then couriered via the commercial LightRun Sanger Sequencing service operated by Eurofins Genomics GmbH, Ebersberg, Germany, **https://eurofinsgenomics.eu/**. Sequencing results with accompanying chromatograms are available online next working day to be downloaded and analysed. A further software package called BioEdit Sequence Alignment Editor (available at **https://bioedit.software.informer.com/7.2/**) was utilised to visualise the data.

4.3.8 Illumina® next generation sequencing

The Illumina® (San Diego, CA) platforms permit massively parallel genomic sequencing-by-synthesis (SBS) of multiple input DNA molecules from multiple samples. Individual DNA molecules bind to the flow cell and are amplified to form clusters. Clusters are sequenced in parallel over 75-300 cycles of SBS with each individual cluster giving a single *read* that, after de-multiplexing, is compiled into the R1 (and R2 in paired end reading) dataset as a FASTQ file for each input sample suitable for downstream bioinformatics analysis. Different platforms have different capacity of SBS cycles, read depth, cluster formation and sample number. This thesis

uses the Illumina MiSeq and Novoseq platforms which offer maximum 300bp and 250bp paired end (PE) reading respectively.

4.3.9 Bioinformatics pipeline

The raw sequencing data requires passage through a number of quality control software packages, filtering, clustering, alignment to the bisulfite converted human genome and subsequent base calling of methylated (cytosine) and unmethylated (thymine) CpG loci. The specifics and set-up of this new pipeline are discussed in results **chapter 5**.

4.4 Gene Target Selection

4.4.1 Gene selection

Of importance, the selection of genes requires them to be evolutionarily neutral in the oesophagus, that is, silent and bearing no benefit or disadvantage to the cellular survival or proliferation.

A tabulated dataset of gene expression was downloaded from the Genotype-Tissue Expression (GTEx) Project website[307, 308] (**www.genome.gov/Funded-Programs-Projects/Genotype-Tissue-Expression-Project**). This was filtered for genes demonstrating no gastrointestinal (GI) expression sorted by RNA sequencing data[309], primarily those expressed in cardiac or skeletal muscle. A transcript-per-million (TPM) value of 0.5 is classed as "below cut-off" for expression, with 0.5-10 "low expression", medium to high expression levels from tissues are in the region of 100s-1000s TPM[310]. A manual cross-referenced dataset screen of expression was conducted on genes identified in the downloaded dataset to either accept or reject based on their GI TPM values. The platforms included the GTEx portal[307] (**www.gtexportal.org/home/**), The Human Protein Atlas (HPA)[311] (**www.proteinatlas.org**), The Expression Atlas[310] (**www.ebi.ac.uk/gxa/home**), GeneCards database[312] (**www.genecards.org**) and the Ensembl Genome Browser[313] (**www.ensembl.org**).

4.4.2 Design of target amplicons

Promoter regions, transcription factor binding sites and CpG islands were targeted within each gene using the Ensembl Genome Browser[313]. Where CpG islands were not listed on Ensembl, the Database of CpG Islands and Analytical Tools (DBCAT, available at **http://dbcat.cgm.ntu.edu.tw/**) was used to locate them[314]. A CpG island is defined as a genomic region >200 bp with >50% guanine-cytosine content and a ratio of observed-to-expected (oe) CpGs >0.6 (60%)[250]. Each sequence was bisulfite converted using the Bisulfite Primer Seeker 12S online tool[315]

(**https://www.zymoresearch.com/pages/bisulfite-primer-seeker**). Bisulfite specific forward and reverse primers were subsequently designed using Primer3[316] (**https://primer3.ut.ee/**) and Methprimer[317] tools (**urogene.org/methprimer/**). Primer-BLAST[318] (**https://www.ncbi.nlm.nih.gov/tools/primer-blast/**) was used to screen each primer set for off-target genomic DNA (gDNA) products. Finally, BiSearch[319] provided a check of primer sequences against the bisulfite converted genome for off-target amplification products which, if present, would significantly affect PCR efficiency and sequencing power. All bisulfite specific primers are designed to be indifferent to methylation status. Where a CpG locus was within the primer sequence a degenerate base of Y (forward [C or T]) or R (reverse [G or A]) was used however, only if within the 5' third of the sequence otherwise the primer was rejected.

4.4.3 Reverse transcription PCR (RT-PCR)

A two-step RT-PCR was performed targeted for each selected gene to exclude RNA expression from BO. The mRNA sequence alongside the gene exon and intron sequences were exported from the Ensembl Genome Browser[313]. Forward and reverse primers were designed to target different exons either side of intron sequences using Primer3[316] software and Primer-BLAST[318]. Therefore, both mRNA transcripts and genomic DNA (gDNA) can be amplified by the same primer set resulting in different amplicon sizes.

Scrapes of BO histopathological sections from microscopy slides were taken. Total cellular RNA was isolated using the RNeasy® Plus Micro kit (Qiagen, UK) according to the manufacturer's protocol following instructions for microdissected cryosections. Quantification and purity were confirmed using the Qubit 4 Fluorometer and Nanodrop One systems (both Thermofisher Scientific, UK).

Reverse transcription (RT) was carried out using QuantiTect® Reverse Transcription kit (Qiagen, UK) as per the manufacturer's protocol. In brief, Total RNA was split into 12 µl aliquots and mixed with 2 µl 7X gDNA wipeout buffer and incubated for 2 minutes at 42^{o}C to eliminate contaminating gDNA. Following this reaction, a sample

is kept aside to act as a no-reverse-transcription (NRT) control in downstream PCR to confirm efficacy of the gDNA wipeout reaction. All other 14 µl samples are added to the RT master mix (1 µl reverse transcriptase, 4 µl 5X RT buffer and 1 µl RT Primer mix) and 1st step RT is performed with incubation at 42°C for 15 minutes to generate copy-DNA (cDNA). The reaction is terminated by inactivation of reverse transcriptase at 95°C for 3 minutes.

Second-step PCR was performed in 25 µl reactions (0.2 units Taq polymerase, 200 µM dNTPs, 10X PCR buffer, 1.5mM $MgCl_2$, 5ul 5X Q solution [all sourced from Qiagen, UK], 200 nM each 5' and 3' target specific primer [Sigma Aldrich, UK] and 1 µl of template) and subjected to denaturation at 94°C for 3 minutes followed by 35 cycles of 94°C for 1 minute, 60°C or 64°C (depending on target) for 1 minute and 72°C for 1 minute. A final extension step of 72°C for 10 minutes completed the thermocycling. After amplification, PCR products were visualised on a 1.5% agarose gel as described in section 4.3.3.

Each gene target for RNA expression was supplemented by: a gDNA positive control, to confirm efficacy of the primer set and thermocycling conditions; a no-template water control (NTC), to exclude RNA contamination; a no-reverse transcriptase (NRT) control, to exclude gDNA contamination; and a *GAPDH* housekeeping gene positive and negative water control, to confirm presence of cDNA. The final *GAPDH* controls are important as all target genes were expected to give no amplification and hence no visible band within the cDNA template reactions.

5 Results

A novel protocol for a targeted, high-resolution, allele-specific methylation sequencing (ASM-Seq) array

5.1 Introduction

The study of epigenetic drift can reveal the ancestral relationship between cellular populations. Gradual changes over mitotic time of the binary methylation sequence (herein called a methylation "tag") that occurs independently of a cell's geno-phenotypic and environmental fitness to survive (or die) and proliferate (or not) can be leveraged to understand the dynamic of normal tissue homeostasis and aberrant carcinogenesis.

Previous work principally interrogated the methylation tags of three genes, *NKX2-5, MYOD1* and *BGN* with primer sets designed to amplify short regions within the 3' untranslated region (UTR) and promoter regions respectively[254, 275, 276, 279, 282]. The resulting amplicons were short and only contained 8, 5 or 9 CpG loci respectively. Therefore, the maximal methylation diversity permutations of unique sequences equalled 2^8 (256), 2^5 (32) and 2^9 (512), but obtaining these numbers was limited by processing technologies at that time where only a handful (<20) of tags (or "reads") were possible.

The low coverage of CpG loci also reduces the chance of detecting stochastic methylation mutation events within such short regions. Assuming ~5000[254-256] alterations per cell division across the 28,000,000 million CpG loci in the genome, this would result in ~0.018% chance of any individual CpG locus being affected each division. In order for the mitotic clock to *tick,* enough variation between target sets of methylation tags has to happen by chance to present an ever-increasing heterogenous pattern. Of course, a homogenous pattern tells its own story of a recent common ancestor or clonal expansion but with limited technical CpG coverage there is a risk this is purely by chance. Furthermore, with respect to *BGN* which is an

X-chromosome locus, this was primarily only reliable for clock purposes in male subjects due to the physiological epigenetic X-inactivation[265] (Lyonisation) of one copy of the XX karyotype in women. With such limited methylation tag reads the sequencing power in females by *BGN* is further diminished by half owing to the bias of one heavily methylated locus.

In line with this, there is also a crucial *just-right* window of methylation density to maximise the effectiveness of any mitotic clock. In particular, in heavily saturated or unsaturated conditions where the total methylation is close to 100% or 0% respectively there is less scope for errors to have a perceivable difference on a background of overall homogeny[267]. More so, there may be some known or unknown biological rules in such gene targets that has been unwittingly selected for, for example altering chromatin state affecting distant gene expression elsewhere and thus the methylation status is stabilised and unlikely to change. The idea here, is to identify gene targets with methylation tags that change in a continual and steady neutral fashion unhindered by the weight of evolutionary selection. Therefore, it is prudent to widen the gene target and sequence length coverage to maximise the chance that chosen target CpGs have a suitable rate of change across the length of the amplicon. These particular type of CpGs are the most powerful in defining a mitotic clock and with respect to BO this was evidenced by Curtius et al.'s[297] paper which computationally identified 67 such age-related drift CpGs from an array of 485,000 screened to infer dwell time.

However, despite the vast CpG coverage delivered by beadchip arrays, the probes are directed against single CpG sites scattered across the genome. Therefore, the allele specific locus-by-locus binary resolution is lost and with it inferences relating to cell-by-cell clonality. The outputs also deliver an averaged CpG locus methylation density/percentage turning a truly discrete data point (0 or 1, unmethylated or methylated) into a continuous variable that represents the entire pool of DNA template molecules. It is impossible to reverse engineer this data back to the discrete form which, if we are interested in the lineage of somatic inheritance between temporo-spatially linked populations is vital to know. Similarly, whole genome bisulfite sequencing (WGBS) usually involves a step of additional DNA fragmentation

prior to hybridisation with the sequencing adapters in the library preparation process[320]. This results in short stretches of DNA, sequenced maximally via 75-150 bp paired-end (PE) reading, becoming independent variables in the processing pipeline and hence degradation being able to characterise the distinct aggregate genome from which they originated. Moreover, the technique is limited by cost and sequencing depth with average reads of 15X per sample[321] hence failing to resolve the issue of how to maximise intra-sample depth to detect methylation tags of low variant allele frequency (VAF). Despite lauding single CpG resolution as an output the end result of WGBS again is that of *averages* rather than *specifics* within and across samples. Note also, these read numbers are just about comparable to the prior methylation tag lineage work albeit with much greater genome coverage than 8, 5 and 9 CpGs[254, 267, 276, 279].

Bisulfite amplicon sequencing permits detection of the single molecule distribution of methylation tags within samples resulting in a discrete and binary output, thereby improving the quantitative capacity that genome wide techniques lack. There is always a trade off with overall coverage being limited to the length of the amplicon insert between forward and reverse primers and the number of single or multiplexed primer pairs. The benefit though is potential for higher throughput of samples at reduced cost and greater read depth per sample. A standard Illumina® MiSeq™ 300 bp PE reading costs circa £2,000 offering 25,000,000 reads divided across a pool of samples. The number of samples that can be loaded on the flow cell to achieve a defined depth is therefore inversely proportional to the number of amplicon sequences in the targeted panel multiplied by the DNA concentration (initial copy number) of that sample. For example, assuming 100% efficiency, a gene amplicon panel of 20 targets achieving a read depth of 20X would use 400 reads per genome copy, thus 62,500 copies (~375ng original DNA extracted template) of the human genome can be loaded for one run. This can equate to in the order of 10s-100s of samples per flow cell (especially from DNA extracted Barrett's glands which are usually <10ng) whereas attempting to achieve similar single CpG resolution analysis with WGBS, the equivalent high read, highest coverage platform would allow 1-5

samples at best but unlikely with full coverage or depth, a well reported cost limitation to its use, indeed repeated runs are commonplace[321, 322].

An additional aspect to consider from prior targeted methylation sequencing work is the impact that PCR and sequencing error has on the final data output. PCR bias and errors are well-recognised in bisulfite sequencing protocols[323]. Extremes of guanine-cytosine (GC) content, high or low, and high adenine-thymine (AT) both characteristics of bDNA or eDNA are key drivers of this, rendering challenges for accurate DNA replication by polymerases. The high AT (~80%) and overall low GC content (~20%) that is concentrated in CpG islands, thus paradoxically high-GC content areas, proves to be particular sticking points for impaired PCR amplification[324]. Bisulfite treatment further compounds the potential bias with relatively indiscriminate DNA fragmentation and degradation eroding the integrity of the original DNA template composition and causing uneven or failed PCR amplification. Ideal outputs would be the uniform amplification of all source DNA template targets such that sequencing read assignment, after equimolar correction, is identical across all components of final library composition. Clearly if one DNA template is over-represented through PCR bias then the veracity and probable lack of diversity of the final methylation tags fall into question. This has always been a concern with the prior colorectal studies that an element of the homogenised patterns may be due to PCR rather than a true reflection of crypt stem cell dynamics[276]. This is especially considering the low read depth that fails to disapprove such a hypothesis.

To address this, alongside next generation sequencing (NGS), I have designed a novel PCR protocol that incorporates unique molecular identifiers (UMIs) to act as a biological barcode attached to every allele template within a sample. The use of UMIs gives an ability to track back to the original DNA molecule accurately, quantify and determine relative abundance of sequences within a sample, discover low frequency (≤1%) sequences, exclude PCR duplication bias, and detect and exclude PCR and sequencing errors through generation of consensus reads from UMI families (discussed in bioinformatics pipeline, **section 5.5**). Protocols involving UMIs have been developed in genomic DNA and RNA studies but never before in a bisulfite or

enzymatically converted template[325-328]. Previous attempts to ascertain allele specific methylation analysis use computational models, however these rely on mathematical probability and can be grossly influenced by the CpG density to read length ratio[329, 330].

My protocol gives a data output where sequencing reads with a different UMI represent different original template bDNA or eDNA molecules, while sequencing reads with the same UMI will represent PCR amplification from a single original molecule. Coupled with increasing the gene target coverage with a newly designed and multiplexed primer array suitable for massively parallel NGS, the results here within lay the groundwork in discovering intricate clonal relationships in BO evolution and mitotic ageing.

5.2 Aims

The aims of this chapter are four fold:

a) Identify a panel of non-expressed genes in GI epithelial tissues with regions of high CpG loci density. Design a primer array to amplify these targets in multiplex PCR.

b) Incorporate UMIs into primer design and PCR protocols to facilitate reliable and reproducible allele specific methylation analysis

c) Establish a NGS bioinformatics pipeline to compile consensus UMI reads, genome alignment, allele specific methylation sequence generation and variant calling.

d) Perform basic sensitivity and specificity testing against a variably methylated and quantitative templates and test applicability of the technique within a cell line population to measure clonal history, dynamics and stochastic methylation mutation rate.

5.3 Building, testing and optimising the target gene panel

5.3.1 Target gene panel

In total, over 150 genes were screened as detailed in Methods **section 4.4.1**. Of these 17 were accepted (**Table 5.1**), the rest being excluded where regulatory elements and/or CpG islands where not found or TPM expression values were unacceptably high. 9 genes have primary expression in cardiac tissue, 3 genes in skeletal muscle, 3 genes shared common expression in both tissue types, 1 gene in connective tissue with the final "gene", *LOC*, representing an intergenic genomic area of the X chromosome and thus has no function expression. One gene, BGN was above the TPM threshold. Otherwise, TPM values were "below cut-off" (≤0.5) in 6 genes, with the remaining 10 genes having an average of 1.75 TPM (range 0.7-3.9; median 1.56). The genes of *NKX2-5*, *MYOD1* and *BGN* that have been used in previous similar studies[254, 276] were included in this total of 17 genes and formed part of the initial optimisation work-flows.

BGN of note did not meet the TPM requirements with maximal expression in oesophageal mucosa at 64 TPM and muscularis at 115 TPM including reported active expression in gastric cells on the Ensembl regulatory builder function. Its inclusion in this thesis was primarily on literature precedent, although ultimately it failed the optimisation steps and was subsequently discarded from further use. As such, its values have been excluded from the averages reported above.

5.3.2 RNA expression analysis

Intron-spanning primer sets were designed (**Supplemental Table S2**) and a two-step RT-PCR protocol was carried out as described in Methods **section 4.4.3**. **Figure 5.1a** shows the H&E section of the BO tissue used for RNA extraction that came from a 61 year old male patient with long segment BO who has never progressed to dysplasia or OAC. 3 targets (*TNNI3, CSRP3, SCN5A*) showed preferential amplification with annealing temperature of 64^{o}C, with 60^{o}C deemed suitable optimisation for the

remaining targets. No gene target showed any RNA expression whereas *GAPDH* RNA target proved positive in all instances demonstrating an appropriately intact cDNA template (**Fig. 5.1b**). 14 targets showed successful amplification product of the gDNA positive control demonstrating the primer sets were functional and efficacious. *CSRP3* failed to give a band in the gDNA lane, this may be due to pipetting error as prior optimisation experiments (presented in **section 5.3.5**) had been successful.

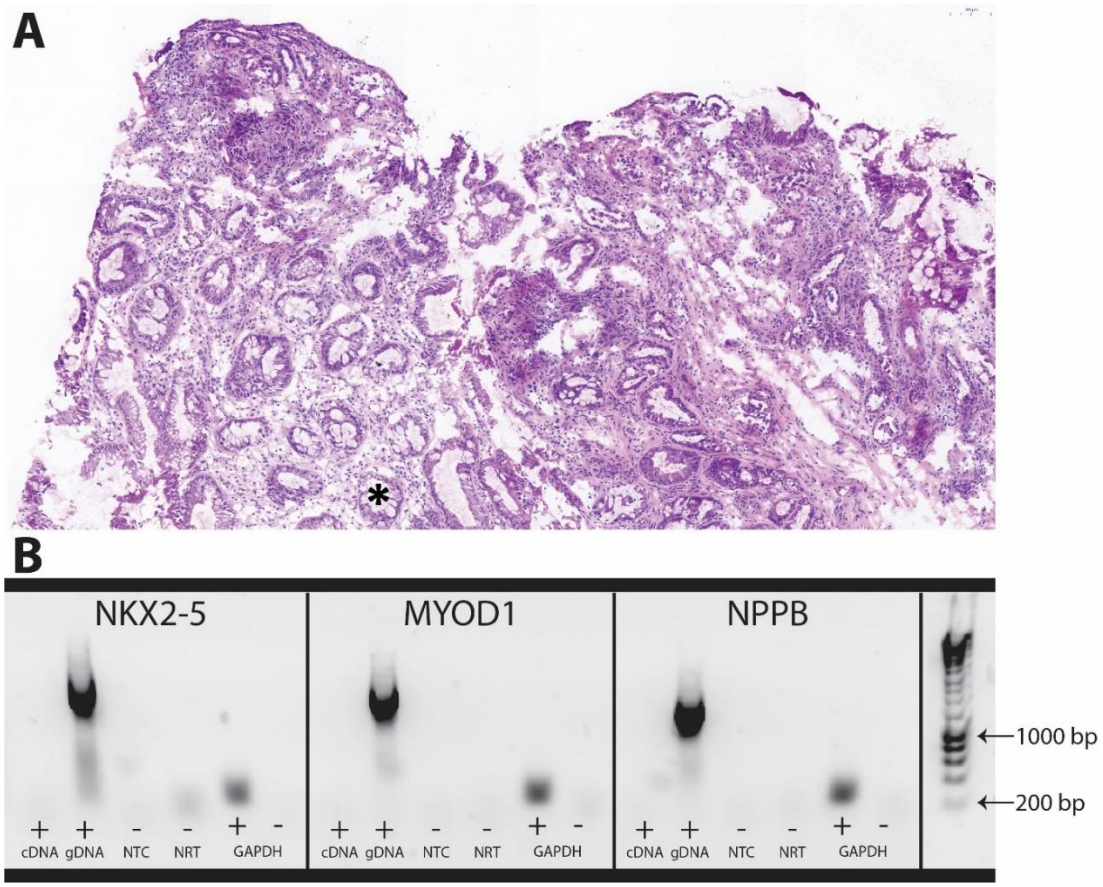

***Figure 5.1**: RNA expression experiments of target genes. (A) H&E stained section of Barrett's epithelium used as the template to demonstrate the non-expression of the target panel of genes with RT-PCR. Note the presence of goblet cell containing glands (asterisk). (B) Merged agarose gels of three example target genes NKX2-5, MYOD1 & NPPB. All show no product in the cDNA lane confirming non-expression in the Barrett's epithelium tissue section. The gDNA lane confirms functioning primer sets and GAPDH confirms an intact cDNA template in all cases. A 1kb Hyperladder™ is shown. cDNA – copy DNA; gDNA – genomic DNA; NTC – no-template-control (water); NRT – no-reverse-transcription control*

5.3.3 Bisulfite specific primer design considerations

Methods **section 4.4.2** details the multi-step process involved in designing primers specific to a bisulfite or enzymatically converted template. All primers were designed to be methylation-indifferent primers (MIP) rather than methylation-specific primers (MSP). MSP primers are usually designed to amplify only when a target locus is methylated or unmethylated[331]. These primers must include a CpGs site located near the 3' end to ensure they bind specifically to representative methylated (Cytosine-Guanine) or unmethylated residues (Uracil-Guanine) necessitating two originals sets of primer pairs to allow this distinction and a subsequent proportional analysis. However, this design is limited to only a few select CpGs and is prone to failure if there is differential methylation of the CpGs within the primer annealing zone[332]. Given the desire to amplify all templates regardless of methylation status optimal design of MIP primer pairs was pursued. Here, any primer overlap with CpG sites is confined to the 5' end which is less crucial with respect to binding specificity[332]. Furthermore, a mixed degenerate base of either "Y" (representing "C" or "T") on forward primers and "R (representing "G" or "A") on reverse primers to incorporated. Nested primer sets were designed for each amplicon target to provide redundancy in case a particular forward or reverse or both primers failed in the outer set.

It is important to note that after bisulfite conversion the original double stranded DNA (dsDNA) template is no longer complementary and hence exists as single stranded DNA (ssDNA). Therefore, a single primer set is specific to one strand only and during the 1st thermocycle of PCR, while the reverse (3') primer will bind, the forward (5') primer will not as there is no complementary template available (**Fig. 5.6**). There is the option to design the counter-part pair of primers for *strand specific* methylation PCR but this was not undertaken for the following reasons. The primer panel was purposely designed directed at densely populated areas of CpG to maximise methylation tag coverage, however this resulted in limited short spans of non-CpG containing sequences in which to fit primers obeying the rules as described above and in **section 4.4.2**. A proportion of primer sets (16/40) also included at least one degenerate mixed base at the 5' end. In addition, a Guanine-Cytosine (GC) clamp was sought for each design located at the 3' end to improve primer binding and

specificity. Thus, a second complementary strand reverse primer would usually have violated these rules, in particular the avoidance of CpG sites at the 3' end complementary to the mixed degenerate base at the 5' end of the first set. Further complexity would be encountered in matching technical considerations for 4 primers versus 2 melting temperature (Tm), GC content (optimally aiming between 40-60%) and avoidance of primer dimer formation and off-target binding. Once the final primer sets were designed an in-silico PCR combining all primer sets was undertaken using the FastPCR (**https://primerdigital.com/fastpcr.html**) computer software programme. This was in view of latter plans to multiplex all primer sets for the ASM-Seq protocol. No significant clashes or alternative product were observed. Full gene sequences, primer annealing zones and technical data on primer design and reaction kinetics including propensity scores to form secondary structures (e.g. hairpins) are available if necessary at request.

Bisulfite also results in all cytosine within the genome being converted to uracil (and subsequently amplified by PCR to thymine) resulting in an AT-rich template with reduced nucleotide diversity. This provides challenges in primer design to obtain suitable on-target specificity and to reduce mismatched primer binding, for example in areas of poly-thymine. Primer dimer formation is also more readily appreciated due to this reduced nucleotide diversity causing greater chance of short AT complementary strings between primers[333]. Bisulfite treatment also fragments and degrades DNA which gives limitations on yields and amplicon lengths, although recent studies show promise with successful amplification of targets up to ~1.5kb in length[334].

5.3.4 The designed primer sets

Across the panel of 17 genes, 22 targets (**Table 5.1**) have been designed directed to promoter regions, transcription factor binding sites, first exon(s), CpG islands and one intergenic region. 5 genes (*NKX2-5, MYOD1, TNNI3, SPTB, SBK3*) have two targets for enrichment by PCR. Both nested sets for each target were tested and optimised with the best set being used for the ASM-Seq protocol, this also included

testing a mix of outer and inner primers when either original pair failed. The primer sequences are available in (**Table 5.1**). The size of all amplicons range from 150 bp to 592 bp (mean 495, median 521 bp). When the less optimal primer set is excluded this changes to 367 bp to 581 bp (mean 479 bp, median 488 bp). The total CpG locus coverage with the experimental primer sets is 752 loci (mean 36 loci, median 34 loci per amplicon) providing more than ample 2^{752} possible sequence permutations per sample. A restriction of <600 bp was enforced for amplicon length for compatibility with the Illumina® MiSeq platform that permits maximal 300 bp paired end (PE) reading. Therefore amplicons over this size would not be fully sequenced with an uncovered insert at their mid-portion.

Gene	Chr	Str	Primary Expression	Max GI TPM	Primer Set Name	Target Sequence Forward	Target Sequence Reverse	Amp Size (bp)	CpGs (#)	Target region
NKX2-5	5	-	Cardiac	≤0.5	**1-NKX2-5 Set 1**	**GYGGTATTATGTAGGGAAGTTGTTAGG**	**ACACCAAACATCTTACATTCTAAACC**	**563**	**42**	TF Binding Site
					1-NKX2-5 Set 2	TATTATGTAGGGAAGTTGTTAGGGGT	AACRCRTATCTCCTCCTCCTAACCCT	504	38	
					2-NKX2-5 Set 1	TAGYGGTAGGATTAGATTTTGGAGTT	CTCAAAATCATATTAAAAACCCCTTC	482	48	Exon 2 3' UTR
					2-NKX2-5 Set 2	**TAGGATTAGATTTTGGAGTTGGTG**	**ATCATATTAAAAACCCCTTCTCCC**	**470**	**47**	
MYOD1	11	+	Skeletal Muscle	≤0.5	1-MYOD1 Set 1	GGYGTYGTTTGAGTAAAGTAAATGAGG	ACCCRCACACTTCCAAAACAAACTAC	455	46	TF Binding Site & Exon 1
					1-MYOD1 Set 2	**GYGGTTGTTTAAGGTGGAGATTTTG**	**CAAACTACCCCAATAACAACTACCCA**	**367**	**40**	
					2-MYOD1 Set 1	TAGTGGGTGGGTATTTAGATTGTTAG	ACRCAAAATCTCCACCTTAAACAACC	584	64	TF Binding Site & Exon 1
					2-MYOD1 Set 2	**GTGGGTATTTAGATTGTTAGTATTTT**	**TCTCAAAAACCTCATTTACTTTACTC**	**516**	**59**	
TNNI3	19	-	Cardiac	1.226	**1-TNNI3 Set 1**	**TAAAGGAAGAGATTTAGATTGGTGGATG**	**CTCTCRCTCCAACTCTTACTTTACAATC**	**540**	**26**	Promoter
					1-TNNI3 Set 2	GATTTAGATTGGTGGATGGGAATGAGG	CTCCAACTCTTACTTTACAATCTACAAC	524	25	
					2-TNNI3 Set 1	TTTGTTTAGAGGGGATTTTAGGGGTT	CCATCCACCAATCTAAATCTCTTCCT	592	29	5' Upstream
					2-TNNI3 Set 2	**TGGTTGGGATTTTTAGGGTTAGGGT**	**CTTCCCATCTATCCCTAAACAAATCC**	**435**	**25**	
CSRP3	11	-	Cardiac and Skeletal Muscle	≤0.5	CSRP3 Set 1	GGTAATAGGGTTATAGTAGGATAAAGATG	TATACCCTCTAACAAATCAATTCTCC	530	25	5' Upstream & Exon 1 5' UTR
					CSRP3 Set 2	**GTTGGTTAAGTTTTATTATGTTTTAAGG**	**ATTCTCCTTTCTAATTCCTTACTTAC**	**451**	**25**	
NPPB	1	-	Cardiac	0.7	**NPPB Set 1**	**TTYGTTGAAGAGAGTAGTTTTTGAGAG**	**CCTACCCTACCATACAAAATTATCTC**	**550**	**45**	TF Binding Site
					NPPB Set 2	GAGTAGTTTTTGAGAGTTTGTTTTAAG	TATCTCTAATTTATCAACCACATTCC	519	44	
MYO18B	22	+	Cardiac and Skeletal Muscle	3	MYO18B Set 1	GAGAGTATGTGTTTGTGTTTAGAGTTG	ACATTTTATTCTCAAATCCTTCCACC	591	27	Promoter & Exon 1 5' UTR
					MYO18B Set 2	**GTYGTTTTTTGGTTAGATTTGGAGTT**	**CCTTCCACCRAAACACTCTTATTTTCA**	**475**	**25**	

Gene	Chr	Str	Primary Expression	Max GI TPM	Primer Set Name	Target Sequence Forward	Target Sequence Reverse	Amp Size (bp)	CpGs (#)	Target region
PXDNL	8	-	Cardiac	0.9	PXDNL Set 1	AATTTGGGTTTTAAGAGGATAGTTGG	AATATAATCCAACATCAAATACATACAAC	525	16	Promoter & Exon 1
					PXDNL Set 2	**AGAGGATAGTTGGAGGTTAAGAGG**	**CAACRAACRATACTCTTAAAACAAAAACACC**	**487**	**16**	
CAMK2B	11	-	Cardiac and Skeletal Muscle	2	CAMK2B Set 1	GTTTTAYGAGGATATTGGTAAGTAAG	AACCCAAATTCCCCAAAAAACCTACA	585	78	Promoter
					CAMK2B Set 2	**AYGAGGATATTGGTAAGTAAGAGTAG**	**AAAACCTACAACAAAAACTCTCCCAC**	**564**	**78**	
ANKRD2	10	+	Skeletal Muscle	0.9	ANKRD2 Set 1	TATTTAGGTTTGGAAGGAGGGATAGA	AATCTCCTCATCCACAAAACCAATCT	437	29	CpG Island of Exon 3
					ANKRD2 Set 2	**TTTGGAAGGAGGGATAGATTTTGGTT**	**CTCCTCATCCACAAAACCAATCTACA**	**426**	**29**	
SCN5A	3	-	Cardiac	≤0.5	**SCN5A Set 1**	**GTGTATGTTAGTGTTTGTTAATGTGAG**	**AAAACTCCRACCRAACCAAAACTACC**	**573**	**72**	Promoter & Exon 1 5′ UTR
					SCN5A Set 2	GTTAGTGTTTGTTAATGTGAGTTTGTT	AAAACACTCRCTCACCTACTAATCCC	534	67	
SPTB	14	-	Skeletal	1.9	**1-SPTB Set 1**	**ATTGAATTGTGTGTAGTGGGGATGTA**	**CCRAACCCTCCTAAAACTCACCTACC**	**581**	**53**	Promoter & Exon 1 5′ UTR
					1-SPTB Set 2	GYGGGAAAGGTTGGAGGGTTTATT	AAACTCACCTACCCTAAAACTAAAAC	540	52	
					2-SPTB Set 1	GTTTTAGTTTTAGGGTAGGTGAGTT	AAATCACTCTCTAAATAACTCCCAAC	590	52	Promoter
					2-SPTB Set 2	**GGTAGGTGAGTTTTAGGAGGGTT**	**CTCACCTATCCCTCCTACCTAAAC**	**523**	**50**	
TNNT2	1	-	Cardiac	3.9	TNNT2 Set 1	TAAATAGTTTATTTGAGTAGTTGGAGG	TCTTAAAACCCAAACCTAACAACCTA	465	11	Promoter
					TNNT2 Set 2	**GTTGGAGGATTATATGGGTTTATATGG**	**ACCTATCCTCTAAAATATAACCTCCA**	**425**	**11**	
MYLK3	16	-	Cardiac	2	MYLK3 Set 1	TTTYGTGTTGGGAATTGGGAGTTTTAG	CATACTCCTAAACAACCACTAAACTA	559	40	Promoter & Exon 1
					MYLK3 Set 2	**GGGAATTGGGAGTTTTAGTTTTATGT**	**AATAAACCTCACCCTCTACCTACCAC**	**509**	**39**	
SBK2	19	-	Cardiac	1	SBK2 Set 1	TAGGGATATTTGTGATTTGGGGTTTG	AACTTAATTCAAACCTACCAACCTCC	553	24	Exon 2
					SBK2 Set 2	**AAGAGGGAGATTGAGTTATAATGAGAAG**	**TAATTCAAACCTACCAACCTCCCCAA**	**489**	**24**	

Gene	Chr	Str	Primary Expression	Max TPM	Primer Set Name	Target Sequence Forward	Target Sequence Reverse	Amp Size (bp)	CpGs (#)	Target region
SBK3	19	-	Cardiac	≤0.5	1-SBK3 Set 1	AGTATYGAGTATAGTATGAGTGTTGGA	ACCTTAAAATCTAATAACTCCAAACTC	546	30	Promoter
					1-SBK3 Set 2	**GTTAGGAGGGGATAAAGTTAGGAAA**	**TAAAATCTAATAACTCCAAACTCCTC**	**461**	**27**	
					2-SBK3 Set 1	TTTGTTTTAGTTTGTTTAGTAGTTGGGA	CAAAAATACTCTCATATTCTCAACACC	472	19	Promoter
					2-SBK3 Set 2	**TGTTTTAGTTTGTTTAGTAGTTGGGA**	**CTCTCATATTCTCAACACCTATTCC**	**462**	**19**	
BGN	*X*	+	Connective tissue expression	115	BGN Set 1	TAAATTGTTTAGGAGTGAGTAGTTGTTTT	CAAAACTAAAATACCAATCACCCAACC	150	10	Promoter
					BGN Set 2	**TAAATTGTTTAGGAGTGAGTAGTTGTTTT**	**AAAAACAACTTAAAACCAACCCTACC**	**489**	**24**	
LOC	*X*	-	Intergenic	≤0.5	LOC-L Set 1	GTTGTGGGATAGGTGTAGGAAT	CCCTTAACCTATCCTACAACC	158	14	Intergenic
					LOC-L Set 2	**TAAGATGGGTGGATGGTTGGAT**	**CCCTTAACCTATCCTACAACC**	**532**	**28**	

***Table 5.1**: Non-expressed gene targets and target specific primer sequences. This table details 17 genes used for the allele specific methylation sequencing protocol, their chromosomal location, strand and expression parameters. 5 genes had two targets designed across the respective CpG island identified, the nomenclature is denoted by the prefix '1-' or '2-'. In total, 22 amplicon targets with 2 primer sets per target were designed, with the sequences as stated in the table. Degenerate bases within any primer are highlighted in red. The amplicon length, CpG loci coverage and binding region of each primer set with the stated gene is detailed. The primer pairs highlighted in bold demonstrated the best amplification efficacy out of each set and were taken forward to the UMI primer design phase. Chr: Chromosome; Str: Strand; TPM: transcripts-per-million; Amp: Amplicon; TF: transcription factor.*

5.3.5 Testing amplification efficacy of target specific primers

All PCR primers designed in **section 5.3.4** were optimised for amplification. Template DNA from redundant cells of cell line DLD1 (colorectal cancer [kindly provided by Dr Angus Cameron, Bart's Cancer Institute]) was bisulfite converted and utilised for the optimisation process. For single-plex reactions, two PCR master mixes were tested, KAPA HiFi HotStart Uracil+ 2X ReadyMix (Roche, Burgess Hill, UK) and Phusion U Hot Start PCR 2X Master Mix (Thermofisher Scientific, UK). Both contain high fidelity proof-reading DNA polymerases which have had a mutation made to their dUTP binding pocket allowing them to read through uracil bases contained in a bisulfite converted template[335]. Without this modification, proof-reading enzymes are prone to stalling at uracil and subsequent amplification failure of the reaction. A high quality polymerase with proof reading ability is also vital in this challenging template in efforts to maintain underlying sequencing fidelity and reduction of PCR mutations[336].

Optimisation required multiple successive trials of PCR including annealing temperature, magnesium concentration and primer concentration gradients. Thermocycling conditions of annealing and extension time were varied along with trials at a lower extension temperature of 68^{o}C (compared to the usual 72^{o}C) which has been previously shown to be beneficial in amplification of a bisulfite template[334]. The final optimal conditions were established for the Phusion U master mix over KAPA HiFi with reaction constituents detailed in **Table 5.2a**. A touchdown thermocycling protocol (**Table 5.2b**) was necessary to reliably reduce off-target product and primer dimers as noted by another group[87]. **Figure 5.2** gives a snapshot representation of some of the variance observed during in optimising PCR conditions in single-plex, additional PCR optimisation outcomes is recorded in the research log and is available at request if necessary.

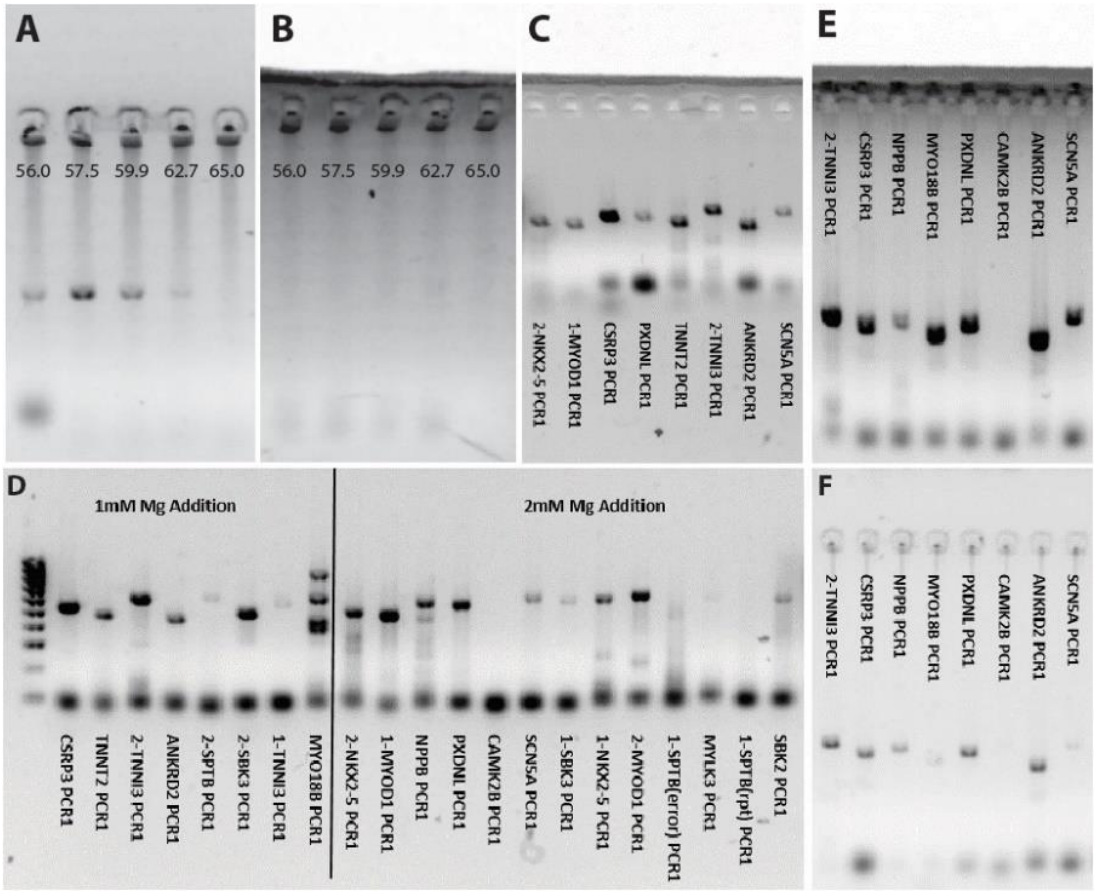

Figure 5.2: *Representative agarose gels from bisulfite specific single-plex primer testing and optimisation. Lanes have been labelled with the relevant primer set on gels C-E, where the prefix 1 or 2 is shown this represents different PCR targets for the gene. (A) and (B) show results from a 56-65°C temperature gradient PCR for two genes, CSRP3 in A and 2nd target for MYOD1 in B, using the recommended protocol for the PCR Ready Mix from the polymerase manufacturer. Note the differing efficacy of the PCR, the varying product concentrations across the temperature range and the presence of primer dimers. (C) Representative example of improved band strength with addition of magnesium chloride – 1mM addition in this case. Some primer sets required addition of 2mM magnesium to achieve suitable amplification (gel not shown). (D) first trial with touchdown PCR across the primer sets, individual primer sets either had addition of 1mM or 2mM magnesium chloride as shown. Touchdown of temperature was performed to 58, 60 or 62°C dependent on the previously determined optimised annealing temperature (Ta) for each primer set (data not shown). To streamline future PCR preparation time all primers were retested with addition of 2mM magnesium chloride followed by a touchdown protocol to 58 degrees (E) vs the previously optimised constant Ta of 58, 60 or 62 °C respectively in (F). (E) and (F) show a representative example of eight of the primer sets clearly demonstrating a stronger band and preference with Ta touchdown protocol over a constant cyclical Ta. CAMK2B however failed in both instances.*

The two primer sets were compared using the conditions in **Tables 5.2a-b**. The resulting agarose gel (**Fig. 5.3, asterisks**) determined which set to proceed with for the long term UMI experiment (also highlighted in bold in **Table 5.1**).

Reagent	**Volume (µl)**
Phusion U 2X Master Mix	12.5
Gene Specific Forward Primer	0.5 [200nM]
Gene Specific Reverse Primer	0.5 [200nM]
Nuclease-Free-Water	10.5
Template (bDNA)	1 [10ng/ul]
TOTAL	25

Table 5.2a: *Reagent constituents for bisulfite specific PCR*

Phase	**Cycle Number**	**Step**	**Temperature**	**Time**
Hot Start	1	Denature	98 °C	1 minute
Touchdown	10	Denature	98 °C	10 seconds
		Annealing	68 °C -> 59 °C Reduce by 1 °C every cycle	30 seconds
		Extension	72 °C	30 seconds
Amplification	35	Denature	98 °C	10 seconds
		Annealing	58 °C	30 seconds
		Extension	72 °C	30 seconds
Final Extension	1	Extension	72 °C	5 minutes
Completion	1	Hold	4 °C	Forever

Table 5.2b*: Thermocycling conditions for bisulfite specific PCR*

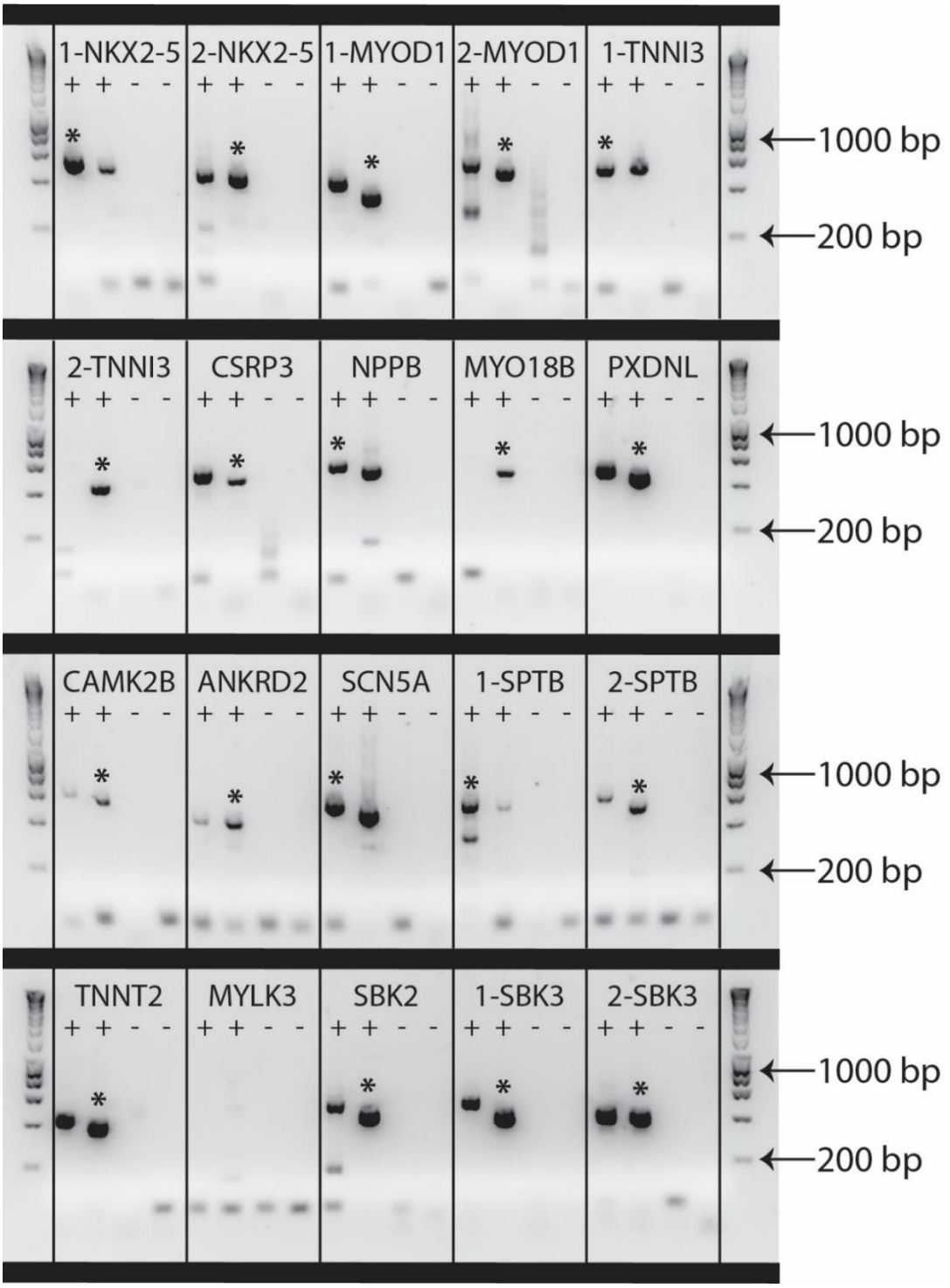

***Figure 5.3**: PCR amplification comparison of primer sets 1 & 2 for each target gene. The 20 gene targets are displayed on this merged agarose gel and labelled in each column. Each gene has four lanes from left to right: i) Set 1 bDNA template; ii) Set 2 bDNA template; iii) Set 1 NTC; iv) Set 2 NTC. Asterisks denote the primer set chosen for the future UMI protocol, product size details are given in **Table 5.1**. Note the presence of primer dimers and non-specific product that is prevalent when amplifying bDNA despite optimised conditions. Also note that both primer sets for gene target MYLK3 failed, this gene has subsequently been rejected. A 1kb Hyperladder™ flanks each gel row. bDNA – bisulfite converted DNA; NTC – no-template-control (water).*

5.4 Derivation of the allele specific methylation sequencing (ASM-Seq) protocol using unique molecular identifiers (UMIs).

5.4.1 Overview

The final optimised protocol involves four steps. The first is assignment of UMIs to individual alleles within the bisulfite converted DNA (bDNA) or enzymatically converted (eDNA) template in a short first round PCR. This is followed by target enrichment and pre-amplification of the 1st round product in a 2nd round PCR augmented by the reverse universal primer that generates multiple copies of each UMI tagged allele (a UMI family). The exact methylation tag of each starting template DNA molecule should therefore be present in high abundance all assigned to the same UMI. If a particular UMI family shows a different methylation sequence amongst the individual DNA molecules then this would represent PCR and/or sequencing error and can be detected for exclusion. These combined errors are estimated to occur at a rate of 1.2-2.5 per 1000 base pairs, this is likely to be higher in a bisulfite or enzymatic converted template[337]. Thirdly, a second and more prolonged PCR is undertaken to fully amplify the product using both the forward and reverse universal primers. Finally, and fourthly, a short library preparation step is required to add the final Illumina® indexes and adapter sequences for binding to the sequencing flow-cell (P7 and P5 respectively). There is a clean-up process between each round of PCR to maximise desired amplicons and reduce non-target product. The full protocol has been termed Allele Specific Methylation Sequencing, hereafter termed ASM-Seq.

5.4.2 Design of unique molecular identifier primers

A number of UMI primer designs have been tested due to the fastidious nature of this experimental protocol and difficulties encountered in generation of specific product suitable for sequencing analysis. Inspiration for the designs has included previous published protocols relating to use of UMI incorporation in gDNA and

ribonucleic acid (RNA) expression experiments[325, 326, 338, 339]. The tested designs are summarised in **Figure 5.4**. All designs involve the gene specific target sequence (chosen in **section 5.3.4**) at the 3' end, a UMI sequence of degenerate bases (N [Any base], W [A/T] and/or S [G/C]) with a universal sequence at the 5' end. The primary obstacles were formation of primer dimers and concatemers alongside off-target product which were more prevalent with certain designs. These particular products would then outcompete the target specific amplicon in the kinetics of the reaction consuming valuable reagents. Following testing, the Abridged-Adapter Universal Sequence (Abr-US) primer structure was taken forward as the most reliable in generating product.

The structure of the Abr-US UMI primers are schematically demonstrated in **Figure 5.4d**. The Abr-US acts as an overhang tail that theoretically does bind or not participate in 1st round PCR but acts as the substrate for amplification with universal primers in subsequent PCR rounds.

A final 21 bp UMI sequence of four "WSNN" repeat elements between 5 fixed bases was selected and incorporated into both forward and reverse primers[340]. This permits exceptionally abundant permutations of UMIs within each primer pool equating to $(2^8x4^8)^2$ combinations per target allele. The fixed bases between each "WSNN" repeat of "… A… T… A… CG" and pattern was initially chosen to deliver a known constant sequence required by the Bartender[341] UMI family clustering algorithm used in subsequent bioinformatics processing. However, evolution of the bioinformatics pipeline over time has resulted in abandoning Bartender, an alternative means of UMI alignment and clustering is now in place (described in **section 5.5**). Knowledge of the exact UMI pattern also allows further robust detection of PCR and sequencing error for exclusion of any reads which do not conform to the pattern. Of note, in testing a series purely comprising of 'N', it was noticed that guanine containing UMIs were preferentially incorporated (data not shown), therefore this pattern also helps to balance the UMI nucleotide make-up.

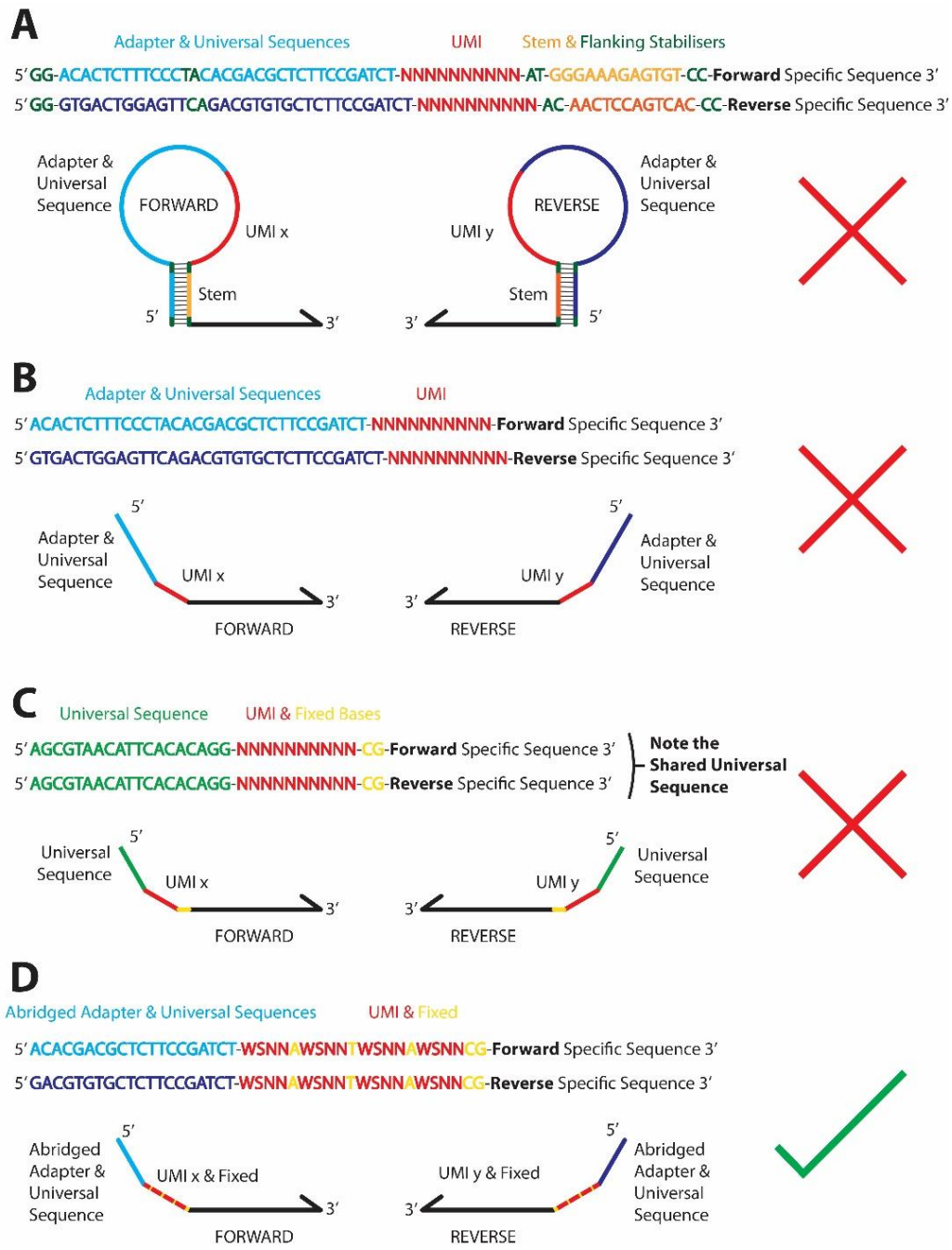

***Figure 5.4**: Tested designs of 1st round unique molecular identifier (UMI) primers. Primer sequences, annotated schematic of structure and testing outcome are displayed. (A) Hairpin universal sequence (HP-US) Primers: incorporate a hairpin structure to protect the UMI and Illumina® adapter sequences from off-target binding and dimerisation[325], failed. (B) Linear Adapter universal sequence (LA-US) primers: the Illumina® adapter acts as the universal sequence, failed. (C) Linear shared universal sequence (L-sUS) Primers: a balanced nucleotide density (GC vs AT) universal sequence is incorporated into both the forward and reverse primer, 2nd round PCR subsequently uses a single universal primer for amplification, this structure gave excellent amplification but due to the shared sequence would theoretically form hairpin structures on the Illumina® flow-cell and likely fail sequencing. (D) Abridged-adapter universal sequence primers (Abr-US): the 3' 20 bases of each Illumina® adapter sequence were incorporated into forward and reverse primers respectively, second round PCR uses 20 bp primers complementary to these sequences for amplification, protocol successful. Also note, the new "WSNN repeats" structure of the UMI in this design.*

5.4.3 UMI assignment and important considerations in 1st round PCR

Initial optimisation focused on a two cycle 1st round PCR performed using the Abr-US primer sets, the optimised reagents and conditions for singleplex reactions are detailed in **Tables 5.3a-b**. A slow ramping rate to the annealing temperature is utilised to mimic a touchdown protocol. Two cycle PCR is the maximal number to avoid incorporation of different UMIs to the same starting template allele. Furthermore, comparing 2, 3, 5 and 10 cycles in an optimised model did not increase the product yield (**Fig. 5.5**). Two cycles on bDNA gives a single strand template (bound to a shorter strand as dsDNA) with UMI incorporation both at the 5' and 3' end of the amplicon, termed double ended UMI (deUMI) product in this thesis (**Fig. 5.6**). The post 1st round PCR will also contain other fragments with a single UMI and Abr-US at the 3' amplicon end (reverse primer), termed single ended UMI (seUMI), however these would undergo inefficient linear amplification in 2nd round universal PCR rather than exponential and are thus outcompeted. They also fail to receive the full Illumina® adapter sequences to become library-ready in 3rd round PCR owing to a lack of the 5' Abr-US and as such should not be sequenced.

The alternative protocol is to consider a single cycle 1st round PCR whereby just the specific reverse primer is bound to the template bDNA followed by an immediate clean-up step to remove any unbound reverse primer UMIs. Following this the 2nd round PCR incorporates both the forward specific primer (without any UMI) and a universal primer complementary to the reverse primer universal sequence (termed Sequencing Primer 2, or SP2 in this thesis). The protocol benefits including avoidance of excessive specific primers (no forward primer) in the first round that is prone to non-target product formation and potential preferential amplification in subsequent cycles and the possibility of an interim *pre-amplification* step of to be determined varying PCR cycle duration accentuated by the use of SP2 alongside the specific forward primer that completed the target enrichment process. Furthermore, the issue of an extra UMI bound to original template DNA molecules, albeit without any theoretical disadvantage, is removed. The final sequencing ready libraries would consist of seUMI tagged DNA molecules with their full length methylation tag

relationships being recompiled during bioinformatic processing of the R1 and R2 sequencing dataset. The final ASM-Seq protocol uses this one-cycle iteration.

Regardless, whether a one-cycle or two-cycle protocol is utilised the importance of discontinuing 1st round PCR prior to a third cycle cannot be overstated. Allowing the reaction to proceed, as demonstrated in **figure 5.6** would result in excessive UMI tagging of single original DNA template rendering the whole purpose of UMI use obsolete.

Reagent	**Volume (µl)**
Phusion U 2X Master Mix	12.5
Gene Specific Forward Abr-US UMI Primer	0.5 [200nM]
Gene Specific Reverse Abr-US UMI Primer	0.5 [200nM]
Nuclease-Free-Water	10.5
Template (bDNA)	1 [Variable]
TOTAL	25

***Table 5.3a**: Reagent constituents for 1st round UMI assignment PCR for singleplex ASM-Seq*

Phase	**Cycle Number**	**Step**	**Temperature**	**Time**
Hot Start	1	Denature	98 °C	1 minute
Slow Ramp UMI Assignment Phase	2	Denature Annealing Extension	98 °C 58 °C – Slow Ramp by 0.2°C/second 72 °C	10 seconds 5 Minutes 30 seconds
Final Extension	1	Extension	72 °C	5 minutes
Completion	1	Hold	4 °C	Forever

***Table 5.3b**: Two-cycle PCR thermocycling conditions for 1st round UMI Assignment PCR in singleplex. Note that only two thermocycling conditions are conducted and the ramping time is slowed for annealing temperature to mimic a touchdown PCR protocol.*

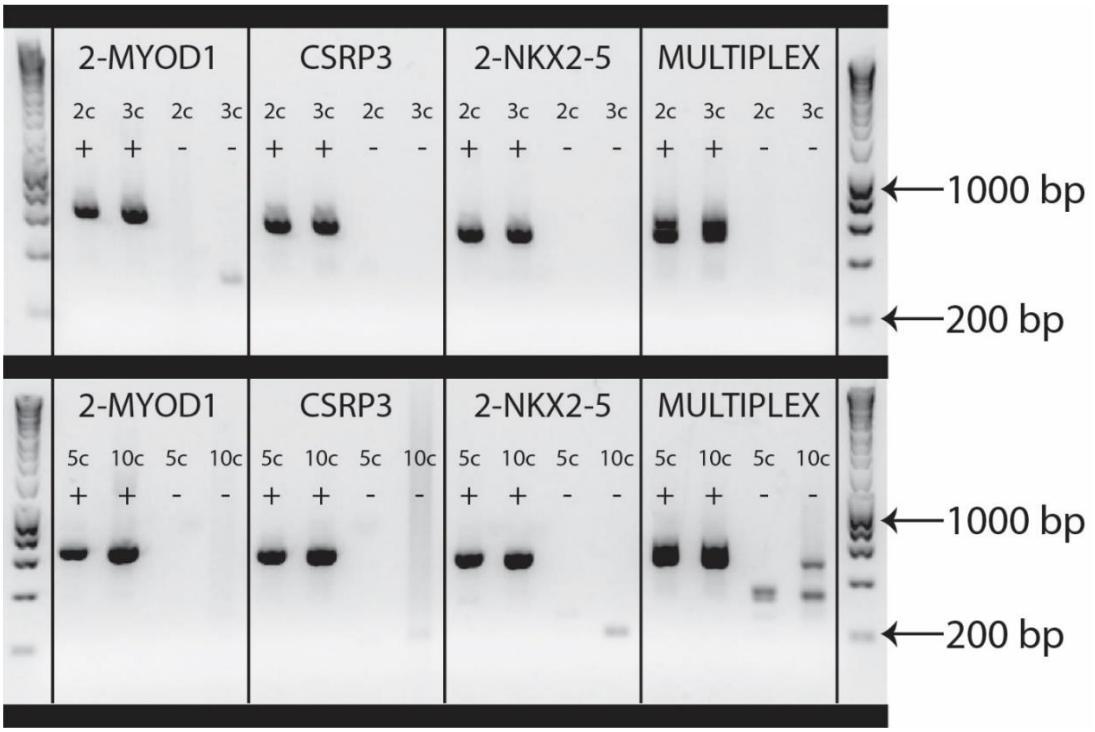

***Figure 5.5**: Comparison of cycle number in 1st Round UMI Assignment PCR. The amplification efficiency of 3 gene targets (2-MYOD1, CSRP3, 2-NKX2-5) and a multiplex of all three targets was tested comparing cycle numbers of 2, 3, 5 or 10 in the 1st round UMI assignment PCR. The primer design used in this instance was L-sUS primers (**figure 5.4c**). + denotes a bDNA template, - denotes no-template-control. The gene target and cycle number are annotated above each lane and the 1kb Hyperladders™ are shown. Note that additional cycles in 1st round PCR did not confer a significant increase in product yield after the 2nd round amplification PCR (section 5.1.7.5). This shows that 2 cycle 1st round PCR is reliable and suitably efficacious. This is beneficial as additional cycles above 2 would result in an over representation of true allele count in the final sequencing analysis due to the assignment of further UMIs to the same starting bDNA template (**see Fig. 5.6**)*

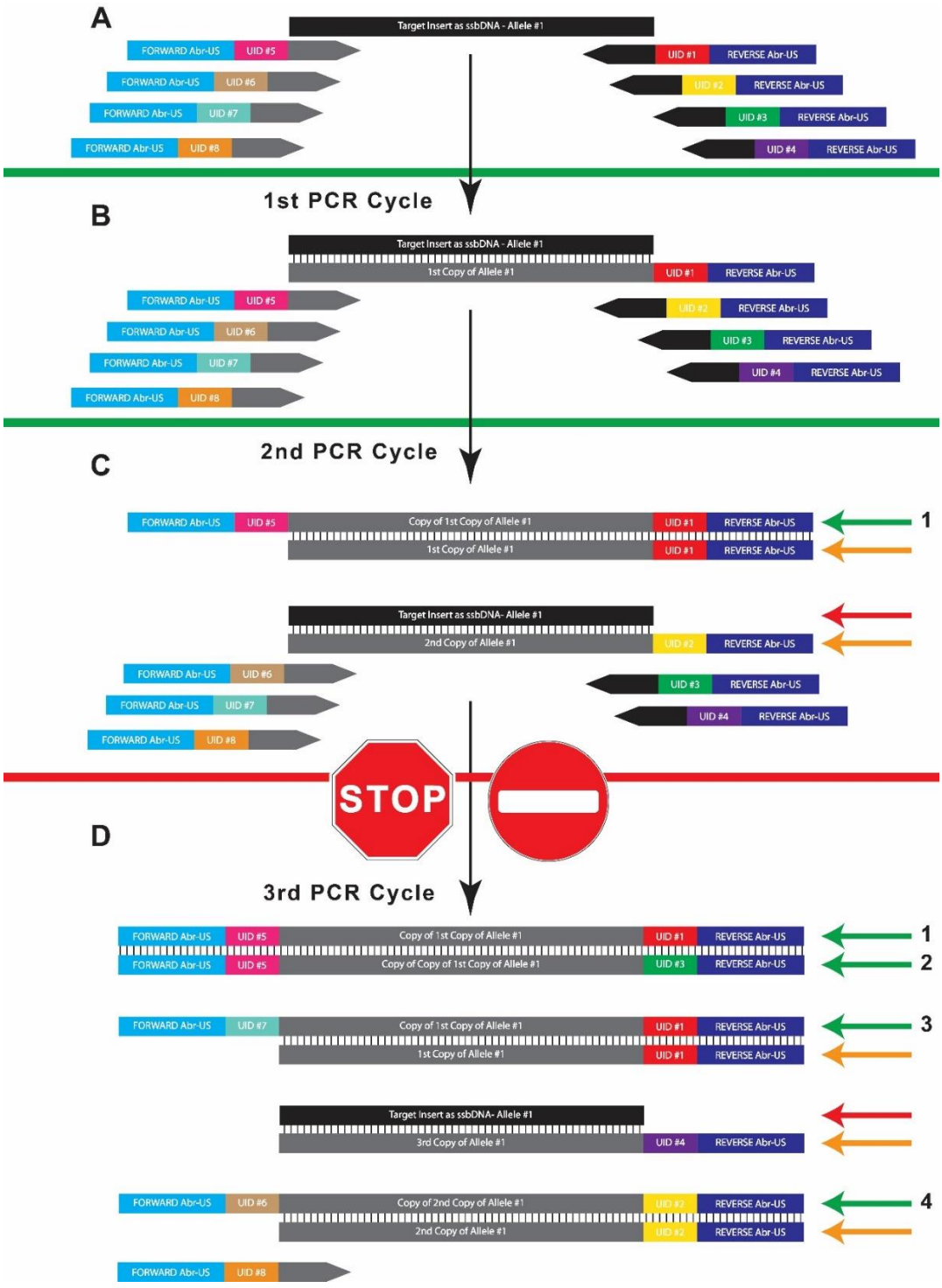

Figure 5.6: *Influence of PCR cycle number on UMI assignment. Full legend overleaf.*

Figure 5.6 *(prior page): Schematic representation of the influence of PCR cycle number on UMI assignment. (A) The starting reaction contains many allele copies of single-stranded bisulfite-converted DNA (ssbDNA), only one allele is shown for the purposes of this figure, the forward and reverse abridged-adapter universal sequence (Abr-US) UMI primers. Each primer has an individual UMI (#1-8 respectively). (B) During the 1st PCR cycle only a single reverse primer (and hence single UMI) can bind to the ssbDNA allele to generate a copy of the template, now as double stranded DNA (dsDNA) with a UMI/Abr-US overhang at the 5' end. (C) During the 2nd cycle of PCR, a single forward primer (in this case UMI #5) can now bind to the copy of the ssbDNA template generated in the 1st cycle to generate a full amplicon sequence with dual ended UMIs/Abr-US (deUMI). This strand is now fully capable of exponential amplification in a 2nd round PCR (green arrow) using universal primers directed to the 5' and 3' Abr-US. Two further strands with single-ended UMI/Abr-US (seUMI) exist (orange arrows) as copies of the original template ssbDNA allele after the 2nd PCR cycle but these would only undergo linear amplification in 2nd round PCR and hence be outcompeted by amplification of the deUMI strand. The original ssbDNA template would not amplify in a second round PCR (red arrow). Therefore, after two cycles of PCR, assuming 100% efficiency of the reaction, every allele would be represented by a single deUMI strand. (D) Demonstrates the problem with a 3rd (or more) cycle of PCR during the UMI assignment phase. This would result in the transformation of the two further seUMI strands in (C) into deUMI strands capable of exponential amplification in 2nd round PCR. Hence the original ssbDNA allele is now represented by four deUMI strands all with a different combination of UMIs. This would cause an over-representation of the allele count in the final analysis and a loss of allele specificity information.*

5.4.4 Testing of individual UMI primer sets

Optimisation in establishing this protocol was initially focused on a select few of the original gene target primer sets: *2-NKX2-5; 2-MYOD1; CSRP3, NPPB, MYO18B, PXDNL, CAMK2B, ANKRD*. These targets provide a good range of amplicon size, CpG density and efficiency in the original bisulfite specific PCR (**section 5.3.4**). **Figure 5.7** demonstrates successful amplification of these eight targets in single-plex, separated on a 1.5% agarose gel. Note that primer dimers remain abundant despite attempts to reduce these as detailed previously.

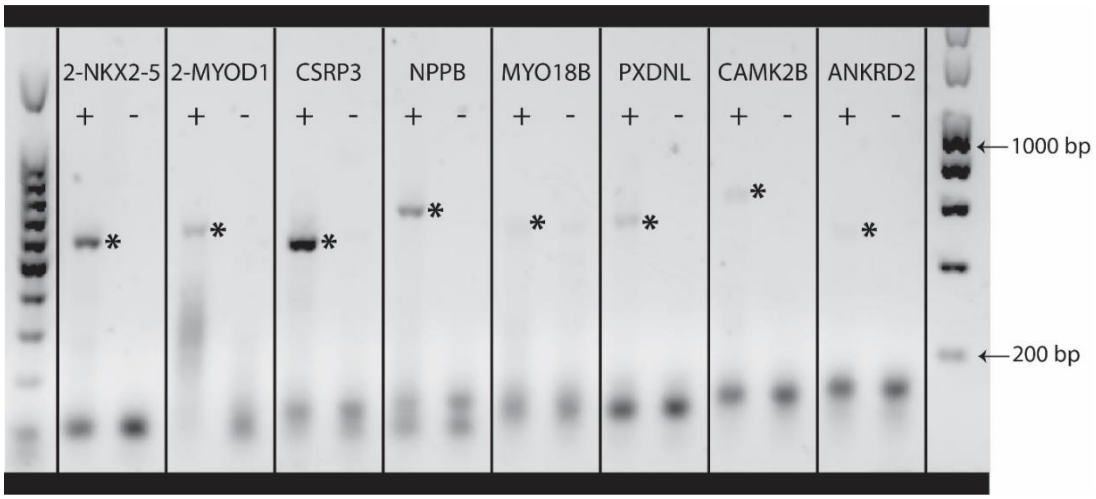

***Figure 5.7**: Example of successful target enrichment directed against bDNA utilising ASM-Seq protocol. Eight targets are shown with their bands marked by asterisks (2-NKX2-5 [512bp], 2-MYOD1 [558bp], CSRP3 [493bp], NPPB [592bp], MYO18B [517bp], PXDNL [529bp], CAMK2B [606bp], ANKRD [468bp]). Lanes are paired for each gene with bDNA template on left and no-template-control on right. A 100bp and 1kb Hyperladder™ respectively flank the gel.*

5.4.5 Optimisation of ASM-Seq for multiplexing

Thus far all steps in the development of the novel ASM-Seq technique were designed and optimised in singleplex, that is, one primer set per PCR well. Until the initial iterations (for example two-cycle 1st round PCR) of the protocol were successful with one primer set, there was little value in further complicating the technical issues such as primer dimers, concatemers, off-target product and outright PCR failure by multiplexing.

Note however, initial combinations of 3 primer sets (2-MYOD1, CSRP3, 2-NKX2-5) demonstrated in **Figure 5.5** were promising, however, on further addition of primer sets, PCR product from a bDNA template became increasingly temperamental and inconsistent with a high rate of failure especially in the context of sensitivity testing low input concentrations of DNA necessary for later BO gland ASM-Seq. Additionally, the lack of gene target amplification was coupled with excessive non-target product, usually <300 bp in length (shortest amplicon in gene panel is 1-MYOD-1 outer set #1 at 367 bp). This is in line with the randomness of bisulfite induced DNA fragmentation and degradation that can result in 90% of DNA affected[324, 342].

Serendipitously, the EM Conversion module (NEB, Massachusetts, USA) has recently been released to market. As described in **section 4.3.2** the protocol involves and oxidation and glycosylation of 5mC catalysed by TET2 to protect cytosine from second step deamination by APOBEC. There is little to no degradation or fragmentation of the DNA which also improves long-range PCR of eDNA amplicons up to 2945 bp in length compared to maximal 1181 bp for traditional bisulfite conversion[343].

An in-house comparison test between bDNA and eDNA inputs was performed down the full finalised ASM-Seq protocol. The differences were in favour of an eDNA template and are presented in the accompanying **Figure 5.8** taken from the subsequent TapeStation analysis of products. The added benefit was improved success on sensitivity testing down to input DNA amounts of as little as 1 ng, equivalent to ~160 copies of the genome. All subsequent DNA processing was completed using the EM Conversion module protocol. Note however, that prior to

this switch a batch of sensitivity and methylation gradient samples on a multiplexed bDNA template had been submitted for analysis and are presented in **section 5.6**.

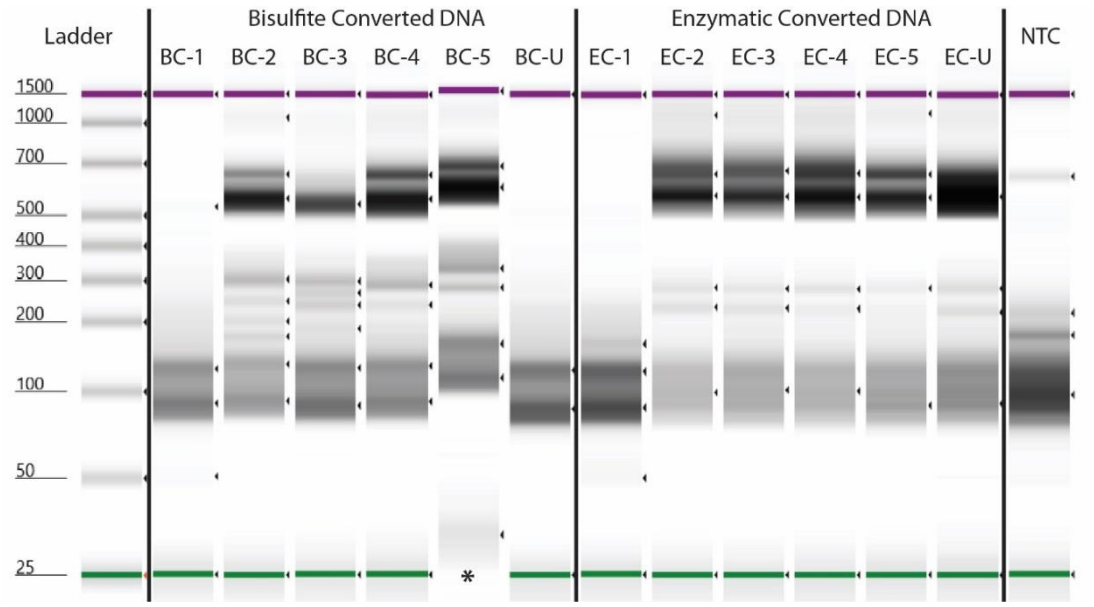

Figure 5.8: *This TapeStation report compares the relative success of the ASM-Seq technique in multiplex PCR when using either a starting DNA template that has been bisulfite converted (left) or enzymatically converted (right). 6 converted DNA samples were tested under each condition (note that BC-1/BC-U and EC-1 unfortunately failed) against a negative control on the far right. The bands of interest are between 526-732bp targeted against 12 genomic regions. Bands outside this range are primer dimers, concatemers and non-specific product, they remain present on this gel as the final AMPure bead clean-up step prior to pooling for sequencing has not yet occurred. Note that enzymatic conversion provides a more balanced representation of the multiplexed targeted amplicons with a reduction in off target products. [BC-# = bisulfite converted Multiplex Pool #; EC-# = enzymatically converted Multiplex Pool #; BC-U = bisulfite converted unmethylated control; EC-U = enzymatically converted unmethylated control; NTC = non-template control; * = technical error with TapeStation where lower marker was not detected appropriately]*

In optimising the multiplex reactions various permutations of primer set pools were trialled based on primer pair individualised success during previous optimisation. These pools were subjected the ASM-Seq protocol with Multiplex Pool (MP) #5 ultimately being selected for future optimisation and the BO glandular work. The full range of pools are given in **supplementary table S3.**

MP #5 consists of 15 out of the total targets designed, in descending order of size: SCN5A, CAMK2B, 1-NKX2-5, NPPB, 1-TNNI3, LOC-L, 2-MYOD1, SBK2, PXDNL, 2-NKX2-5, 1-SBK3, CSRRP3, 2-TNNI3, ANKRD2, 1-MYOD1. The range of amplicon sizes is 367 bp to 573 bp (mean 488, median 489). The CpG locus coverage is 583 (range 16-78) if all targets are successful. A multiplex pool of separate reverse and forward aliquot primers was mixed and diluted to a stock concentration of 0.625 µM concentration for each primer. 0.5 µL of this solution is used in each PCR reaction, equating to ~15 nM concentration of each primer in each reaction well. This is a marked reduction from the singleplex optimisation concentration (200nM) in efforts to further reduce primer dimers. Total volume was also reduced to 20 µL reactions to increase the relative concentration of template DNA in solution.

5.4.6 Multiplexed gene target UMI assignment – 1st round

The 1st round optimised reagent and PCR conditions for multiplexed ASM-Seq are detailed in **tables 5.4a-b**. Herein, all protocol conditions will be described for multiplex ASM-Seq.

Reagent	**Volume (µl)**
Phusion U 2X Master Mix	10
Gene Specific Reverse Abr-US UMI Primers Master Mix – MP #5	0.5 [~15 nM per primer]
Nuclease-Free-Water	Up to 9.5 [Variable]
Template (bDNA)	Up to 9.5 [Variable]
TOTAL	20

***Table 5.4a**: Reagent constituents for 1st round UMI assignment PCR in multiplex ASM-Seq. Note the gene specific forward primer no longer forms part of the 1st round reaction when in multiplex.*

Phase	**Cycle Number**	**Step**	**Temperature**	**Time**
Hot Start	1	Denature	98 °C	1 minute
Slow Ramp UMI Assignment Phase	1	Denature Annealing Extension	98 °C 58 °C – Slow Ramp by 0.2°C/second 72 °C	10 seconds 5 Minutes 30 seconds
Final Extension	1	Extension	72 °C	5 minutes
Completion	1	Hold	4 °C	Forever

***Table 5.4b**: One-cycle PCR Thermocycling conditions for 1st round Multiplex ASM-Seq. The annealing temperature is unchanged from singleplex protocols and again a slow ramp is mandated to mimic the touchdown PCR effect identified during early phase optimisation.*

5.4.7 Removal of the gene specific reverse Abr-US UMI primers

Removal of the gene specific reverse primer, any primer dimers and concatemers formed in 1st round PCR is crucial to success of this protocol to prevent UMI retagging of original DNA template molecules and to also shift the balance of 2nd round PCR towards the larger and less abundant targeted amplicons. The optimised protocol uses 3 µl (60 units) per well of Exonuclease I[326] (NEB, Massachusetts) to digest linear single-stranded DNA in the 3' to 5' direction which includes the original bDNA or eDNA template. The enzyme is inactivated at 80°C for 20 minutes (**Tables 5.5a-b**) to prevent carryover activity to 2nd round PCR. Confirmation of Exonuclease I efficiency to degrade the remaining UMI primers is shown in **Figure 5.9** whereby failure to perform this clean-up results in a product amplified by the 1st round UMI primers in 2nd round PCR.

Reagent	**Volume (µl)**
Post 1st Round PCR Product	20
Exonuclease I	3 [60 units]
TOTAL	23

***Table 5.5a**: Reagents for exonuclease I clean-up step*

Phase	**Cycle Number**	**Step**	**Temperature**	**Time**
Digestion	1	Incubate	37 °C	30 minutes
Termination of Reaction	1	Denature	80 °C	20 minutes
Completion	1	Hold	4 °C	Forever

***Table 5.5b**: Thermocycling conditions for exonuclease I clean-up step*

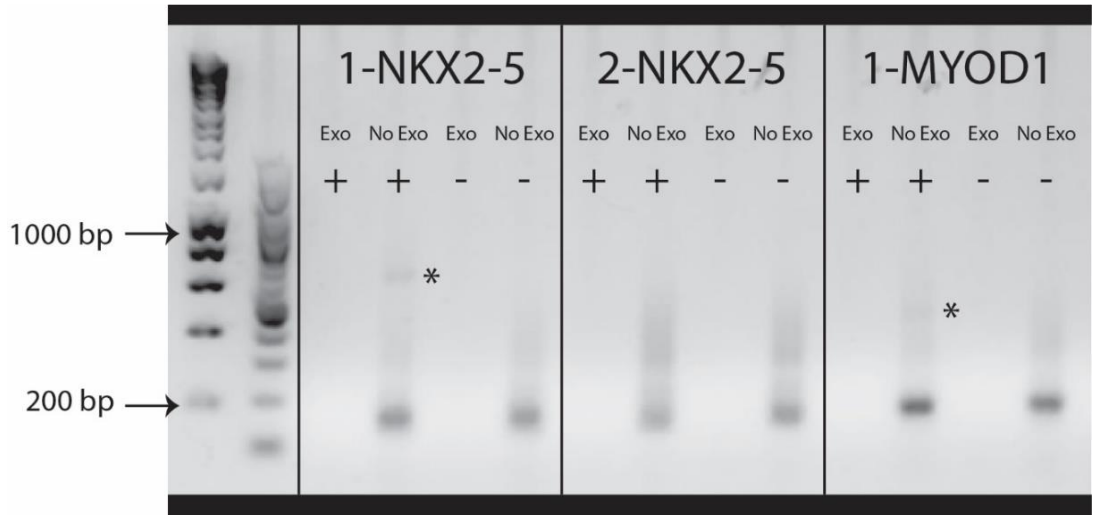

***Figure 5.9**: Importance of exonuclease I clean-up between 1st and 2nd round PCR. The amplification products from three gene targets are shown (1-NKX2-5, 2-NKX2-5, 1-MYOD1) on an agarose gel alongside a 1kb and 100kb Hyperladder™. This test was conducted with both the forward and reverse Abr-US UMI primers in a singleplex two-cycle 1st round PCR. All reactions proceeded from 1st round PCR into a 40 cycle PCR. Each gene target has four lanes: i) bDNA template with exonuclease I clean-up between 1st round PCR and the 40 cycles; ii) bDNA template with no exonuclease I clean-up; iii) no-template control (NTC) with exonuclease I clean-up between PCR rounds; iv) NTC with no exonuclease I clean-up. A lack of exonuclease I clean-up between PCR rounds results in a target specific product (faint bands for 1-NKX2-5 and 1-MYOD1, asterisks) amplified by the 1st round Abr-US UMI primers. This would affect the final allele specific methylation analysis (**Fig. 5.6**) if exonuclease I clean-up is not conducted. No clear band is visualised in lane 2 for 2-NKX2-5, this is likely to represent a low yield undetectable product as the 2nd round PCR amplification conditions here were not optimised for the 1st round primers. Lanes 1 demonstrate exonuclease I efficiently digests the 1st round primers and hence no product is seen on the gel. The presence of a lane 2 band (asterisks) in targets 1-NKX2-5 & 1-MYOD1 also confirms that exonuclease I is appropriately denatured and inactivated.*

5.4.8 Target enrichment and pre-amplification PCR – 2nd round

For each designed reverse Abr-US UMI primer (**section 5.4.2**) a 2nd round primer complementary to the universal sequence was designed and tested, called SP2. Furthermore, phosphorothionate bonds are incorporated between the final three bases of the universal primers at the 3' end. This is to prevent their degradation should there be any residual exonuclease I activity[326].

2nd round PCR target enrichment and pre-amplification is undertaken for 10 cycles (**Tables 5.6a-b**). The 23 µl post exonuclease I 1st round reactions are combined with a 7 µl master mix as detailed in **Table 5.6a.** This master mix includes 5 µl of Phusion U Multiplex PCR 2X Master Mix (Thermofisher Scientific, UK), 0.5 µl of the gene specific forward Abr-US primers multiplex mix (equating to ~10 nM per primer per reaction) and ~333nM (1 µl 10 µM stock) each universal primer to make 50 µl reactions. Following PCR, AMPure XP bead clean-up at a 0.9X proportion is conducted to remove the forward Abr-US primer. Samples are eluted into 12.5 µl of nuclease free water with 11.5 µl of the supernatant immediately carried over to amplification PCR (3rd Round).

Reagent	Volume (µl)
Phusion U 2X Master Mix	5
Gene Specific Forward Abr-US Primers Master Mix – MP #5	0.5 [~10 nM per primer]
Abr-US Reverse Primer – SP2	1 [~330nM]
Nuclease free water	0.5
Post 1st Round Clean-up Products	23
TOTAL	30

***Table 5.6a**: Reagents for 2nd round target enrichment and pre-amplification PCR*

Phase	Cycle Number	Step	Temperature	Time
Hot Start	1	Denature	98 °C	1 minute
Amplification	10	Denature Annealing Extension	98 °C 58 °C – Normal Ramping 72 °C	10 seconds 30 seconds 30 seconds
Final Extension	1	Extension	72 °C	5 minutes
Completion	1	Hold	4 °C	Forever

***Table 5.6b**: Thermocycling conditions for 2nd round target enrichment and pre-amplification PCR*

5.4.9 Amplification PCR – 3rd round

3rd round PCR follows a more conventional PCR thermocycling protocol (**Tables 5.7a-b**). The post AMPure beads 11.5 µl samples are combined with a 13.5 µl master mix of Phusion U Multiplex PCR 2X Master Mix (Thermofisher Scientific, UK) and 200 nM (0.5 µl 10 µM stock) of each universal primer, the SP2 and also SP1 that is complementary to the universal sequence overhang at the 5' end of the gene specific forward Abr-US primer. Following PCR, a second AMPure XP bead clean-up at a 0.7X proportion is necessary. Samples are eluted in 15 µl with a total of 14 µl removed from the supernatant. 4 µl of this is used in subsequent library preparation indexing, the remaining 10 µl are stored at -20°C and can be re-used or re-purposed if or when necessary.

Reagent	Volume (µl)
Phusion U 2X Master Mix	12.5
Abr-US Forward Primer	0.5 [200nM]
Abr-US Reverse Primer	0.5 [200nM]
Post 2nd Round AMPure Bead Cleaned-up PCR Products	11.5
TOTAL	25

***Table 5.7a**: Reagents for 3rd round amplification PCR*

Phase	Cycle Number	Step	Temperature	Time
Hot Start	1	Denature	98 °C	1 minute
Amplification	40	Denature	98 °C	10 seconds
		Annealing	60 °C – Normal ramping	30 seconds
		Extension	72 °C	30 seconds
Final Extension	1	Extension	72 °C	5 minutes
Completion	1	Hold	4 °C	Forever

***Table 5.7b**: Thermocycling conditions for 3rd round amplification PCR*

5.4.10 Library construction for sequencing

In total, sixteen Index 500 (P5-i5) series and twenty-four Index 700 (P7-i7) series primers were designed with their 3' end targeted to the 5' universal sequence of the 3rd round PCR product. The combination of indices permits a potential multiplexing of 384 samples on a single sequencing run, assuming an appropriate sequencing depth per sample is available and achievable. These library preparation primer sequences for the ASM-Seq protocol are available in the **supplemental table S4.**

Q5® High-fidelity 2X Master Mix (NEB, California) 12.5 µl is combined with 200 nM (2 µl of 2.5 µM stock) of each P5-i5 and P7-i7 primers, 4.5 µl of nuclease free water and 4 µl of the cleaned-up post 3rd round PCR product. The thermocycling conditions are in **Tables 5.8a-b**. The resultant product is cleaned-up with AMPure XP beads at 0.7X proportion and eluted in 40 µl of nuclease free water. Samples are now deemed *library ready* and suitable for sequencing on the Illumina platforms.

Reagent	Volume (µl)
Q5® High Fidelity 2X Master Mix	12.5
P5-i5 Library Primer	2 [200nM]
P7-i7 Library Primer	2 [200nM]
Nuclease free water	4.5
Post 3rd Round AMPure Bead Cleaned-up PCR Products	4
TOTAL	25

***Table 5.8a**: Reagents for library preparation PCR*

Phase	Cycle Number	Step	Temperature	Time
Hot Start	1	Denature	98 °C	30 Seconds
Library Preparation and Amplification	5	Denature Combined Annealing & Extension	98 °C 65 °C	10 seconds 75 Seconds
Final Extension	1	Extension	65 °C	5 minutes
Completion	1	Hold	4 °C	Forever

***Table 5.8b**: Thermocycling conditions for library preparation PCR*

5.4.11 Preferential incorporation of appropriate adapter sequences

The Abr-US design consists of 20 base pairs of which 13 bases at the 3' end of the forward and reverse primers share sequence homology. Hence, there is potential in the 3rd round library preparation for amplicons to be tagged with a P5-i5 or P7-i7 at both the 5' and 3' ends. It is important to note that this issue exists for all commercially based library preparation methods also. However, a trial comparing 4th round amplification of 3rd round product against P5-i5/P7-i7 or P5-i5/P5-i5 or P7-i7/P7-i7 primer pools demonstrates superior amplification efficacy of the P5-i5/P7-i7 pool (**Fig. 5.10**). Furthermore, a dual ended P5-i5 or P7-i7 amplicon would also fail to sequence appropriately on the flow-cell.

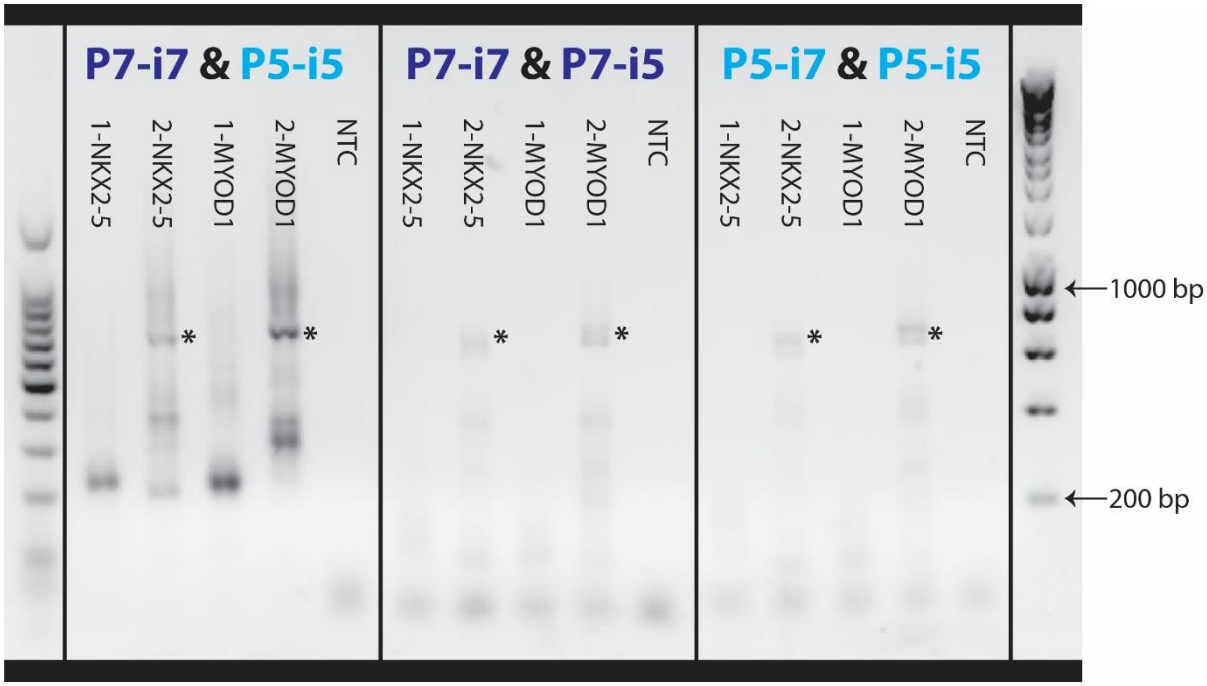

***Figure 5.10**: 1.5% agarose gel depicting the preferential incorporation of P7-i7 and P5-i5 sequences during library preparation compared to dual-ended P7-i7 or P5-i5. Four gene targets are shown (1-NKX2-5, 2-NKX2-5, 1-MYOD1, 2-MYOD2) in single-plex with a post-3rd round PCR template compared against a no-template-control (NTC). Only gene targets 2-NKX2-5 & 2-MYOD1 were successful in this instance, bands marked by asterisks. Band intensity is greater in the P7-i7/P5-i5 column vs P7-i7/P7-i7 or P5-i5/P5-i5 columns demonstrating greater yield of Illumina® compatible product and a greater amplification tendency towards a library with the correct final structure.*

5.4.12 Sanger sequencing

Sanger Sequencing was employed to demonstrate the entire amplicon structure (**Fig. 5.11a**) is correct including the variable UMI sequence evidenced by multiple chromatograms representing the four nucleotides in this region (**Fig 5.11a**). Furthermore, when a differentially methylated template is amplified and sent for Sanger sequencing, superimposed cytosine and thymine chromatograms can be seen at distinct CpG loci (**Fig. 5.11b, asterisk**). Finally, amplicon product sequences align appropriately with the gene specific target sequence enriched for in the PCR (**Fig. 5.11b**).

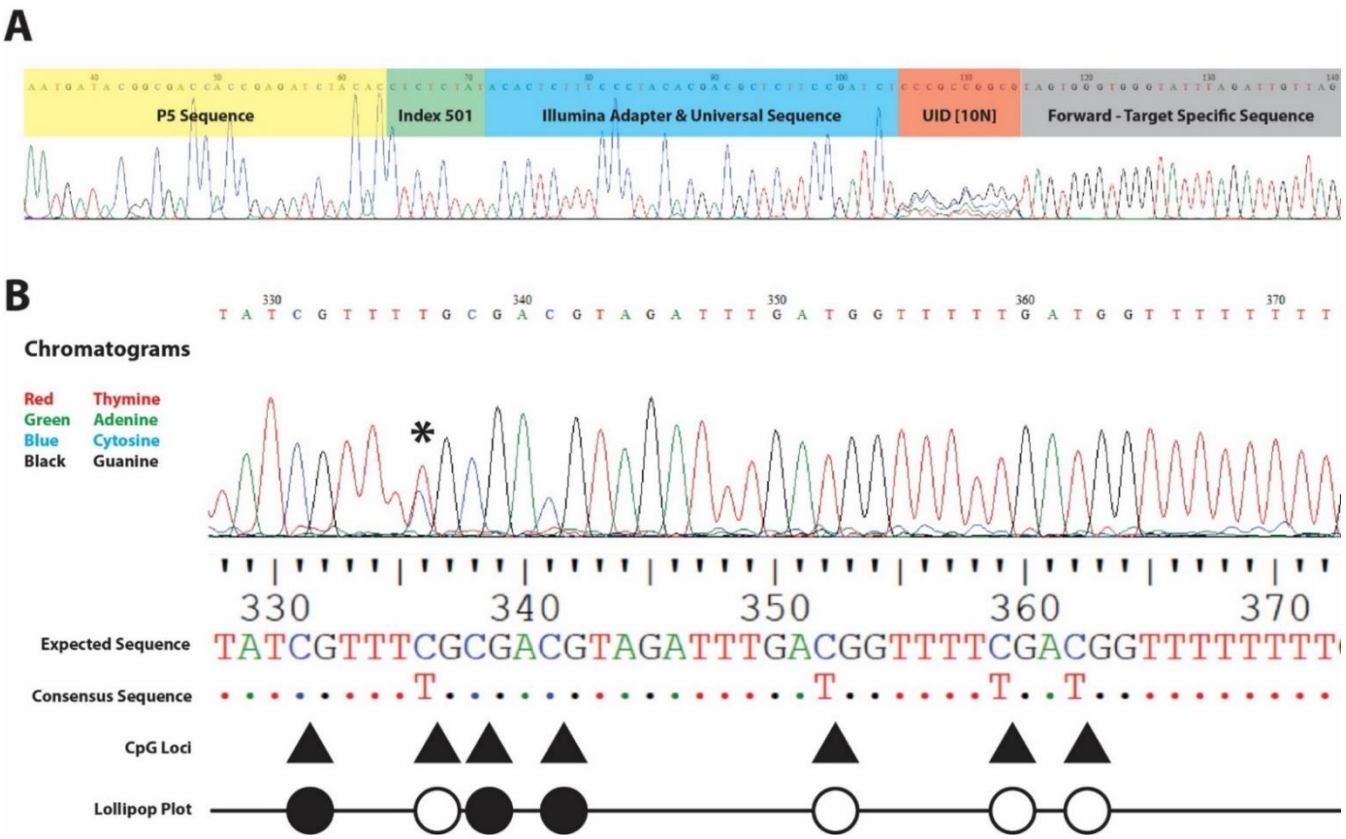

***Figure 5.11**: (A) Sanger sequencing confirms the presence of an appropriately structured amplicon compatible with the Illumina® systems. The UMI in this instance is in the format of a series of 10 Ns and is represented by a mixture of the four nucleotide chromatograms at varied amplitudes. (B) A sanger sequencing example showing a 45 bp stretch of the 2-MYOD1 gene target. Note that the consensus sequence aligns appropriately with the expected sequence only differing at CpG loci where thymine (T) is present instead of cytosine (C). This differential sequence resolves the methylation pattern (lollipop plot) of this 45 bp stretch where 7 CpGs are located (arrowheads). Also note that the second CpG locus has a mixed T and C chromatogram (asterisk) demonstrating the heterogenous methylation state of the pool of amplicons within the sample.*

5.4.13 Preparation for submission for next generation sequencing

Qubit® quantification and quality control 4200 TapeStation analysis on HS D1000 Screentapes® (Agilent Technologies, Santa Clara, CA) is undertaken. This allows visualisation of multiplexed PCR products as bands seen between the expected target range for the specific library ready amplicons. Additionally, an electropherogram product distribution can be overlaid with the main region of interest, determined as between 450 bp – 900 bp to account for the additional UMI, Abr-US and Illumina® flow cell adapter sequences. Here, samples that had poorly amplified, had poor electropherographical distribution or without clear peaks within the expected range were excluded from further down-stream pooling and sequencing. Samples are then pooled in equimolar concentrations to a total 4nM concentration. A calculation to take into account the initial measured Qubit measured gDNA copy number appropriately assigns enough reads per target (15 targets) per copy number to normalise samples across the spectrum and avoid sequencing bias in the final data output. Pooled libraries were submitted for Illumina® MiSeq v3 300 bp PE sequencing. A single cohort of bDNA control samples was submitted for Illumina® NovoSeq 150 bp PE sequencing prior to the switch to exclusive eDNA conversion. Despite 150 bp PE reading not covering the full length of target amplicons the purpose here was to assess technique sensitivity and methylation density analysis and design the bioinformatic workflow rather than understand the full methylation tag. Pooled libraires were spiked with 20% PhiX to generate enough background cluster diversity to optimise base calling of the targeted amplicon sequencing. The downside of PhiX is reduction in sequencing power and depth as at least 20% of total reads are assigned to it, a loss of 5,000,000 on the MiSeq v3 300 bp PE platform.

5.4.14 Summary schematic of protocol from tissue to data output

Figure 5.12 overleaf gives a summary schematic of the workflow for the allele specific methylation bisulfite sequencing (ASM-Seq) protocol from eDNA sample to libraries ready to be pooled for sequencing.

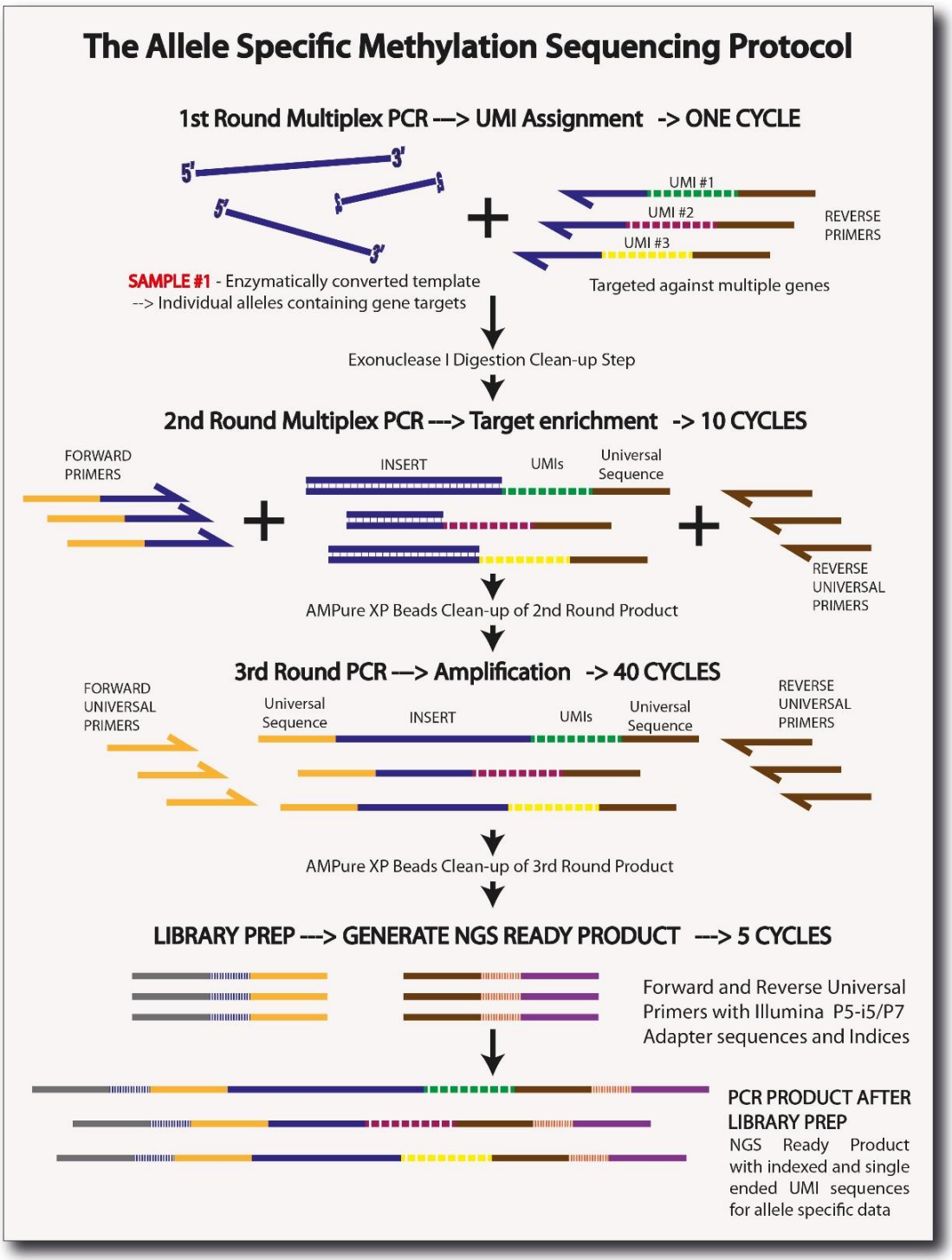

Figure 5.12: *The Allele Specific Methylation Sequencing Protocol. This schematic represents the key steps of the novel ASM-Seq technique described to this point in the main text.*

5.5 Bioinformatics workflow

5.5.1 Overview

The novel ASM-Seq protocol, that combines UMIs to a bDNA or eDNA template for the first time, requires a new specific bioinformatics pipeline to process the significant volumes of massively parallel NGS data into a more recognisable format for subsequent clonality, stem cell and mitotic clock analysis.

The R1 and R2 sequencing data is downloaded and decompressed as Fastq files. There are two key stages to the designed pipeline. First, a pre-analysis quality control, trimming and filtering processing pathway occurs with subsequent UMI family identification, clustering and calling of consensus and base call CpGs which is translated into a more user friendly format. The data is saved as text files demonstrating arrays of binary code representing the underlying methylation tags. The second stage interprets these matrices to further reformat them and divide the data into the key components of total read depth, UMI information, raw consensus aligned methylation tags, unique consensus methylation tags, methylation density tables and mean pairwise distances between methylation tags both within and between samples.

The pre-analysis processing is written as shell script with the subsequent analytical processing script written in the Python 3 language. All coding scripts and data was uploaded to the Queen Mary University High Performance Computing Cluster (HPC) called Apocrita[344] and an individualised initiator script was written to commence the two step processing.

The master scripts are available for review, critique and editing at:

https://github.com/hacket01/ASM-Seq-Files.

5.5.2 Designing and actions of the pre-analysis pipeline

5.5.2.1 Quality control

The R1 and R2 fastq files are first processed by FastQC[345] which performs simple quality control checks on the data to provide an overall impression of whether the protocol has been successful or now. In particular, single base phred scores define the probability of an incorrect sequencing base call, scores of 20 represent a 99% chance of correct base call with a score of 30 representing 99.9% and probabilities improving by a factor of 10 for every 10 points higher phred rating[345]. As Illumina SBS cycles progress generally phred scores deteriorate with data at the end of the 150 bp or 300 bp paired end reading being of less reliability if it is even called at all. These areas represent the middle of the amplicons in the ASM-Seq protocol with the longer amplicons more prone to having a region of miss called bases or absence of bases centrally. FastQC also provides plots for GC % content and per base sequence content which demonstrates expected high levels of Adenine and Thymine residues The GC % are more variable given the amplicons are targets at CpG islands. This base sequence variability leads FastQC to determine that the sequencing is of poor quality when in fact it is consistent with a bDNA or eDNA template and the gene regions targeted.

5.5.2.2 pRESTO

pRESTO[346] is a bioinformatic toolkit that can process raw sequencing reads from multiple data types to sort sample and convert the data where necessary. Crucially, pRESTO has an inbuilt function to manage UMIs. Multiple opensource software packages capable of managing UMIs were explored in the optimisation of bioinformatics pipeline but none performed quite to the standard or flexibility offered by pRESTO. Such trialled packages included Bartender[341], UMI-tools[347], UMIc[348], FgBio[349] and Starcode Master[350]. All these packages have been designed and optimised for UMI gDNA or RNA sequencing which typically with shorter range

single or paired end reads and appropriate sequence complexity along the length of the contig important for the underlying coding algorithm that uses k-mer stretches of sequence to aid genome alignment and hamming distances or inherent similarities to define the probability of two sequences coming from the same UMI family suitable for clustering. The characteristic AT-rich sequences and loss of genomic diversity causes many open source packages to stumble at the first hurdle being unable to even align sequences properly to either a reference genome or themselves[351]. Alternatively parameters surrounding the construction and location of the UMI sequence were too rigid and prohibited further progress, for example, UMIc[348] would not permit a seUMI on the R2 (reverse sequences) alone without a corroborating R1 (forward) UMI sequence. It was not possible to manipulate the underlying code sufficiently to rectify this issue.

pRESTO offered a flexible environment to complete the bioinformatic processing up to and including UMI consensus reads. The FilterSeq.py script is used to remove any sequences with mean quality phred scores of <20 followed by 3' trimming of all poorly called bases (mean phred score <20) in a 5' direction that as stated represent the middle of the amplicon. It was noted during the FastQC analysis that a number of sequences were abruptly curtailed with multiple >100s bases missing from the output data. They were usually of erroneous base calls, or sometimes adapter sequences and would always fail an alignment step but the issue arose when performing UMI processing with the algorithm expecting sequences all of a certain length. Thus, to deal with this issue, a further FilterSeq.py pass is performed to discard all sequences <100 bases in length. Pairseq.py script is then utilised to reunite the mates 1 (R1) and mates 2 (R2) sequences into mate pairs using the Illumina coordinates in the read headers. The UMI sequences are then extracted and added as an annotation to the headers of both the components of the mate pair. Reads that do not meet this structure will later fail alignment or clustering within UMI families. 20 bases of the 5' primer sequences are then removed as additionally there can be quality loss in this region, especially where degenerate bases (Y or R) were used in the primer design, that can result in failure to cluster appropriately into a UMI family due to hamming distances that fall outside the threshold criteria under the

threat of an erroneous base call. The shortening of the reads here has no bearing on the overall CpG coverage as the primers were designed away from such sites. UMI clustering and collapse of all sequences within a UMI family occurs to form a unique consensus read, additionally the read count that forms that consensus read is maintained in the final data output. This is achieved through the BuildConsensus.py function and setting a maxerror of 0.1 which sets the maximum threshold for sequencing differentials between two sequences (equivalent to hamming or pairwise distance). Variations of maxerror were used in the optimisation process of this pipeline, in general, a higher max error is less stringent in terms of integrity of calling consensus reads but with greater potential numbers, with the opposite being true where maxerror is excessively low and tightly controlled (data not shown but available on request). In this bDNA/eDNA template it is important to recognise that some leniency is necessary given the greater propensity for base calling mismatching (despite addition of PhiX) that could excessively remove otherwise perfectly reliable sequences. Additionally, there is no clear precedent set for this maxerror rate as UMI use in bDNA/eDNA has not been done before. Following generation of the unique consensus reads matched along the length of the full amplicon the mates 1 and mates 2 are separated back to individual data files to permit alignment to the bisulfite genome.

5.5.2.3 Bismark

The Bismark opensource package is a bioinformatics tool that allows alignment of bisulfite (or equivalent) sequenced samples to the bisulfite reference genome and subsequent methylation status calling[352]. The two UMI consensus reads fastq files are passed from pRESTO to Bismark where they are aligned to a bisulfite converted genome. To speed up this process, rather than Bismark checking alignment against the entire genome, a custom fasta file was created for Bismark to use solely with the sequences of the specific gene targets of interest. Bismark attempts alignment against all permutations of the bisulfite converted and non-converted positive and negative strands of DNA which do not have complementary sequences. Once aligner,

the bismark_methylation_extractor function is used to generate the methylation base calls along the length of the re-paired forward and reverse sequences. This data is outputted into a single text file where methylated residues are denoted “1” and unmethylated residues are “0” alongside their CpG locus location within the specifically named gene target amplicon. Bismark also generates a bam file that can be used to visualise the methylation called on the Interactive Genomics Viewer (IGV) genome browser software package (**https://software.broadinstitute.org/software/igv/**) amongst many others. The Samtools package (**http://www.htslib.org/**) is necessary to finalise the sorting and indexing of the bam file for visualisation.

5.5.3 Post processing data analytics

The text files give the raw methylation sequencing data though binary CpG calls. However, it remains in a cumbersome and poorly organised state, albeit organised into single files of by sample with long column lists of seemingly random 0’s and 1’s within.

5.5.3.1 Python 3 reorganisation

The individual text files are split into their aggregate parts with removal of excess and irrelevant syntax data. The binary elements are sorted by their accompanying gene target data first and then numerical CpG loci. The counts that reflect the number of individual reads forming a consensus UMI read are extracted for later appending to the reorganised data set. The data exists cohorted into its individual gene targets, thus a function is defined to re-organise into a format akin to the lollipop plots demonstrated in Chapter 1 where each row represents a single methylation tag and each column a CpG locus within the overarching gene target e.g. 1-MYOD1 or 1-NKX2-5 within a single sample. The UMI counts are re-appended and the first tranche of raw data is downloaded.

5.1.8.2.2 Dealing with missing data

Despite the careful filtering and trimming described in the pre-analysis pipeline there remains patchy data loss throughout. Most of this is non-random, occurring at particular gene target sites, for instance ANKRD2, CSRP3, SBK2 and 1-TNNI3 are all poorly represented across multiple datasets suggesting an inherent problem with early in the multiplexed ASM-Seq protocol. Note, for example, that CSRP3 had previously performed well in singleplex testing. Other causes of non-random data loss are at the whole sample level. Despite screening out poorly amplified samples using the 4200 TapeStation system some suboptimal samples were still passed to go on to pooling and NGS, ultimately with an expectation they could fail to provide any meaningful data. Additionally, while the 4200 TapeStation provides reasonable visualisation of the PCR product there is little way of knowing if the visualised band and electropherogram represent the actual gene targets or non-target contaminants that would be filtered out in the pre-analysis bioinformatics. There is also persistent middle amplicon (3' ends of the forward and reverse) data degradation in the observed methylation tags. Which again is non-random loss representing exhausting of the Illumina SBS process as it reaches its final cycles. The FilterSeq.py pRESTO quality trimming phred score algorithm had mostly dealt with this poor data but the result is of a heterogenous representation of the middle CpG loci across the longer gene targets. The uneven length of methylation tags has downstream consequences where pairwise distance measurements expect the identical CpG locus to compare against and the same length tag. Further, other missing data elsewhere in the methylation tags can erroneously cause analytic processing failure. Thus, the middle amplicon data for individual gene targets was visually inspected and clearly obvious lack of CpG locus data was excluded for all reads and samples for that gene target.

Resultantly, the frequencies of CpG discard rate per gene target were as follows:

ANKRD2 : 0; CAMK2B : 14, CSRP3 : 0; LOC-L : 6; 1-MYOD1 : 0; 2-MYOD1 : 0; 1-NKX2-5 : 14; 2-NKX2-5 : 0; NPPB : 1; PXDNL : 6; SBK2 : 9; 1-SBK3 : 3; SCN5A : 13; 1-TNNI3 : 9; 2-TNNI3 : 2.

This totals 68 middle amplicon CpGs bring the total coverage of the multiplex panel down to 583.

Following this exclusion remaining missing data was deemed to be *missing completely at random*. Thus a computational function was defined to screen each missing data point (defined as NaN in Python 3 parlance) and make a comparison with all the other data points in the set. In particular, if the sample arose from a control DNA such as 100 % methylated tags or 0% methylated tags then an x-axis (row of the lollipop plot) correction would apply whereby if ≥80% of the other data points on the axis share the same methylation call then the NaN would be replaced with that call (0 or 1). For patient gDNA the axis is inverted to the y-axis (columns) but the same threshold applied. The reason for y-axis correction is that all the reads are ancestral relations of one another and thus the methylation status of the particular CpG locus in question is somatically inherited and more likely to be similar to the other gDNA templates within the sample pool than the CpGs along its own x-axis (row). For any NaN that failed to reach the 80% threshold then entire sequence (x-axis/row of methylation tag or y-axis/column of CpG locus) was discarded.

5.5.4 Final creation of the analytic tables

The dataset has been further filtered and any missing data corrected or discarded further. The final step is to create the unique methylation sequence, intra and intergland pairwise distance and methylation density tables.

5.5.4.1 Unique methylation sequences

Each consensus methylation tag is compared along its entire length with all the other tags within a particular gene target. Where the sequences are exactly the same these tags are further collapsed into unique methylation sequences. The output table is subsequently downloaded.

5.5.4.2 Pairwise distance

Two computation functions are defined to measure absolute aggregate pairwise distance and collapse this data into more manageable mean pairwise distance. This is completed separately on two fronts for BO samples: ***intra***-gland and ***inter***-gland. For intragland all combinations of the cohort of methylation tags within a single gene target are compared along their length CpG by CpG. Where CpG sites are differentially methylated a score of 1 is ascribed with the sum of the total score recorded, e.g. three differentially methylated CpG sites between two methylation tags would give a pairwise distance of 3. Once all combinations have been completed the mean of the sum of the total scores is calculated and recorded on a gene target by gene target basis. For intergland pairwise distance, first all the combinations of individual glands within the sequenced cohort are compared. Then between a pair of glands each individual methylation tag in sample 1 is compared with each individual methylation tag in sample 2. The mean CpG pairwise distance is called as described for intragland comparison. The data of inter-biopsy inter-gland comparisons is reorganised later in a separate script in line with the clinical BO data defining which gland belongs to which patient, which timepoint and location within the BO segment. This all takes exceptional computer processing power, the 384 BO glands that were compared result in over 73,500 individual combinations, however, this results in excellent depth of data to be discussed in the next chapter.

5.5.4.3 Methylation Density

A computational function is defined that calculates mean methylation density across both the x-axis (methylation tags) and the y-axis (CpG locus) for each gene target within each sample. This mean is recorded and populates a separate methylation density table for later review.

5.6 Further Experimental Testing and Validation

5.6.1 Sensitivity testing

The aim of the protocol is to examine Barrett's glands that are composed of a few thousand cells equating to ~5-10 ng of gDNA template. Some, if not >50%, of this template is lost in processing, especially due to bisulfite conversion[334, 353]. Although this should now be mitigated against by switching processing to the enzymatic pathway. Sensitivity demonstrated efficacy of the protocol down to a 5 ng starting template as evidence by bands seen on the agarose gel (**Fig. 5.13**). Thus, assuming satisfactory glandular gDNA yield and extraction, the protocol is sufficiently sensitive.

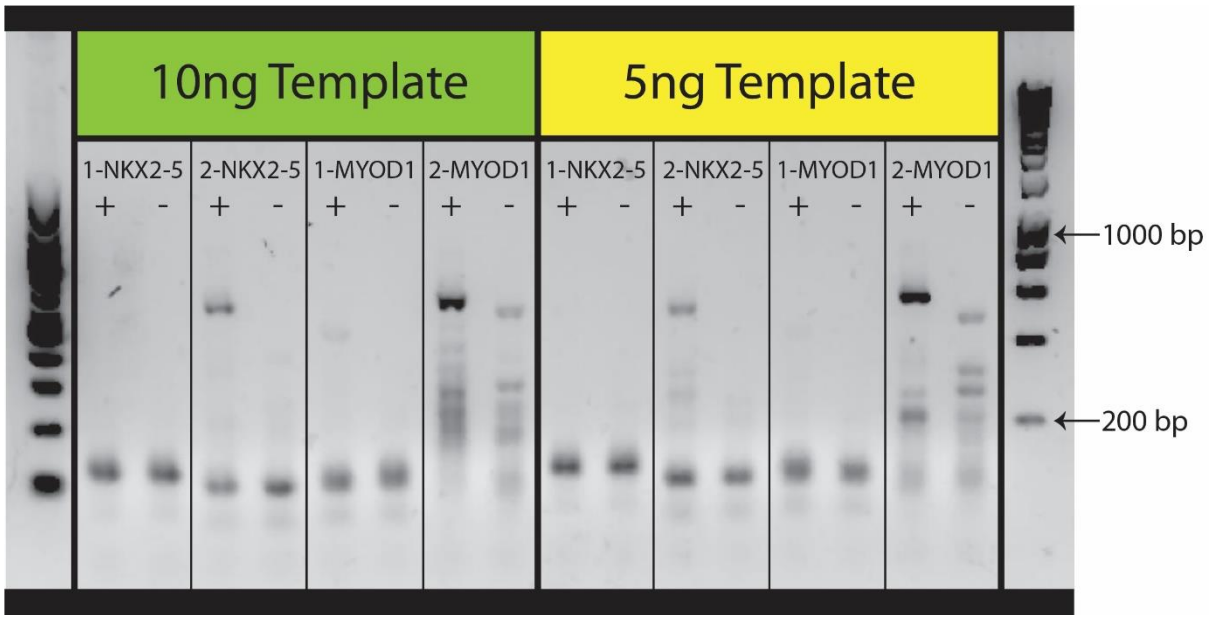

***Figure 5.13**: Sensitivity testing of the ASM-Seq protocol. Successful amplification is seen down to 5ng of starting bDNA template in three of the four gene targets here (2-NKX2-5, 1-MYOD1, 2-MYOD2). It is not clear why the target, 1-NKX2-5, failed to amplify in this instance. Lanes are in pairs with bDNA template on left and no-template-control on the right. A 100bp and 1kb Hyperladder™ respectively flank the gel.*

5.6.2 Sanger sequencing

Prior to the first samples being submitted for NGS, Sanger Sequencing was employed to demonstrate that the entire amplicon structure (**Fig. 5.11a**) had formed correctly and the presence of a variably sequenced UMI. This would be evidenced by multiple nucleic chromatograms representing the 4 bases seen in this region (**Fig 5.11a**). Furthermore, when a differentially methylated template is amplified and sent for Sanger sequencing, super-imposed cytosine and thymine chromatograms can be seen at distinct CpG loci (**Fig. 5.11b, asterisk**). Finally, amplicon product sequences align appropriately on the BioEdit software programme with the gene specific target sequence of *MYOD1* (**Fig. 5.11b**)

5.6.3 Testing exonuclease I efficiency

The degradation of the reverse UMI primer after 1st round PCR is very import to prevent re-tagging of the same original DNA template molecule that would give an erroneously elevated UMI count. To test the efficiency of exonuclease a simple experiment was devised where the 2nd round PCR cycle number was altered from 10 cycles to either 2, 4, 6, 8 or 10 cycles (control). 100% methylated bisulfite converted control DNA (EpiTECT^{TM}, Qiagen) was used as a template for the reactions and ASM-Seq was carried out as before except with this modification. If exonuclease fails to fully degrade the reverse UMI primer then the unique methylation tags in the final data output would increase proportionally with the increasing 2nd round PCR cycles. The ideal result would be a static unique read count from 2 to 10 cycles as a horizontal line on **figure 5.14**. Comparator modelling of PCR conditions with normal exponential amplification and three models of continues presence of reverse UMI primer with expected sequencing output copy number are also shown in **figure 5.14.** An ordinary one-way ANOVA statistical test was performed on the slopes of all the gene targets against the described modelling conditions. All gene targets were significantly (p value <0.001 or <0.0001) different to the normal exponential growth model, however they remained non-significant against the other models of amplification. The findings suggest that there may be a small degree of hangover

reverse UMI primer that is not fully degraded by exonuclease I. However, given the SP2 primer is spiked in the second round at a much greater abundance this would generally outcompete any remaining reverse primer.

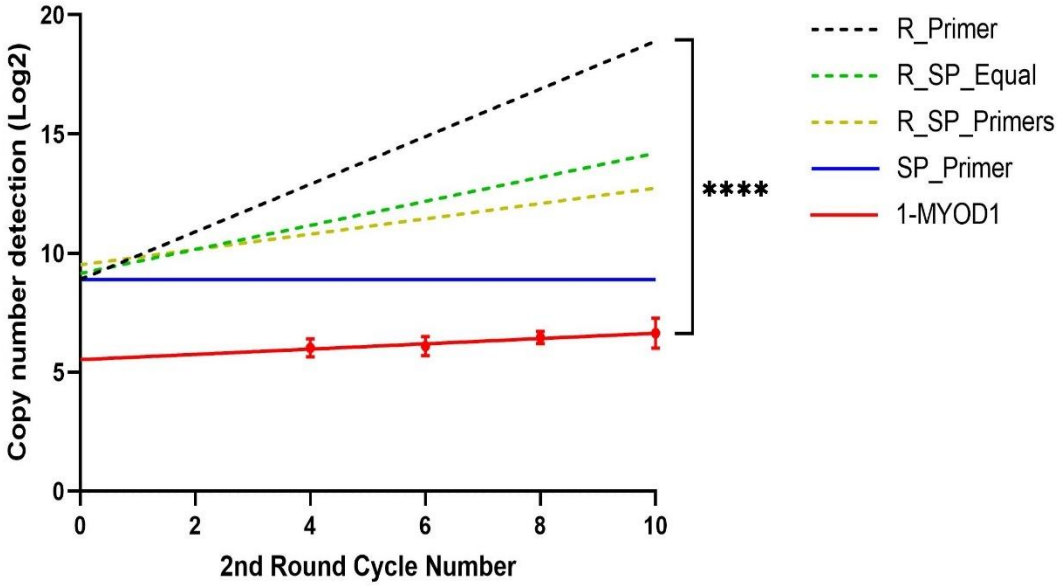

***Figure 5.14**: Testing exonuclease efficiency. A representative plot for the gene target 1-MYOD1 is shown with 2nd round PCR cycle number on the x axis and abundance of uniquely identified original template molecules in the final sequencing output when the 2nd round is terminated at the specified cycle number. Also plotted are 4 models of PCR amplification and their expected unique read sequencing output in normal conditions: R_Primer – exponential growth model; R_SP_Equal – equal concentration of the reverse UMI primer and SP2 primer in the 2nd round PCR; R_SP_Primers – actual concentration of the two primers assuming no activity of exonuclease I in degradation; SP_Primer – desired output with complete degradation of the reverse UMI primer, a horizontal line is drawn. 1-MYOD1 unique sequence abundance did rise with increasing cycle number however the final abundance was significantly different to the exponential growth model and non-significant against the other models. This suggests there may be some residual reverse primer in the 2nd round PCR although not at levels that significantly alter the unique read calls.*

5.6.4 Methylation gradient

As described bisulfite conversion of DNA can introduce biases to PCR. In particular with respect to templates methylation status where there can be preferential amplification. Both a bias to hypermethylated templates and hypomethylated templates has been reported[324, 354]. As a control experiment to test ASM-Seq's tendency or preference for one target or another known quantity of either 100% or 0% methylated control DNA (EpiTect™ Control DNA, Qiagen) were mixed to form a methylation gradient of DNA template molecules at 0%, 25%, 50%, 75% and 100% abundancy respectively. Both single-plex and multiplex pooled primers were used to test the primers under the different reaction well conditions (**Fig. 5.15**).

The goal Is for uniform amplification of the methylated DNA in line with the gradient such that a linear function where $x = y$ and R^2 value of 1. The deflection of the represented methylation by the laboratory technique can be model and given a value of *b*. The equation to calculate the *b* value is below:

$$y = (1 * b * x)/(b * x - x + 1)$$

A *b value* of 1 is optimal. Where *0 < b < 1* occurs, this represents bias towards and unmethylated template and where *b > 1* represents preference for methylated targets. If methylation bias is found to be present then this *b* value can be used in effort to correct the sequencing data[355].

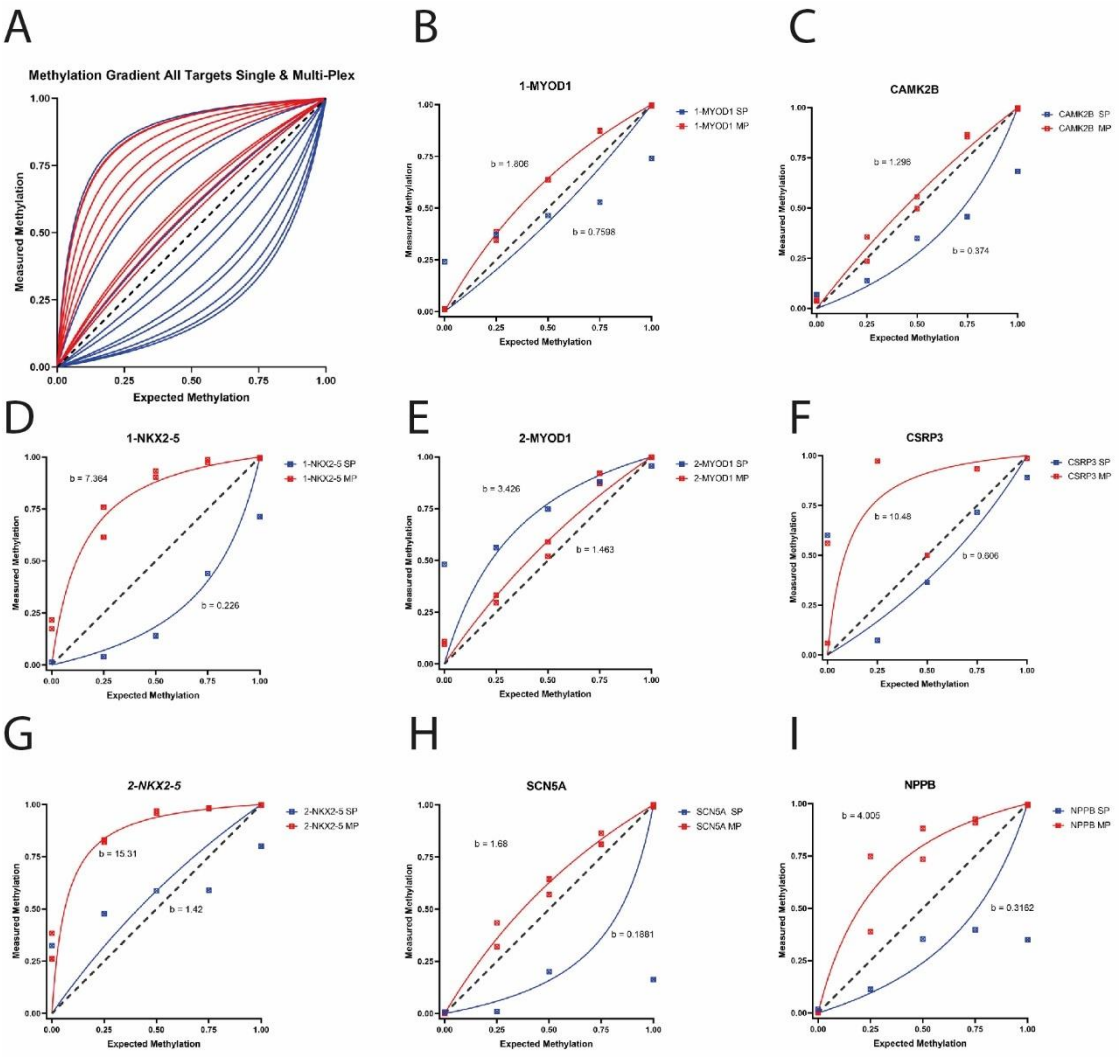

***Figure 5.15**: Methylation gradient plots. 9 panels are represented of both single-plex (blue curves) and multiplexed (red curves) methylation gradient reactions. The expected and measured ratio of 100% vs 0% methylated templates is on the x-axis and y-axis respectively. Panel A demonstrates an overlay of all gene targets primer sets in singleplex and multiplex reactions. Panel B to I give individual targets as titled and their respective singleplex and multiplex curves, note the differences (discussed in main text). Curves are drawn and analysed by non-linear regression, a value b is then calculated which represents the deflection away from the linear function. A positive deflection >1 represents preferential methylated DNA amplification with 0 < b 1 values representing unmethylated DNA preference.*

5.6.5 Testing ASM-Seq on cell lines.

In collaboration with Dr Freddie Whiting (BCI, Centre for Cancer Genomics and Computational Biology), a cell line experiment was devised to generate data regarding the rate of methylation infidelity during mitosis, currently estimated to be 10^{-4}-10^{-5} per CpG per cell division, but possibly even as high as 5%[255]. The cell lines HCT116 (MSI [microsatellite instability] colorectal cancer) and SW620 (MSS [microsatellite stable] colorectal cancer) were used for this experiment. All replicates were initially diluted to a concentration of 1/10 cells per well, thus on average 1 in 10 wells should have a single ancestral cell from which the cells were cultured. In the 12 hours following seeding a visual inspection ensured that chosen wells only contained one cell. These were then cultured to confluence within a 96 well plate. 4 colonies from each cell line were sampled separately and subsequently divided 50:50. One half was counted on a haemocytometer, pelleted and stored at -80^{o}C to act as a single time-point, the other half was re-seeded on progressively larger vessels and re-grown to confluence again with a repeated 50:50 split between each successive plate. In total six individual time points exist, three replicates of HCT116 (one replicate was lost during processing) and four replicates of SW620.

DNA was extracted, quantified and enzymatically converted to eDNA as detailed in **sections 4.2.6, 4.3.3 and 4.3.2**. Each sample underwent ASM-Seq as described using the primer panel MP #5. Samples were quality control checked on TapeStation (**Fig. 5.16**), subsequently pooled in equimolar concentrations and submitted for sequencing on Illumina® MiSeq v3 Platform at the Genome Centre.

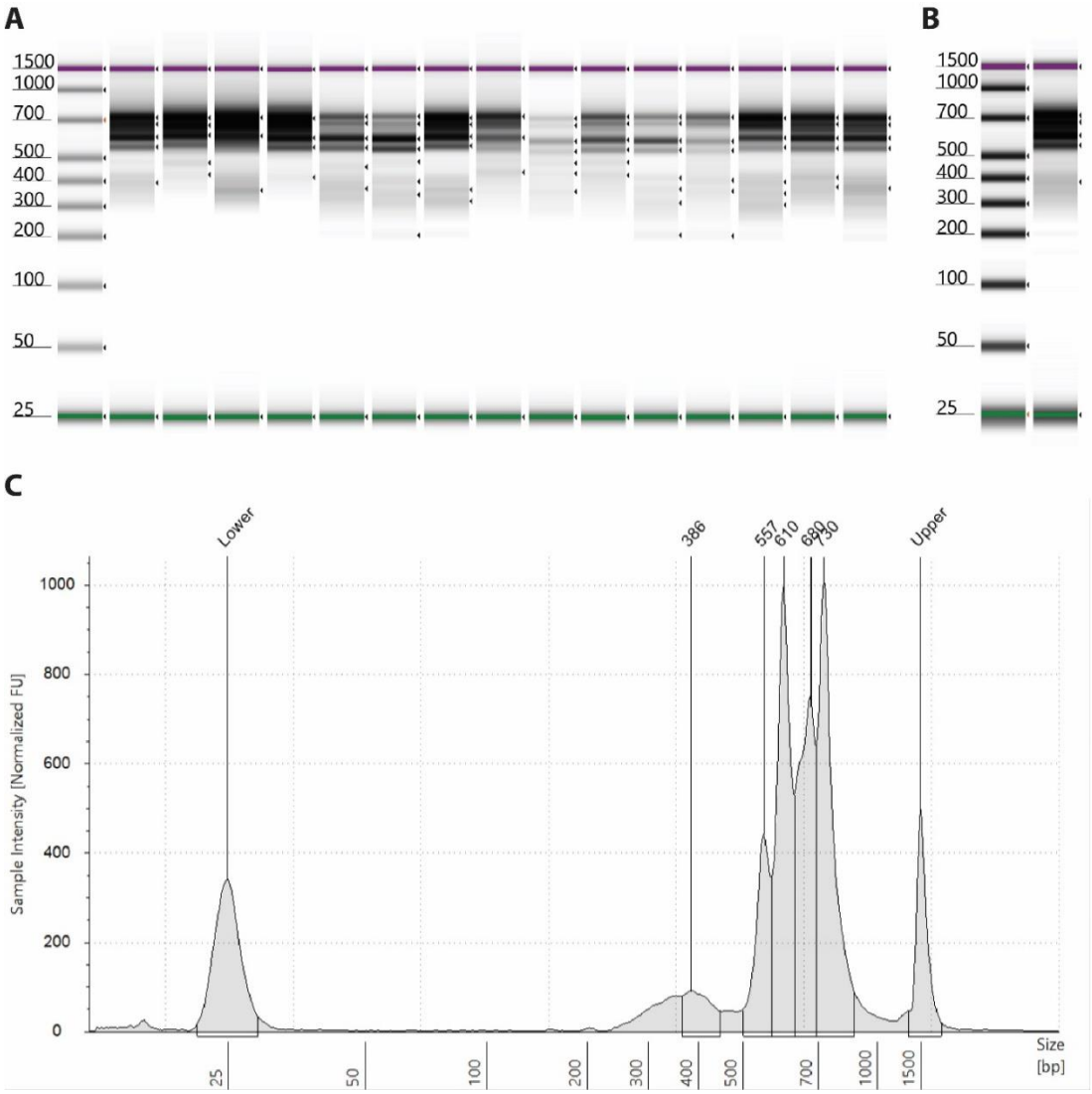

Figure 5.16: *Cell Line pooled libraries that have been sent for sequencing. Enzymatic conversion of DNA was employed. These gels/electropherogram demonstrate a fully functioning ASM-Seq technique for multiplexed, targeted methylation sequencing that incorporates UMIs. (A) This gel shows an example of successful library preparation of 15 sequencing ready individual cell line samples prior to pooling. (B) Gel of all pooled libraries - this includes 5 growth timepoints run in duplicate of 2 distinct colorectal cancer cell lines also run in duplicate (40 samples total). (C) Electropherogram of the final pooled library demonstrating the distribution of product. For this experiment, the multiplex primer pool #5 of 15 targets has been used with a library-ready size between 526-732 bp.*

5.6.6 Analysis of cell lines

The cell lines grew up over 5 further collection time points until the end of the experiment. The cell count numbers and days between re-seeding are recorded. With this information, computational modelling was kindly conducted by Dr Freddie Whiting to ascertain the birth and death rate of each line. Subsequently the number of cell divisions between each timepoint can be calculated with the following formula:

$$(b - d) = \frac{\ln\left(\frac{Nt}{N0}\right)}{dt}$$

Where *b* is the birth rate, *d* the death rate, *Nt* is the population size at time *t*, *N0* is population at time 0 and *dt* is the change of time in days. The ratio of *b:d* is then calculated to resolve the equation at each time point. With the net birth and death rate identified a ratio can be applied for the total mitotic time between the 5 timepoints. These are plotted in **figure 5.17**.

There is an increase in mean pairwise distance as the cell line proliferates across the samples as expected under cell division. However there is an abrupt fall in diversity mid-way through the experiment. It is not clear exactly why this is the case. Whether this represents a particular more rapid clonal expansion of one of the ancestral clones (from the single cell) could be a possibility as this would cause a relative homogenisation of the methylation patterns detected. The death rate may also have increased unwittingly resulting in a loss of diverse clones, I note that the drop occurs at transition to the 4th flask and whether there was a laboratory workflow or technical problem at this time is not known to me. Nevertheless the purpose of using the cell lines was to test the ASM-Seq protocol in another media and use the cell line data to optimise the bioinformatic pipeline which has been achieved.

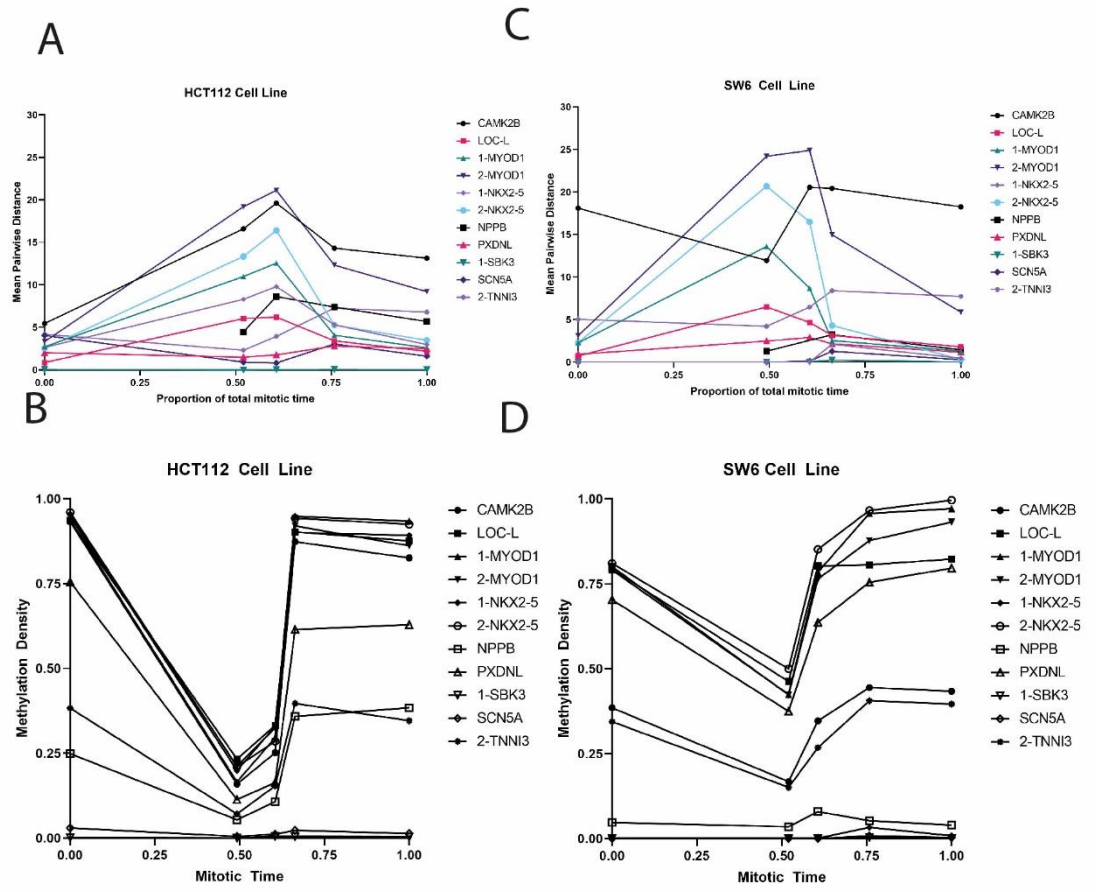

Figure 5.17*: Cell Line plots using the ASM-Seq technique. Left sided panels are HCT112 cell line data and right sided panels are SW6. A and C represent the change in mean pairwise distance over time of the experiment, normalised to a fraction of 1., there is a gradual increase of diversity over time as expected although an abrupt change is encountered at time point 4 (discussed in main text). Panels B and D show methylation density over the course of the experiment, at baseline it is high which is not uncommon in cell lines given their mitotic histories. There is a similar abrupt change in methylation density in the middle of the experiment that is not fully accounted for and subsequently recovers on following timepoints. Nevertheless, these plots represent a pipeline and analytic workflow that is now in place for the ASM-Seq protocol.*

5.7 Discussion

UMIs offer the opportunity to establish the true allele specific methylation patterns for more robust clonal dynamics and mitotic clock analyses. This protocol, the first to incorporate UMIs into a bisulfite template, has potential to be used in multiple tissue types and starting template to infer those dynamics, not just in epithelial tissues such as Barrett's. However, owing to the difficult starting template, the optimisation of the protocol has proven exceptionally fastidious. It is clear that primer design is the most important step to reduce the chance of primer dimerisation and off-target product as much as feasibly possible. Even small amounts of erroneous binding early in the PCR cycles results in a significant shift away from the target specific product and failed PCR. The elongation of 1st round primers >70 bases has also posed a challenge presumably through self-binding and concatemer formation. This is compounded by a UMI of degenerate bases which is free to exacerbate this problem. Designing the shortest length of primer possible without rendering the inherent idea of the protocol impossible or impractical is the key. A low concentration starting template also shifts the reaction towards primer dimerisation and mismatch, this has been overcome by reducing primer concentrations and improving AMPure Bead clean-up steps between 2nd and 3rd round PCR to better eliminate these erroneous products. Subsequently, sensitivity is acceptable down to 5 ng template and suitable for microdissected samples. If high concentrations of DNA are used the protocol becomes even more reliable in generating a library-ready product demonstrating its potential utility in other experiments where DNA template is in abundance e.g. bulk tissue sampling, cell line experiments.

The methylation gradient experiments are satisfactory. Of course some targets do demonstrate high *b* values though and these may need to be revisited in the fullness of time if repeated gradients are consistent with these findings.

The whole design and optimisation of this protocol has proved challenging due to the nature of it needing to be sensitive enough to detect and UMI tag small quantities of generally poor quality gDNA. However through perseverance and repeated reagent and mutable variable assessment and troubleshooting, a widely translatable technique has transpired. Finally, optimisation has included a wide range of permutations to reaction constituents and thermocycling conditions (**Supplemental Table S5**). Nevertheless the above protocol is suitable for onward analysis in the experiment proper to examine the clonal dynamics and mitotic age of Barrett's oesophagus through high resolution, high throughput allele specific methylation sequencing.

6 Results

Utilisation of the allele specific methylation sequencing protocol in predicting risk of progression to cancer, the origins and clonal dynamics of Barrett's oesophagus

6.1 Introduction

It is not fully understood how and when Barrett's oesophagus arises. The prior chapter focused on design, set-up and optimisation of the novel Allele Specific Methylation sequencing protocol that hopes to reveal the life history of BO over both time and space and define rates of mitotic ageing in cohorts of non-progressors and progressors.

6.2 Revisiting the origins and clonal mosaic of Barrett's

Barrett's is a clonally derived lesion[356] and its true ancestral origins remain controversial and much debated. It is not possible to observe its inception *in vivo* and the majority of patients are unaware that they have the condition.[55] Recent compelling evidence suggests that gastric cells from the cardia are the founders for BO[85, 125] and it is suggested the gastric cardia glands are the pioneer that create a reparative epithelium in the context of GORD induced ulceration[24]. However, rather than a *bottom-up* origin, theories remain regarding *top-down* beginnings from transdifferentiation of squamous epithelium progenitors[162] or potentially a multifocal *outside-in* approach from submucosal oesophageal glands[153]. These are certainly the most plausible and mainstream explanations regarding the histogenesis of BO.

In addition, BO exists in a state of dynamic equilibrium[184] whereby there is a natural waxing and waning of clonal patches over time and space but generally, unless at the precipice of progression[185], the lesion appears to be relatively indolent with little to

no clonal superiority. This benign façade however belies significant genotypic aberrancy[168, 180] that is ostensibly held in check via means that are poorly understood but likely relate to local clonal competition[357]. Intrinsic factors such as *TP53* or *SMAD4* mutation acting as a trigger for more widespread genomic catastrophe are important but do not explain the full story or how one reaches this point and the relative importance of extrinsic factors at play such as clonal interaction or expansion through glandular fission that could drive or suppress such a transition. In 2006, Maley et al.[238] demonstrated that the degree of genomic diversity was correlated with risk of progression to OAC suggesting for the first time that the specific underlying mutation itself was less important than the heterogenous milieu that it was surrounded by. This lends credence to the idea that there is more to uncover regarding the interplay and house-rules across the clonal mosaic of Barrett's.

With this in mind, the study of epigenetic drift has proven a useful tool in understanding proliferation and stem cell dynamics in other tissues such as the colon[254] and hair follicles[358]. Colonic crypts, like Barrett's glands, are maintained by a population of stem cells that define the clonal unit. As the colonic epithelium ages, through replication and stochastic hypo- and hyper- errors in DNA methylation status the stem cell population presents with increasing epigenetic mosaicism both on an intracrypt and intercrypt. The diversity of intracrypt methylation tags and thus presence of multiple unique sequences reflects the number of underlying stem cells present in the stem cell niche. In addition, by comparing intracrypt and intercrypt values both between near and far entities inferences can be made regarding the clonal expansion timing and history of that epithelium and how related crypts are. Additionally, similar comparisons over time informs how the epithelium changes and evolves or alternative remains in a state of indolence with a paucity of proliferation and crypt or glandular (in the case of Barrett's) expansion through fission.

6.3 Ageing the Barrett's lesion

When discussing whether *age* of Barrett's is a risk factor, to be clear, we speak of *mitotic* age rather than *chronological* age in the context of this thesis. Any inferences made regarding Barrett's dwell time or risk of progression is relative to the number of cell divisions that have occurred rather than the true passing of time. As an example, if we take two patients who are both 60 years old and have clinically identical BO and at index endoscopy (chronological time zero) a tissue biopsy gives a mean pairwise distance of evolutionary neutral gene targets of say 5 (arbitrary number) in both of them then we can say that their BO is mitotically the same age at that snapshot in chronological time, i.e. roughly the same number of cell divisions have occurred resulting in the same degree of epigenetic mosaicism. What this does not tell us is the current rate of cell division. One patient may have accrued this degree of mosaicism over 20 years, the other over 10 years. The question is whether the latter patient is at greater risk of imminent progression with a more rapid rate of accrual driven by excessive cellular turnover. It is important to restate that these mitotic clock CpGs should have no biological bearing on the differentiation, function, fitness or turnover of the cells but are merely passengers observing and documenting the rate of such actions. For this detail of mitotic rate between our two hypothetically identical patients, an interval of chronological time is necessary to repeat the mean pairwise distance and plot the slopes between the two points for an individualised assessment of their trajectory. If there is a significant difference in the gradient between our two slopes then we can determine that despite being clinically and phenotypically the same, they have different speeds of their mitotic clock. Whether a slow (shallow gradient slope) or fast (steeper slope) mitotic clock correlates with enhanced risk can only be determined with subsequent prospective follow-up against the clinical outcomes of our patients. It is also important to note here that the defined slope (mitotic rate) is not a static variable remaining on the same trajectory for ever more but actually, and more likely, represents a dynamic system. Illustratively, it could be that our patient with the greater mitotic rate has actually just undergone a short-lived burst of *punctuated evolution* and clonal expansion with resultant rapid cellular turnover, whereas the other patient remains in an indolent

state of *gradualism* with their individual mitotic clock slowing ticking over[356]. However, at the next endoscopic review, the roles may have reversed such that the two mean pairwise distances are equal again and with that the slope between the data points are compatible with a now synchronised mitotic clock. The benefit to researchers with clinical Barrett's surveillance practices is that it affords the archetypal model *in vivo* to study such matters as patients attend for interval endoscopy every 2-5 years (in non-dysplastic BO). This can be done both prospectively and retrospectively.

Clearly the idea is more useful, robust and reliable to calibrate mitotic clocks when multiple timepoints of good quality tissue samples are available however that does not mean that single timepoint biopsies render no useful information. Going back to our two hypothetical patients, once sufficient data is available and validated to define the *normal* expected mean pairwise difference for a given chronological age is when a snapshot determination of risk could be provided. Take colorectal cancer for example, Woo et al.[287] demonstrated that colorectal cancer appears to evolve from a baseline of mitotically older cells with the theory being that stem cells (from which cancer arises) that are mitotically older have reached a period of exhaustion, impaired tissue function and ability to maintain their status in competition with other stem cells such that the tissue micro-environmental dynamic has changed and focal proliferation of adjacent stem cell(s) with a relative fitness advantage occurs, of which this could be a neoplastic phenotype. This proliferating stem cell(s) has the same mitotic age by virtue of existing side-by-side in the clonally derived niche however their counterpart is likely heading towards cellular senescence while they continue to thrive released from local competitive pressures. What makes this stem cell thrive and the other reach growth arrest and exhaustion is debatable, however putatively in the context of tumourigenesis the surviving stem cell may harbour all the necessary genomic aberrations that are now selected for and useful to it with the change in micro-environment. This theory has been proposed as a mechanism for why cancer rates increase with chronological age. From an epigenetic methylation tag perspective, this competitive release would be measurable as a reduction in mean pairwise distance as the single (or smaller group) or stem cell(s) comes to

dominate the epithelial landscape through mitosis and clonal expansion. Indeed, when sampling early colorectal cancer methylation tags across the tumour are relatively uniform consistent with a so called "flat" clonal expansion with subsequent stalling or reversion to a the gradualistic model of drift once the new tissue microenvironment dynamic and boundaries have been defined between newly competing populations of cells. This new dynamic may or may not include neoplasia, or on the converse with an exceptionally advantaged stem cell the new dynamic could be invasive and metastatic cancer, as has been seen in some unfortunate cases of small early colorectal cancers[286]. This theory of cancer evolution aligns with the *born-to-be-bad* hypothesis and may explain why some patients with BO rapidly progress to OAC within the first year of index endoscopy.[58-60]

This prior work sets the precedent and example for examining BO to the same degree of detail in efforts to understand how the lesions does or does not evolve and change over time and space. The novel ASM-Seq protocol will allow a greater depth of single DNA strand and thus reads of methylation tags along with a wider genomic coverage to reach a resolution of the technique not achieved before. In addition to reducing confounding factors of PCR amplification bias and sequencing error via incorporation of the UMIs.

6.4 Aims

a) Establish a fresh frozen tissue bank of multi-level and multi-timepoint biopsies from a diverse population of patients with BO. Assign patients into a progressors or non-progressors cohort dependent on their clinical histories and presentations.

b) Establish a mitotic clock model using the stochastic epigenetic drift of methylated CpGs to estimate the individual dwell time of BO and whether progressors have an "older" lesion predisposing them to an elevated cancer risk.

c) Examine the clonal and cellular origins of the BO lesion by *timing* to glandular architecture across the segment through intraglandular and interglandular crosswise comparison.

6.5 Hypotheses

a) That a mitotic model of tissue aging in Barrett's demonstrates that patients who progress to cancer have an "older" and more diverse methylation pattern that predicts progression to cancer.

b) The origin of Barrett's arises upwards from the GOJ and will be characterised by mitotically older populations of cells/glands in the more distal aspects of the lesion.

6.6 The Barrett's cohorts

6.6.1 The patient cohort from the Royal London Hospital

302 patient encounters between February 2015 – May 2021 have provided a total of 1132 fresh frozen tissue samples. These samples come from 150 individual patients. The sex distribution is 112 male to 38 female, a ~3:1 ratio commensurate with the male predominance and prevalence of BO. The age at enrolment range is from 27 to 89 with a mean of 62.2 and median of 63 (data missing from two patients). 69 patients have more than one timepoint with a range of 1 to 42 months between enrolment biopsy and most recent biopsy. Total patient-years of current follow-up since individual enrolment is 623 patient-years. The range is 30 to 91 months with a mean of 50 months and median of 55 months. 15 patients have been referred to other hospitals for treatment of dysplasia or cancer or ongoing surveillance local to them, 9 patients have been discharged (6 for not having Barrett's; 1 at patient's request, 1 due to comorbidities and age; 1 due to failure to attend appointments) and 7 patients have died (1 from OAC, 6 from other causes).

102 patients are non-progressors with at least two timepoints (including pre-enrolment clinical timepoints) of histology showing gastric or intestinal metaplasia only. 22 patients have a maximal worst histology history of indefinite for dysplasia (IFD = 16) or low grade dysplasia (LGD = 6). 12 patients have a history of progression to high grade dysplasia (HGD = 3) or oesophageal adenocarcinoma (OAC = 9). 36 patients are currently undefined with only one timepoint both pre and post enrolment into this study, however all have either gastric or intestinal metaplasia and non-progression to date.

At the time point of progression (HGD or OAC) these patients are usually referred to the tertiary referral centre University College Hospital (UCH), therefore they are subsequently lost to ongoing fresh frozen prospective tissue collection. Furthermore, the long term clinical outcome of these patients e.g. curative therapy, future surveillance and/or death is not available.

6.6.2 The patient cohort of archival fresh frozen specimens from University College Hospital

All samples are currently stored in liquid nitrogen. 1438 tissue specimens have been obtained taken between September 2001 – April 2009. All samples are from patients who have progressed to either HGD or OAC. 358 patients were enrolled over this time with 131 patients having 2 or more timepoints (range 2 to 11 timepoints, mean 3.8, median 3). The biopsies comprise a mixture of squamous epithelium, distal and proximal BO, GOJ, gastric cardia, direct sampling of areas of HGD and OAC, and endoscopic mucosal resections (EMR).

6.6.3 Patient selection for processing

The aim here was to achieve a relatively balanced cohort of non-progressors versus progressors for comparison with use of ASM-Seq. Furthermore, choosing cases to ensure there were additional surveillance timepoints, this was clearly more difficult with the progressors who mostly did not have any prior surveillance tissue available under our recruitment. Some of these cases were identified either as *de novo* cancers in a background of Barrett's and were re-attending for repeat biopsy at which point they were consented for additional biopsies of the adjacent non-dysplastic region and dysplasia if available at that time, there would be no intention of the patient coming back to RLH for surveillance on clinical grounds as these cases are usually managed elsewhere (UCH), thus serial biopsies over time for the progressor cohort is more limited than non-progressors. True surveillance progressors, that is a documented history of non-dysplastic BO in the past, total 3 out of the 12 patients in the RLH cohort with HGD or OAC.

6.6.4 Progressor cohort

The 3 true progressor cases were chosen alongside a further 6 RLH progressor cases and 5 progressor cases from the UCH cohort of samples where it could be confidently

ascertained that no endoscopic therapy had been given prior to the research biopsies, as this could confound any subsequent findings. A total 14 progressor cases, age range of 57 to 86, mean age 70; median 68). 2 out of 12 had HGD with the remaining 10 progressing to OAC. One case was short segment BO, the others were long (≥3cm). 3 had tissue available from at least 2 timepoints (1 HGD, 2 OAC), the rest were single timepoint biopsy samples. Biopsies were a variable mixture of non-dysplastic BO both adjacent to the area of dysplasia or OAC and at distant sites to it elsewhere in the BO segment. Of the 14 progressor cases, a total of 12 had glandular sections cut on the LCM for onward processing, optimisation and either acceptance or rejection for sequencing (see **section 6.7**). The 2 cases not processed were from the UCH cohort and unfortunately the tissue biopsies were far too degraded to visualise any meaningful glandular architecture following sectioning.

6.6.5 Non progressor cohort

18 cases were chosen from the database, 14 of these had at least two timepoints of biopsies available for processing (range 1-6 timepoints). Age range was 47 to 73 with mean and median of 61. 4 cases had LGD at their most recent endoscopy, there was occasional IFD reported in the longer histories but never any higher grades. 1 case had not had surveillance before and thus was an index case of LGD. The other 3 had documented non-dysplastic BO at least >1 year prior to diagnosis of LGD. Of these 18 cases, 11 (8 non-dysplastic BO, 3 LGD) proceeded through the entire protocol. The drop out of the other cases was mostly due to poor quality biopsies without clear glandular morphology, which is not uncommon in a fresh frozen template, thus the cases were abandoned. 3 cases had suitable tissue sectioning but did not reach the point of processing prior to focusing on bioinformatics and data analysis.

Unfortunately, across the progressor and non-progressor cohorts there are only two female cases, one in each cohort due to unexpected drop out of other selected cases.

6.7 Workflow

6.7.1 Histopathology and immunohistochemistry

Biopsies samples from both cohorts across multiple timepoints were collected and submitted to BCI Pathology Services for sectioning as per the protocol described in **section 4.2.1**.

On receipt back of the sections on PALM membrane slides, the sections underwent CCO immunohistochemistry as described in **section 4.2.2** to provide contrast for tissue architecture identification and opportunities to cut differentially expressing CCO glands for clonal analysis with Sanger sequencing of mtDNA when available on the section.

6.7.2 Laser capture microdissection (LCM)

H&E reference slides were reviewed prior to commencing LCM (**Figure 6.1**). This was to allow prioritisation of tissue microdissection to sections and biopsy specimens with excellent post-freezing morphology and to ensure that the majority of biopsies within a single timepoint conformed to a reasonable standard such that the end resulting cohort of Barrett's glands was balanced in temporo-spatial axes. For example, at some timepoints for patients, only one biopsy out of the 3-4 that were available was suitable for LCM. As an example, case number 172 (non-progressor) was a particular disappointment with all 16 biopsies spanning 3 timepoints being unsuitable for LCM and thus the case was dropped.

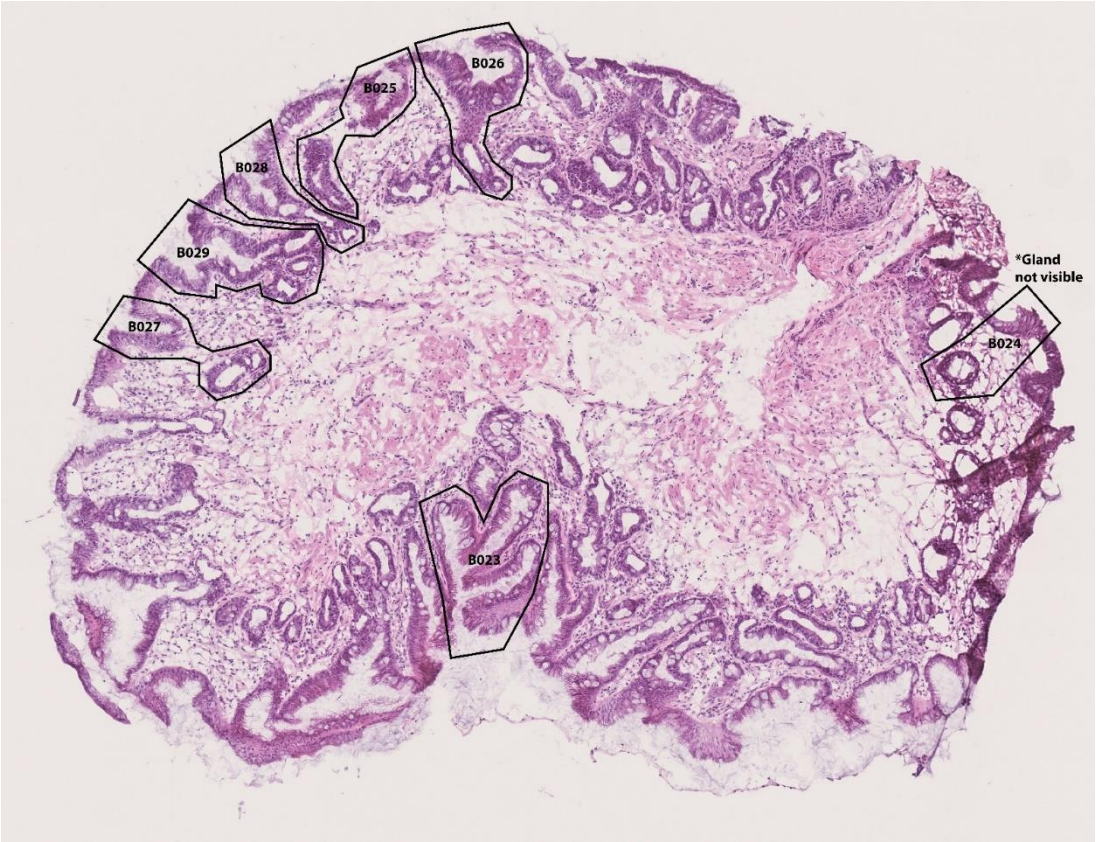

Figure 6.1: *Haematoxylin & eosin (H&E) stained reference slide of the glandular epithelium of Barrett's oesophagus (BO). This is an annotated H&E section of biopsy 128-JE23-B2. This particular biopsy is taken proximally within the patient's BO segment and represents the first of three timepoints collected within the fresh frozen tissue bank. This particular patient had progressed to high grade dysplasia by timepoint three. Seven individual BO glands have been marked up prior to cutting the 12 serial sections on Laser Capture Microdissection (LCM) slides. Samples are labelled B023-B029 respectively. The H&E's serve as valuable references for case and sample tracking. They provide a detailed historical record of the architecture, spatial mapping and permit an assessment of basic glandular phenotype prior to formal immunohistochemistry where necessary. Taken together, this is useful when conducting intragland versus intrabiopsy versus intergland analysis and clonal ordering experiments.*

Nevertheless, out of a total 247 biopsies across 65 total timepoints in 33 patients (progressors and non-progressors), 112 biopsies across 48 timepoints had sections cut on the LCM to continue down the workflow.

The goals of LCM were to achieve a wide breadth of glandular phenotypes with aiming for 4-6 glands cut per biopsy expecting that 1-2 may fail the pre-sequencing quality control as per **section 5.4.13**. Therefore, across a timepoint of 2-4 biopsies this would result in 8-24 total biopsies respectively. A precedent with a similar breadth of coverage was presented from our lab in the colorectal adenoma lineage tracing study by Humphries et al.[276] in 2013. Assuming all planned glands are sequenced the lowest target end would provide 2^8 (256) glandular combinations for intergland analysis. In this cohort of patients, at the lower end of target biopsies an output of >384 glands is to be expected to achieve suitable cohort coverage. Furthermore, a simple power calculation using a lifetime risk of ~9% for dysplasia or OAC estimates that ~80 samples are necessary per each progressor and non-progressor cohort to generate suitable power for comparison.

Following identification of suitable phenotypes on the reference H&Es cutting proceeded as described in **section 4.2.5** on the LCM. **Figure 6.2** demonstrates an example of the serial capture of an individual gland across multiple BO sections. The individual glands then have their DNA extracted (**section 4.2.6**), quantified (**section 4.3.3**), underwent enzymatic conversion of 5mC (**section 4.3.2**) followed by ASM-Seq as per chapter 5 followed by analysis on the 4200 TapeStation system prior to suitable selection for Illumina sequencing.

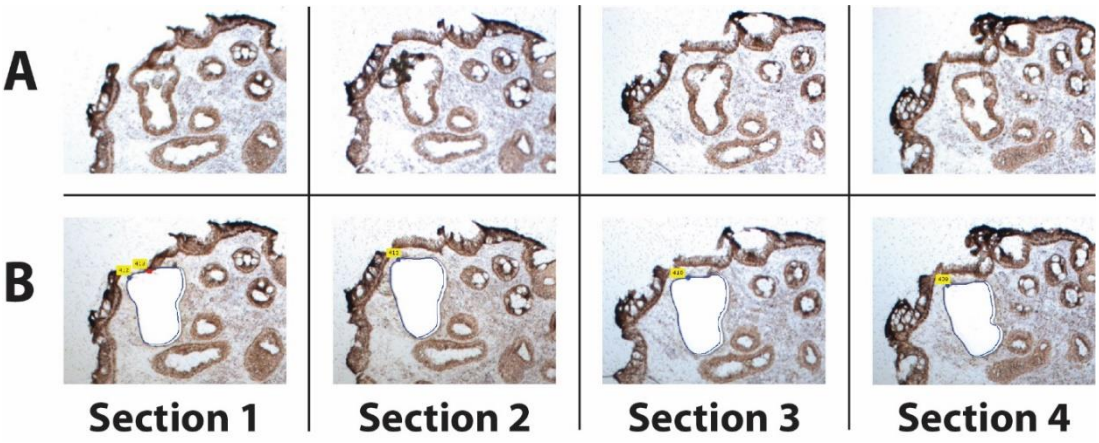

***Figure 6.2**: Gives an example of the LCM technique to isolate single glands. Four sections individual sections are shown. Row A shows gland before LCM cut is taken. Row B shows the tracing of the cut for each section and the disappearance of the targeted gland which has been catapulted into the collection tube. All sections are serial.*

6.7.3 Barrett's gland dataset

In total 480 individual Barrett's glands were micro-dissected and underwent ASM-Seq to the TapeStation checkpoint. Of these, 384 passed the visual quality control check with respect to there being a suitable PCR amplification product identified between the expected bounds of the multiplexed amplicon pool. Average yield of DNA was 6.35 ng / gland prior to enzymatic conversion equating to roughly 1000 copies of the genome. The sensitivity of ASM-Seq was such that only half of each sample was necessary with the rest stored at -20 ^{o}C or used in the event of a technical failure. **Figure 6.3** is a representative TapeStation gel demonstrating the efficacy of the technique with low quantities of input eDNA from glands.

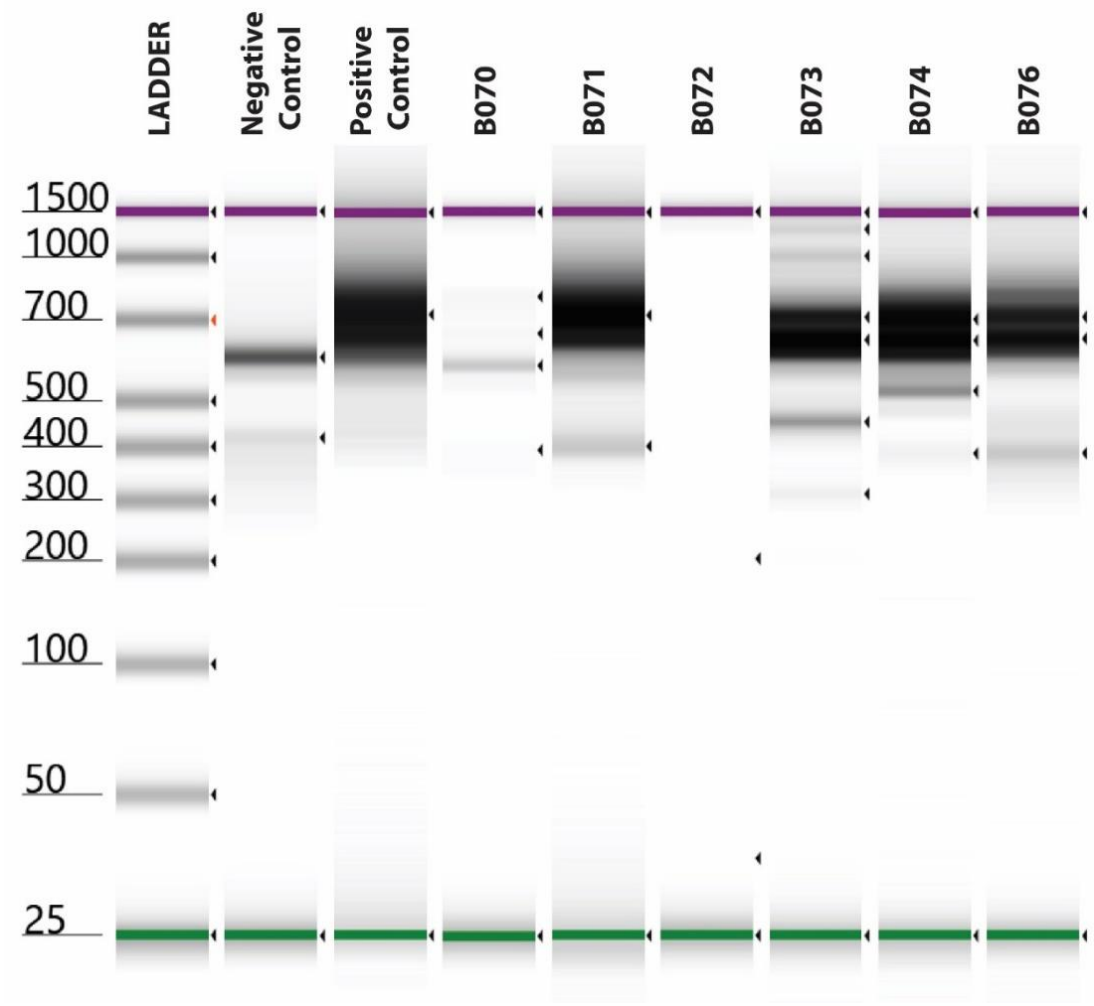

Figure 6.3: *Example of successful amplification of individual Barrett's Oesophagus (BO) glands using ASM-Seq. This gel demonstrates the sequencing ready libraries for six BO glands [B070; B071; B072; B073; B074; B076] that await pooling with the wider experiment and submission for MiSeq 300bp paired end sequencing. A negative and positive control have been included in this example. Unfortunately, glands B070 and B072 have failed to amplify. A technical error is likely to have occurred either in the laser capture, DNA extraction, enzymatic conversion or the PCR protocol itself. The individual total amount of extracted DNA of these six laser captured samples prior to enzymatic conversion ranges between <2.4ng – 5.6ng demonstrating sufficient sensitivity of the technique with small inputs of eDNA template.*

6.7.4 Read assignment

The cohorts of samples were all sequenced on the Illumina® MiSeq v3 300 bp PE platform. The following calculation was used to assign a suitable number of reads per sample per gene target per DNA template aiming for at least 20 reads per target amplicon to ensure good depth would be obtained. This is also in the knowledge of a challenging substrate to sequence (AT-rich) with amplicons at the edge of the sequencing range in terms of length. Furthermore, a spike in of 20% PhiX, as recommended (details available at **emea.support.illumina.com**) and as is necessary given our low complexity sequences, would still result in over-jealous preferential sequencing of its genome reducing the power to our Barrett's samples. All sample quantification was converted to copy number using the Thermofisher DNA copy number and dilution calculator available online (**www.thermofisher.com/**)

Thus:

MiSeq v3 Chemistry	25,000,000 reads / 15,000,000,000 bases
Minus PhiX 20% spike-in	20,000,000 reads
Read assignment per UMI	>20 reads for every original strand DNA
Input DNA (allele copies)	Average ng / sample = 500 copies
Number of target amplicons (TA)	15 target amplicons

Therefore acceptable maximal sample number per sequencing run:

$$= \frac{\text{Maximal Total Reads}}{\text{UMI read assignment} \times \text{Allele copy \# per sample} \times \text{\# TA}}$$

$$= \frac{20{,}000{,}000}{20 \times 500 \times 15}$$

= **up to ~133 Samples per sequencing run**

6.8 Results

Four MiSeq sequencing runs covered all Barrett's samples including additional peripherals of positive and negative controls, a repeat methylation gradient (plots for each gene target are provided in the **supplementary figure S1**) and a few cardia and squamous samples that did not form part of this analysis.

All data files were downloaded and bioinformatically processed as described and subsequently analysed with GraphPad Prism 9 (**www.graphpad.com**).

6.8.1 Exclusion of failed gene targets from datasets

It was quickly apparent that unfortunately not all gene targets had been successful in target enriching and generating a product. As such much of their data was missing and so any data where they were successful was removed wholly. 4 genes were affected, these were *ANKRD2, CSRP3, SBK2 and 1-TNNI3.* This represents a loss of 104 CpGs and reduction of total panel coverage to 479 CpGS. From herein these targets are no longer used.

6.8.2 CpG correction

Because of the variable length of the multiplex primer panel and variable coverage of CpGs this resulted in variable mean pairwise distances in the dataset. For example, *CAMK2B* contains 78 CpG loci compared to *PXDNL* which only has 16. Thus there is for these outliers to skew the data and as such a CpG correction factor was calculated computationally based on the average number of CpG loci that formulated the mean pairwise intraglandular dataset. This dataset was chosen because normalisation of CpG number had already had to happen at this point in the bioinformatic processing to permit appropriate alignment of CpGs across samples for comparison. It was important to computationally derive the correction factor drawing on the original CpG numbers that fostered the data in the first place. This is because, for example,

CAMK2B usually spans 78 CpGs but during intergland analysis just an average of 61.25 CpG loci were used per combination. Thus, using a correction factor akin to the expected CpG coverage of 78 would only serve to swing the pendulum too far the other way in terms of attempting to normalise data across the panel.

Each gene target's correction factor is listed in **table 6.1**. All analyses using the mean pairwise distance of intraglands or interglands was subsequently corrected by dividing this value into the value of the mean PWD. The resultant effect was a *CpG-normalised* dataset that had also become much tighter in absolute terms with decimalisation.

GENE TARGET	CpG LOCI	CORRECTION FACTOR
CAMK2B	78	61.25
LOC-L	28	18.89
1-MYOD1	40	36.36
2-MYOD1	59	57.34
1-NKX2-5	42	26.28
2-NKX2-5	47	44.59
NPPB	45	41.85
PXDNL	16	6.877
1-SBK3	27	23.86
SCN5A	72	52.04
2-TNNI3	25	22.64

Table 6.1: *CpG Correction factor table. Gene targets successful from the multiplex pool and their respective usual CpG number across length of amplicon is shown alongside a computational defined CpG correction factor.*

6.8.3 Methylation density plots

Methylation density plots were created. There was however no observed correlation between age of the patient and the methylation density of their samples. Simple linear regression gave close to zero slopes, all p values were non-significant for a correlation with age (**Fig 6.4a**). On the contrary though, gene targets did demonstrate a correlation. When plotted against each other, there was strong positive correlation between methylation density of each gene. P values were <0.0001 between all gene targets. **Figure 6.4b** represents this correlation.

This finding potentially suggests that while epigenomic methylation density is not correlated across the patients in this cohort, the degree of methylation is correlated within patients at their own independent rate. Of note, patients at the different spectrum of ages (range 48 year old vs 84 years old) had comparable overall methylation density across the gene panel. This finding is in contrast to previous work[254, 282] in colon crypts where the authors reported a gradual trend towards hypermethylation. Although on closer review, their discrete pattern of methylation data does match those presented in this thesis albeit with a more convincing line of best fit. R^2 value has not been quoted though in terms of fitness.

The correlation of individual gene targets is primarily drawn statistically from lower ends of the methylation range where there is significant clustering of data <0.25. It is possible that this is skewing the data and regression line due to a paucity of similar data further down the x and y axes. On reviewing the R^2 values of all regression lines they are poorly correlative with the linear function. A plot of the mean and standard deviations of all R^2 values is given (**Fig. 6.4c**). Previous work by the same groups above did not demonstrate a correlation with their gene targets (MYOD1 and NKX2-5) but had far less of this bottom left clustering effect. On balance the data distribution appears similar to previous published work, however this needs further examination ahead to ensure that these previously proven neutral targets are truly neutral in their epigenetic drift.

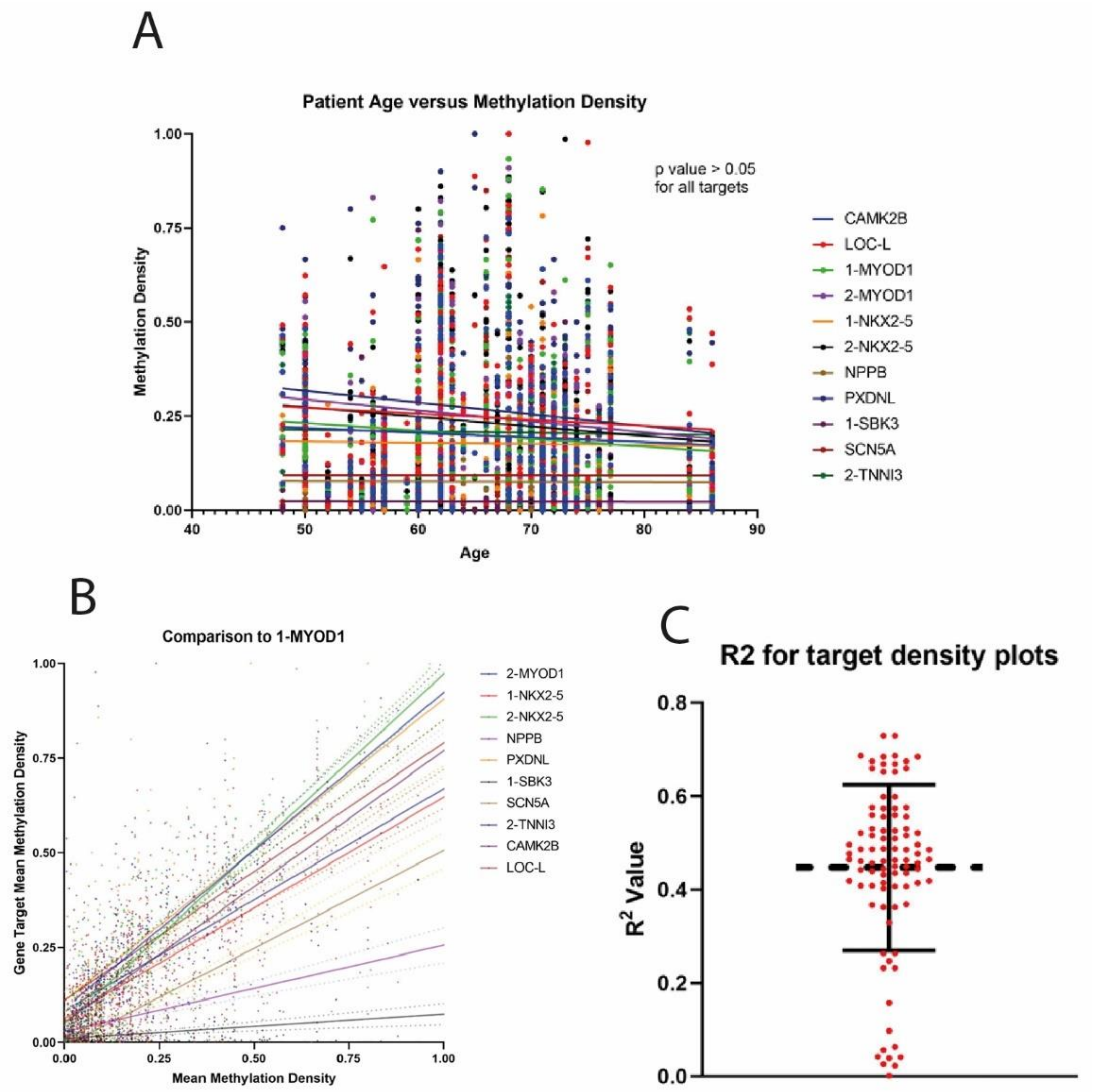

***Figure 6.4**: Methylation density analysis of the entire Barrett's patient cohort. Panel A is methylation density over time (age), there is no correlation observed with all age groups presenting with the same degree of methylation density across the gene panel. Panel B demonstrates a linear correlation between 1-MYOD1 and all the other gene targets. R^2 is 0.5755 which is typical of the cohort of targets in pairwise comparison plots. Note the excessive clustering in the bottom left-hand corner which may be adversely skewing the data given the general paucity of data points >50% methylation. Panel C shows mean and standard deviation of R^2 values for all gene target methylation density comparisons.*

6.8.4 Variant methylation density over age separates the two cohorts

The two cohorts of non-progressors and progressors were split into their constituent parts in the dataset. Methylation density over age was then replotted to see if there

was any difference in the two groups that was being averaged out and hidden from view in the plots in **figure 6.4a**. **Figure 6.5** demonstrates this analysis. This shows that over time methylation density in patients who progress to HGD or OAC appears to decrease. Simple linear regression analysis demonstrates a significant p value of 0.0341 compared to the horizontal. Epigenetic drift can occur bidirectionally over time that is to both hypo and hyper methylated state though the stochastic errors of DNMT. This balance may account for the appearances of a horizontal plot in **figure 6.4a**. The finding of progressive hypomethylation in the progressors is in line with a recent paper from Jammula et al. where they defined a distinct hypomethylator subtype of OAC characterised by significant levels of genomic instability.

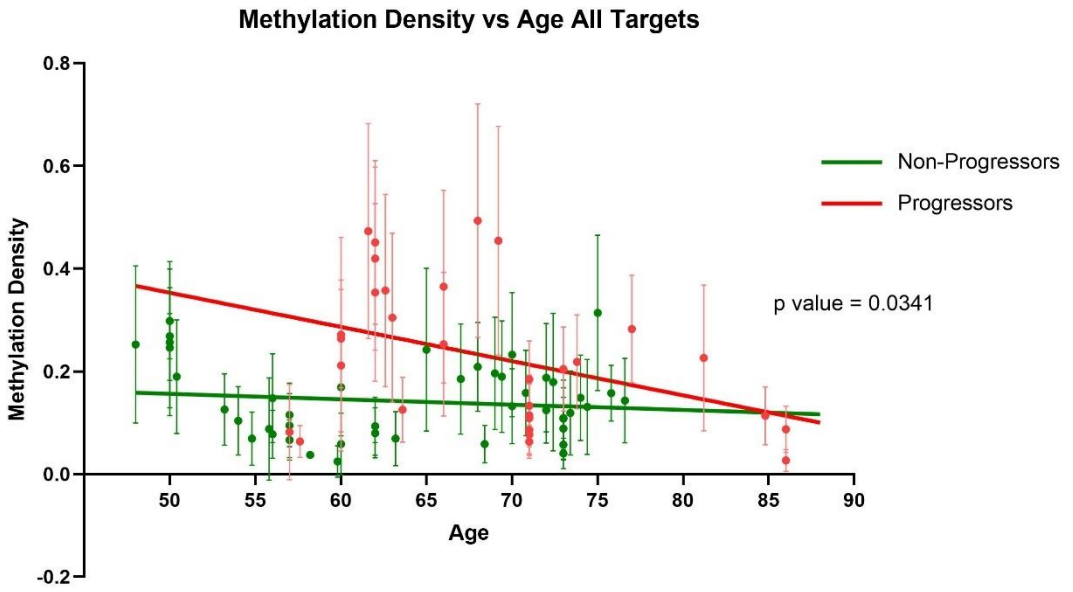

***Figure 6.5**: Progression is characterised by hypomethylation. Age is plotted against methylation density of the ASM-Seq gene targets. The progressors and non-progressors have been separated out and demonstrate differential methylation density time passes. This suggests that hypomethylation may be a marker of clinical progression.*

6.8.5 Read depth

The average number of total reads per Barrett's gland was 36,949 reads. On average this produced 240 UMI consensus families per sample with the most consensus UMI

reads being 1883 in B068. Note that the mean copy number per sample was in the region of 500 copies of the genome, thus this number of average UMI sequences would be comparable with other previously published yields for barcode sequencing in gDNA. The remaining 11 gene targets favoured variably in their sequencing credentials with NPPB the worst performing target left in the pool with just unfiltered 14,000 reads collapsing to 979 UMI consensus reads across the entire cohort of 386 samples. A graphical representation of total reads, UMI consensus reads and unique methylation tag reads is given in **figure 6.6**. To correct this uneven representation of gene targets in the future further protocol optimisation is necessary with particular focus on multiplex primer pooling and altering the ratios in line with this data. It is also important to note that 1-MYOD1 is the shortest amplicon and thus will be favoured the most in PCR and NGS. On reflection, a graded primer concentration pool may be a possible mechanism to correct for some of these factors.

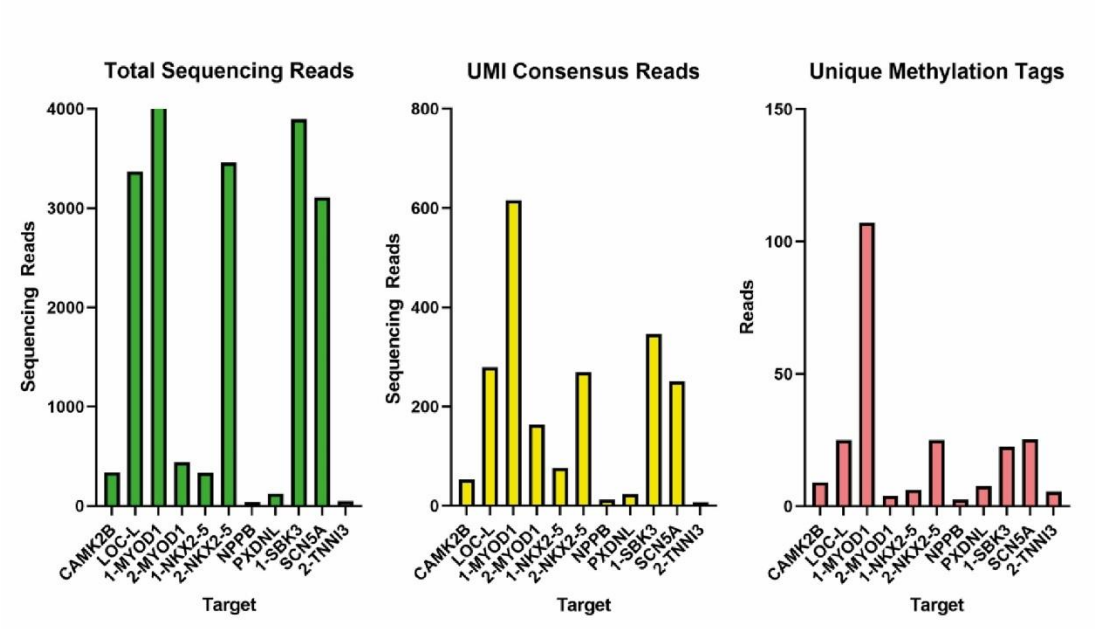

***Figure 6.6**. Differential read counts delivered across the primer pool by ASM-Seq. On the left green panel are mean total unfiltered sequencing reads per Barrett's gland. Please note that for scaling purposes 1-MYOD1 has been curtailed. The value here is 21,796 reads compared to 1-SBK3 which is next best at 3893 reads. The middle yellow panel is the number of consensus UMI reads brought together as UMI families during pre-analytical processing. The Red panel shows mean unique methylation tags that are collapsed down from the UMI consensus reads where two tags share the exact same binary series of 1's and 0's.*

6.8.6 Epigenetic diversity over progression

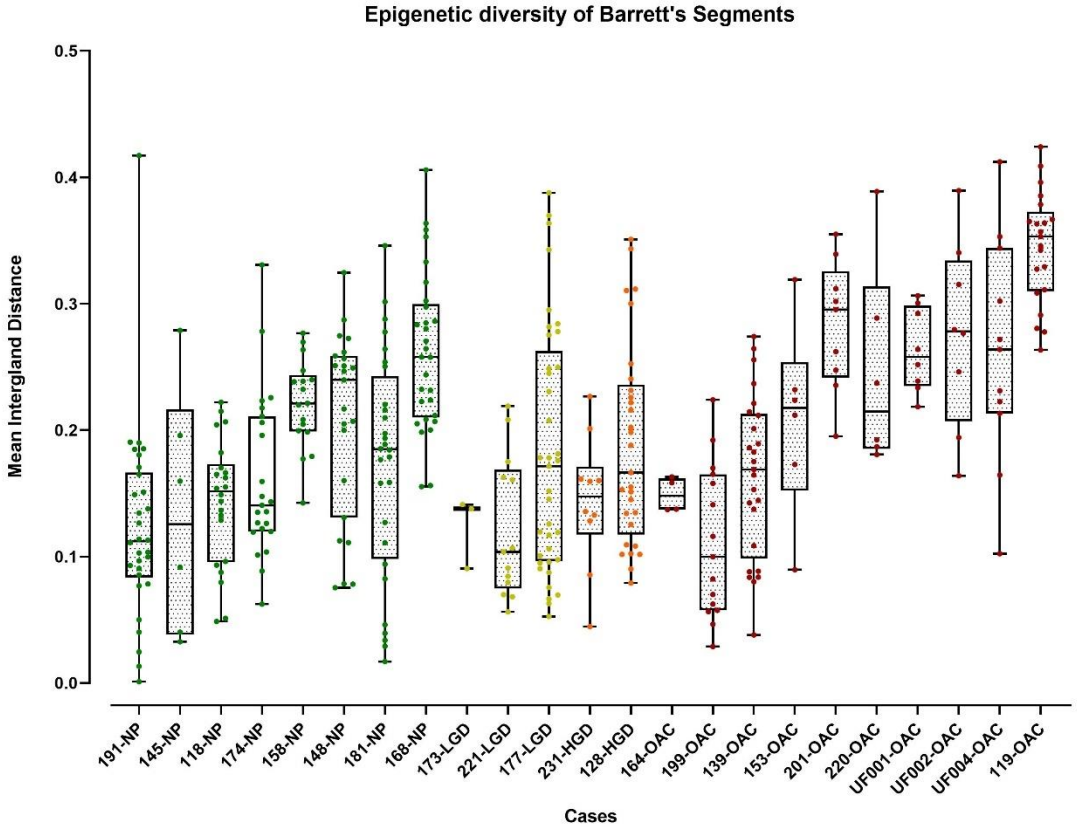

***Figure 6.7**: Cases summary plotted against mean intragland pairwise distance (PWD). This demonstrates the 8 non-progressor cases organised on the left of the chart and the OAC cases on the right with yellow/orange LGD/HGD cases respectively in the middle. The cases have been purposely ordered in this fashion to demonstrate a signal of increasing epigenetic diversity with each stage of progression that will be explored further in the next sections through intragland and intergland dynamics.*

6.8.7 Intragland analysis

Intra-gland analysis forms the basis of the mitotic clock idea as discussed previously. The older the population of cells the more mitosis that has occurred and the greater chance of generating epigenetic diversity that breeds greater mean pairwise distance (PWD) that can be measured. Whether there is differential mitotic clock rate or mean pairwise value between different cohorts of Barrett's patients is not known.

The Barrett's glands data was subjected to a number of different permutations to represent different aspects of the ageing model in Barrett's. In overview, the following sequence of plots demonstrate that mean pairwise distance can be used as a measure of risk to progression.

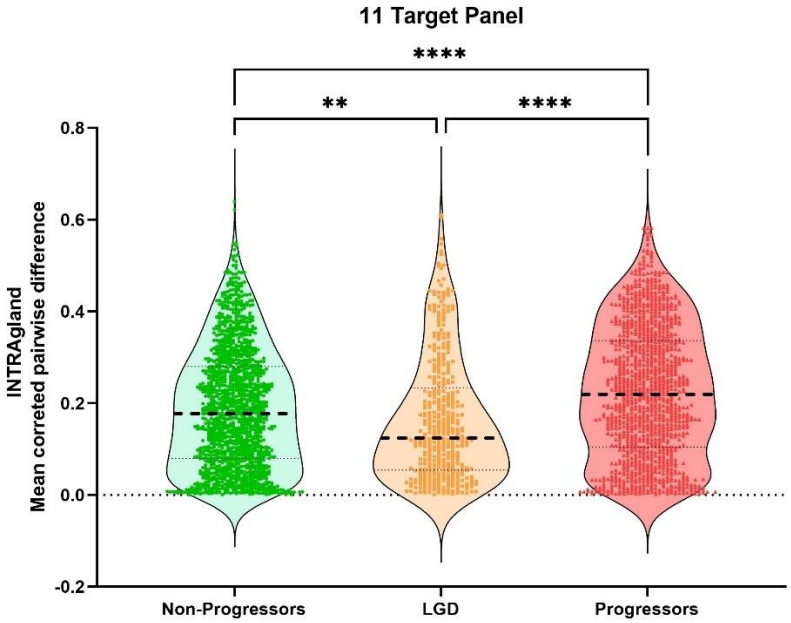

Figure 6.8: *Progressors have an older Barrett's lesion. Here, the entire cohort is separated into non-progressors, low grade dysplasia and progressors. This figure demonstrates that progressors have a higher intragland mean PWD suggesting that on average across the cohort they have a mitotically older Barrett's segment. Interestingly the LGD patients have significantly lower intragland distance, this may represent a recent clonal expansion in the context of their dysplasia driving this. **=P<0.01, ****P<0.0001*

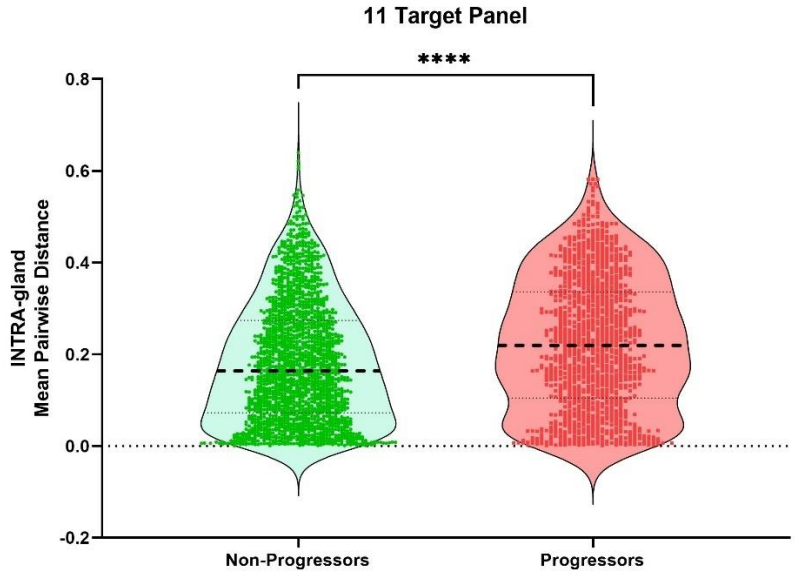

***Figure 6.9**: Intragland distance is greater in progressors. Here the cohort is separated into just progressors and non-progressors. The low grade dysplasia patients (n=3) are also classed as non-progressors. The statistical significance remains. P<0.001*

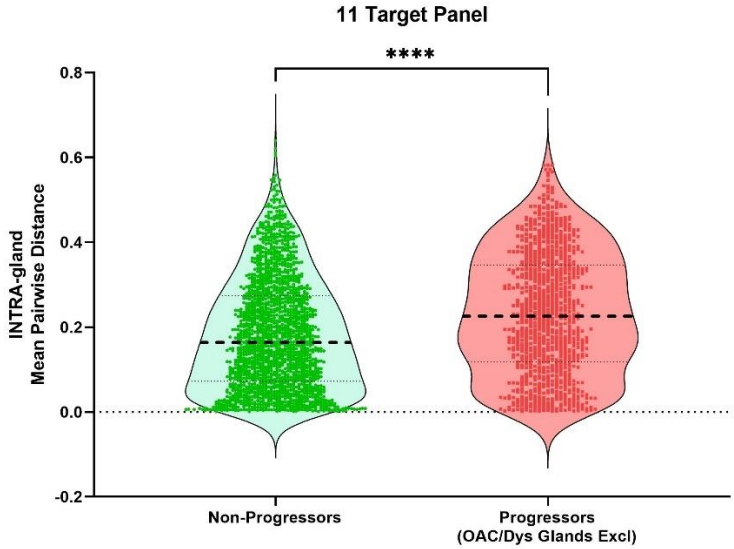

***Figure 6.10**: Non-progressors versus Progressors with all dysplastic samples removed from the from cohort. To ensure the significance of differential pairwise difference was not being caused by dysplastic or outright cancerous glands these were filtered out of the dataset. This figure demonstrates that the mean intragland PWD maintains its statistical significance. In particular these samples are only non-dysplastic BO glands. Some of these glands are within the same timepoint and biopsy series as the dysplasia or cancer samples but none are intra-biopsy samples. This suggests that high mean PWD detected in non-dysplastic BO remains discriminatory. P<0.001*

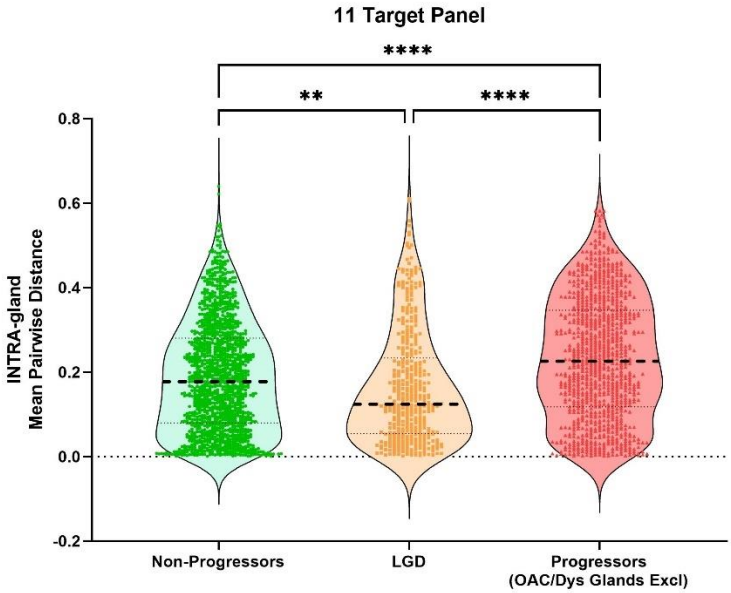

***Figure 6.11**: This plot demonstrates the statistical significance of **figure 6.10** is maintained when LGD patients are separated from the non-progressor cohort given they could be erroneously pulling down the mean PWD. . **=P<0.01, ****P<0.0001*

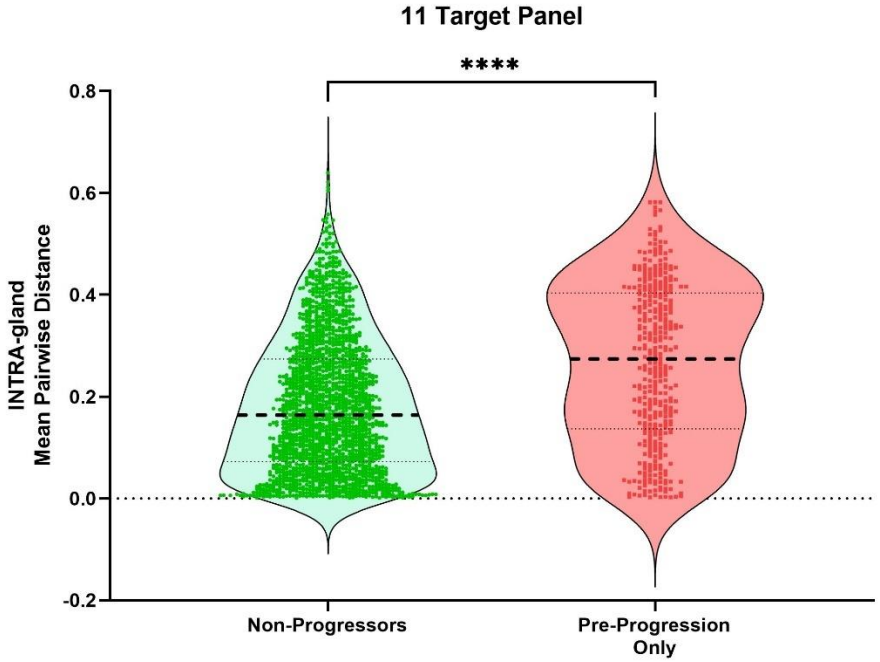

***Figure 6.12**: Pre-progression can be detected by mean intragland PWD analysis. A further filtering step of the data has occurred whereby all glands from the timepoints associated with OAC or HGD have been removed. Thus the pictorial significance here is that all samples are non-dysplastic BO taken at least >1 year prior to the presentation with progression., ****P<0.0001*

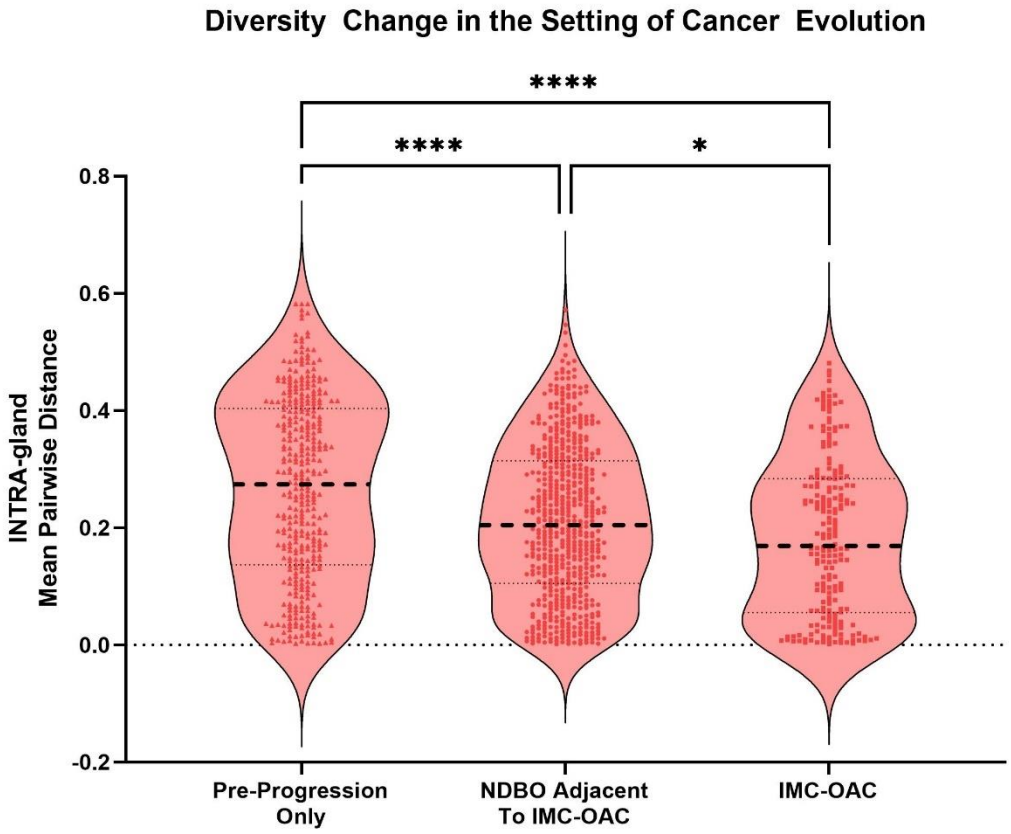

Figure 6.13: *Mean PWD falls at the onset of progression. Here is represented just glands taken from the progressor side of the cohort. They have been filtered into three further cohorts i) the pre-progression cohort from* ***figure 6.12****; ii) Histologically normal appearing BO taken within the same timepoint and thus same Barrett's lesion at the endoscopy of progression; iii) Just the glands that are histologically dysplastic or adenocarcinoma. This figure demonstrates a step wise reduction in mean PWD occurring within the 1-2 years before presentation with cancer. Such reduction in PWD would be explained by a general homogenisation of the methylation tags through clonal expansion of the dysplastic and then malignant phenotype. *=P<0.05, ****P<0.0001*

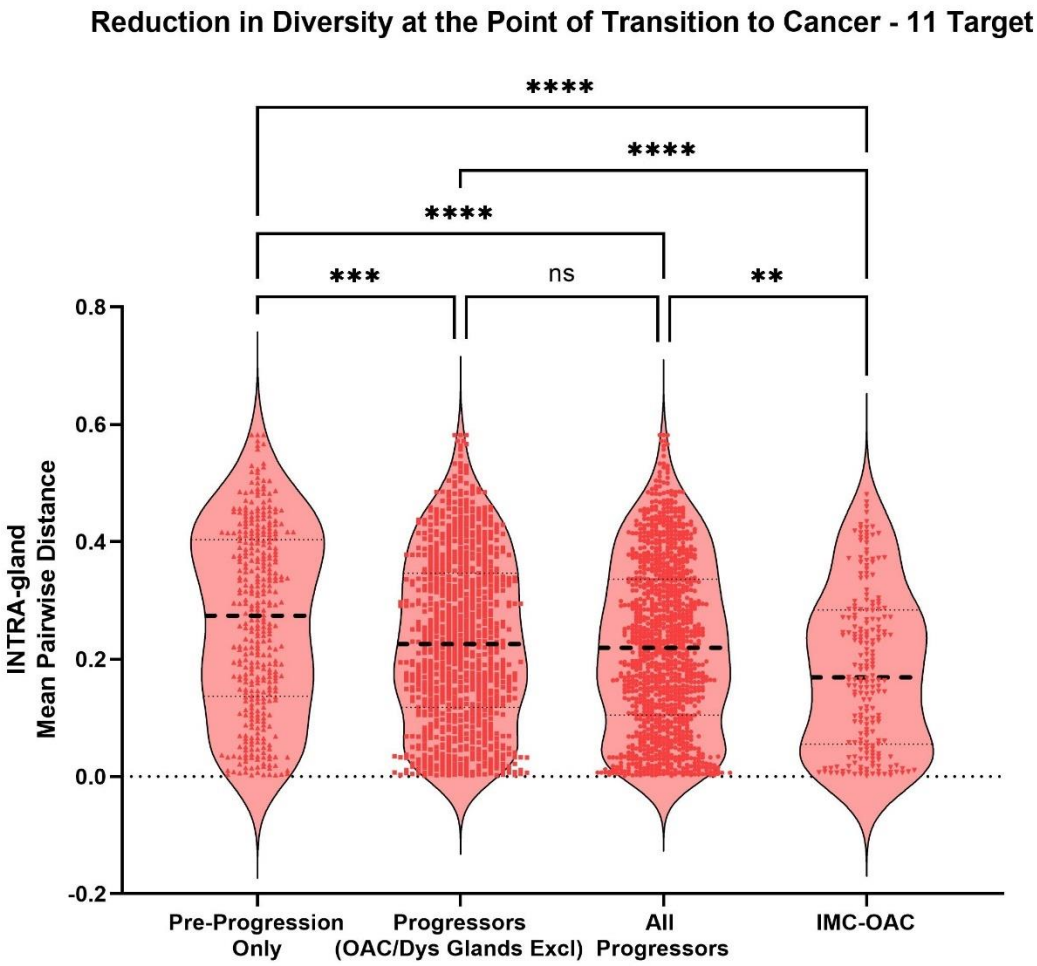

Figure 6.14*: Cancer is associated with a low of methylation pattern diversity. This plot further illustrates the points made in* ***figure 6.13****. The progression to cancer results in a clear reduction in epigenetic diversity regardless of how the cohort is filtered. In particular there is strong significance (****=p < 0.0001) of a drop in diversity presumably as IMC or OAC expands out of the general background aberrant Barrett's tissue represented here by the "OAC/Dys Glands Excl" cohort. This significance is also maintained when the same data is replotted adjacent to IMC-OAC giving a further signal to a demonstratable loss of epigenetic heterogeneity within the glands of the malignancy. **=P<0.01*

6.8.8 Intergland analysis

Intergland analysis is useful in determining intra-lesional clonal dynamics. That is, how diversity patterns alter over time echoes what is happening at a clonal level with the Barrett's. In particular intergland analysis measures the age differential between two spatially separated glands. They may be adjacent to each other or they could be at either ends of the oesophagus (SCJ vs GOJ), whatever, the mean pairwise difference can tell us something about their relationship to each other. If the patterns are similar in diversity terms then the two glands are recently clonally related. If there is significant pattern variability between the two glands then they represent more distant relatives. A high pairwise distance would mean more ticking of the mitotic clock has occurred since the two glands potentially clonally expanded together. Low diversity indicates a younger gland to the one that is being compared. Perhaps this gland arose as a later clonal expansion in the segment either under intrinsic genomic or extrinsic environmental factors such as the need to heal ulceration at the GOJ.

With respect to the wider spatial dynamic and environment across the segment, interglandular analysis informs can inform on how active the lesion is generally. Activity may include local clonal expansion or invasion, regression, cell death or loss of a clone and with it its ancestral history. In these cases of dynamism there will be greater pairwise distance seen multifocally across the segment where-ever such activity is taking place. If instead there is relative homogeny and lack of diversity then it suggests relative indolence.

The benefit of Barrett's surveillance is the capability to take multiple biopsies through the segment to try and discover the full cacophony of clonal interaction through this intra and interglandular analysis.

Herein are presented relevant plots that reveal and confirm prior literature that Barrett's is a relatively inactive and indolent lesion.

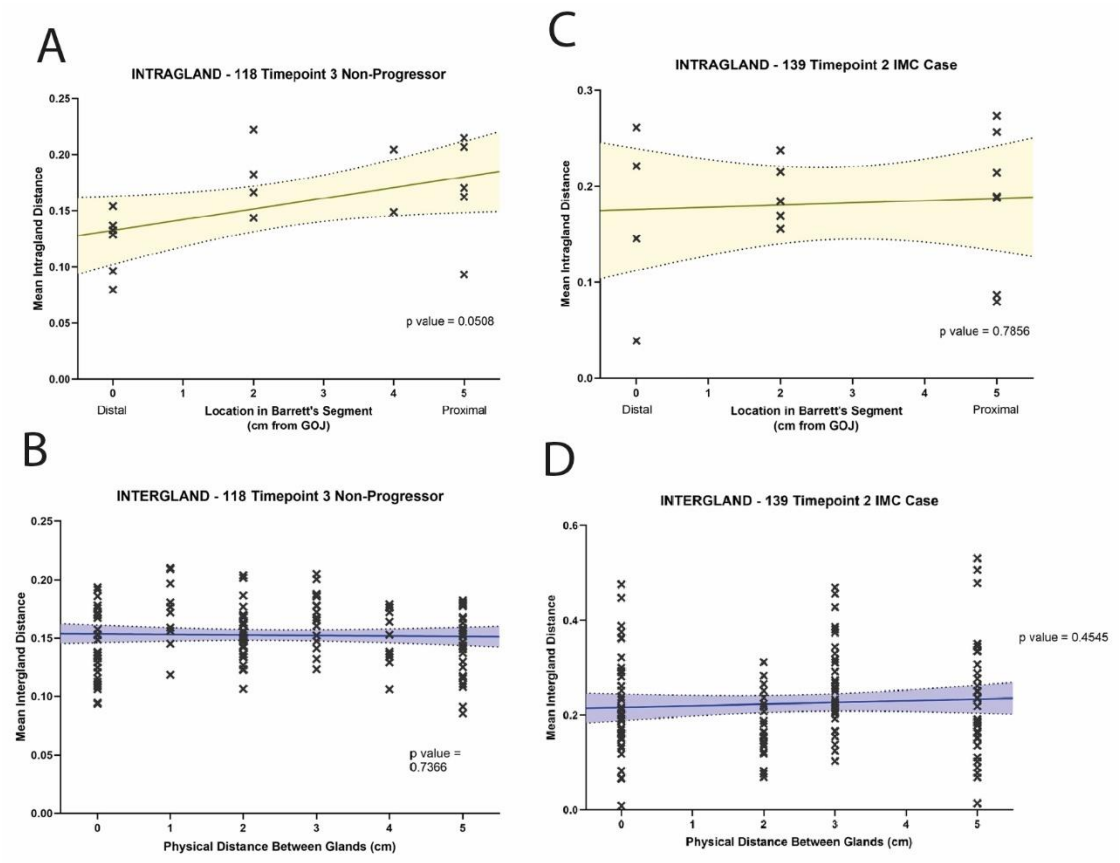

Figure 6.15*: Intragland distances along the Barrett's segment (A & C) and Intergland physical distances against pairwise distance (PWD) (B & D) for two patients, Case 118-non-progressor (A & B) and Case 139-progressor (C & D). Above are representative plots of a multiple analysis that was undertaken comparing temporo-spatial relationships against intragland and intergland pairwise distances separately across the entire cohort of progressors and non-progressors and their respective different timepoint endoscopies. All plots generated are similar to the above. In A & C the intragland PWD is plotted against the length of the patient's Barrett's segment on the X axis. Distal (near GOJ) biopsies are on the left, and proximal on the right. Simple linear regression analysis then plots the line of best fit, p values are given, none were significant for variant pairwise distance over the segment, although panel A was borderline at p=0.0508. In B and D for the same respective patients all combinations of all glands are compared for pairwise distance analysis and the mean between two individual samples is plotted. A known measurement between two glands forms the x-axis thus the plots examine the clonal relationship of physical distance (not physical location) within the segment and how divergent their intergland methylation patterns are. In both cases of intra and intergland analysis here through space there is no significant variation in the mean PWD. This indicates that the Barrett's is in a state of indolence without much clonal expansion, contraction or mitosis. Furthermore, the distant glands are just as related as the near glands suggesting that a burst clonal expansion happened filling the physical space with glands followed by indolence.*

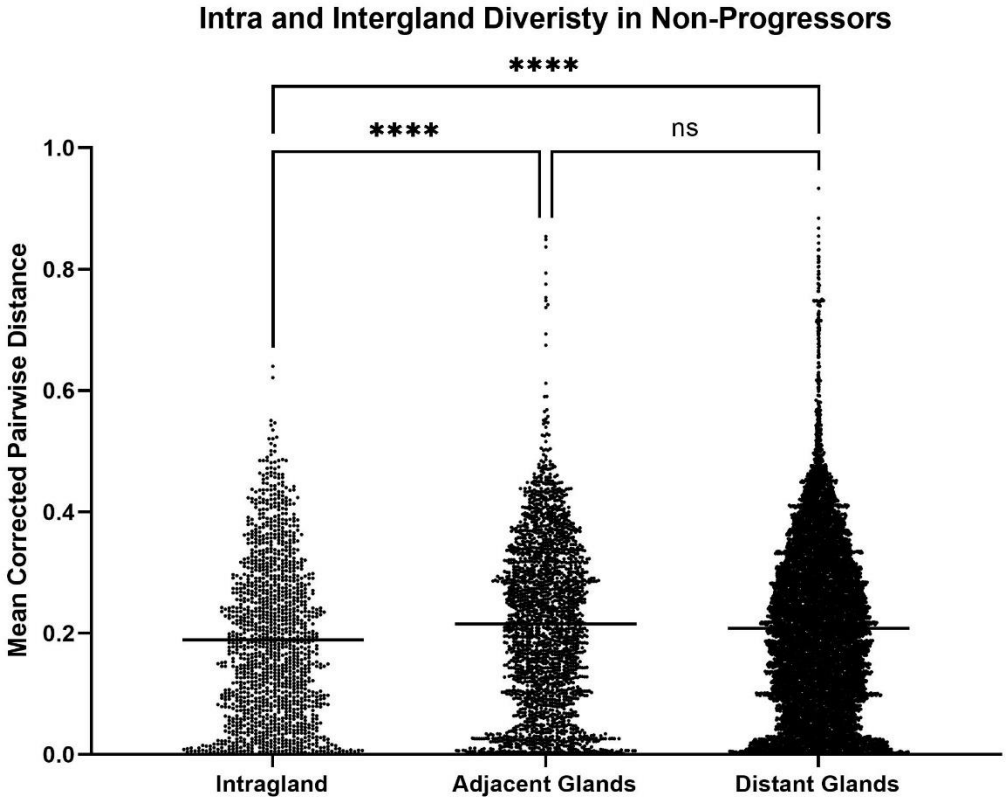

Figure 6.16*: Intra and intragland dynamics in non-progressors.*

Here, all the intragland mean PWD are plotted on the left for the non-progressor cohort, these values are then compared with interglandular mean PWD of glands that are physically close to each other, that is from the same biopsy, for example in ***figure 6.1*** *and glands that are further away, different biopsy and over time. This plot demonstrates that within glands that are maintained by stem cells there is little diversity. However glands at greater distances have a somewhat divergent pattern. This points toward an older population of glands across the segment without any significant clonal sweeps that would result in homogenisation and reduced diversity. There is low intragland PWD because the stem cells are relatively indolent without any mitotic drift, but there is high intergland diversity because over time of significant indolence in these non-progressor patients the drift has been significant enough to cause divergent stem cell patterns between glands. This represents a stable Barrett's lesion without clonally expansions with greater fitness such as dysplasia or malignancy., ****P<0.0001.*

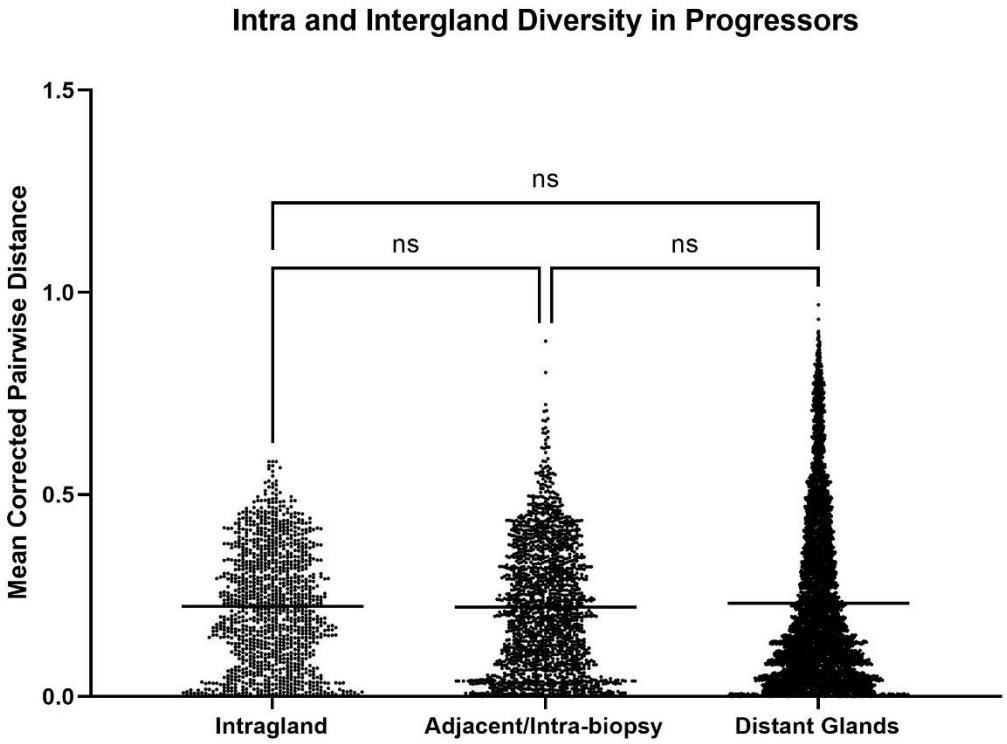

***Figure 6.17**: Intra and intergland diversity in progressors. Contrast figure 6.16 with this figure. This is the same preparation of the data only in progressors. There has been a relative reduction in the interglandular mean PWD compared to intragland mean PWD such that the grossly significant p value seen in non-progressors is now non-significant. This suggests that there is a clonally more active Barrett's lesion resulting in homogenisation of methylation patterns across the segment. The other alternative is the intraglandular stem cells may have become more active in their replication, mitotic indices or expansion dynamics such that the prior suggestion of reduction in intergland PWD is actually an increase in intragland activity. This plot points towards a more sinister environment within the Barrett's lesions proven by the virtue that these cases progressed to cancer.*

6.9 Discussion

These data have revealed that a mitotic clock model of Barrett's oesophagus is possible to achieve with the use of the ASM-Seq protocol.

First, when comparing patients who progress to cancer and those who do not, they possess a Barrett's lesion that is already mitotically older at least 1-2 years prior to the onset of their progression. Furthermore, when that progression occurs there is a rapid and stepwise reduction in methylation tag diversity and thus presumably tissue diversity by virtue of a monoclonal expansion of the malignant phenotype into the space that results in this measurable epigenetic homogenisation. It seems that the malignant transition is thus characterised by high diversity leading up to its precipice. We know from somatic DNA and RNA sequencing studies that diversity is a risk factor in progression[166, 238, 243] but these data now reveal that high epigenetic methylation tag diversity measured by mean pairwise distance is also a risk factor and points towards a mitotically active cellular population.

This is taken in the context of the intragland analysis work too which states that the Barrett's lesion is mostly indolent, especially in non-progressors. But that in the progressor cohort a shift in pairwise distance and reduction of diversity is seen to occur across the Barrett's segment evidenced by the endoscopic mapping biopsies taken throughout this work.

Lastly, there was the relatively unexpected finding of flat methylation density patterns across the entire cohort that was not correlated with age. As stated, both hypo and hyper methylation can occur as part of epigenetic drift that may account for some of this finding in addition, there was the confirmatory finding[215] of a significant fall of methylation density in the progressors as they age that can characterise some subtypes of OAC. This fall in the progressor cohort may account for some element of flattening in the dataset as a whole, although the non-progressor cohort did not meet significance in progressive hypermethylation over age.

These data have revealed key new insights into the clonal competition, ordering and expansion dynamics in Barrett's. While this data stops short of defining differential rates of the mitotic clock between progressors and non-progressors there is clear and highly significant evidence (p value = <0.0001) that elevated PWD and heterogenous methylation patterns which are purely driven by the act of mitosis can predict risk in progression to malignancy. To augment these findings further with histopathological data phenotypic analysis on our ASM-Seq data follows.

7 Results

The evolution and dynamic relationship of Barrett's gland phenotype

7.1 Introduction

Barrett's oesophagus is the only known precursor condition of oesophageal adenocarcinoma and as discussed previously is the metaplastic replacement of the normal oesophageal squamous epithelium with a columnar, glandular phenotype. Diagnostically, the presence of goblet cells as a marker of intestinal differentiation has been used to define Barrett's in several countries,[19] however in the UK Barrett's is defined as any columnar epithelium within the anatomical oesophagus[3]. When we consider the various theories on the origins of Barrett's, it is becoming clear that the role of repair of oesophageal ulceration due to chronic reflux from the normal gastric epithelium is the most likely cellular source of Barrett's [85, 101, 359]. The evolution of intestinal metaplasia in Barrett's has been previously linked to stratification of cancer risk in Barrett's but there is strong evidence that this is not always the case[63, 103, 104] (and researchers from my host laboratory have demonstrated that cancer can evolve from epithelial glands that do not show any intestinal differentiation[101]. Indeed this work revealed lineage tracing of *TP53* mutations across the non-goblet phenotype across the lesion and formed the clonal source of the subsequent cancer. It is therefore clear that understanding the role of Barrett's gland phenotype in the evolution of the disease is important if we are able to fully understand how Barrett's develops into cancer.

Barrett's itself displays a rich, diverse epithelial landscape with several gland types present. These range from glands that contain entirely gastric differentiated cells to those comprised of entirely intestinal differentiated cells and importantly, those that contain an admixture of both. We do not understand fully the distribution of these phenotypes in Barrett's however each has been relatively well defined. The corpus

gland type displays all the differentiated cell types of gastric corpus glands (including parietal cells and Chief cells), the oxyntocardiac gland only displays parietal cells but no Chief cells and the cardiac gland displays a simple foveolar pit with mucous secreting cells at its base. Additionally, glands can display both intestinal goblet cells (that express MUC2) and gastric foveolar cells (MUC5AC) and are therefore termed 'specialized'. These are typically diagnostic for Barrett's however some glands only contain intestinal lineages such as goblet cells and Paneth cells [107, 305]. The significance of each gland phenotype has been demonstrated in a paper published from my host laboratory with data produced from this thesis, that demonstrates that gland phenotypes reflect an evolutionary process where one phenotype can transition into another at due, presumably to environmental pressures. Furthermore, we were able to demonstrate that diversity of phenotypes within a Barrett's lesion is associated with dysplastic progression[305]

To date, the significance of Barrett's gland phenotype has not been fully appreciated and we do not completely understand the distribution of Barrett's gland phenotypes at the gastro-oesophageal junction (GOJ) and nor do we know the mechanism by which Barrett's glands transition from one phenotype into another. Furthermore, we do not understand the dynamics and mitotic age of each phenotype. I have demonstrated in the previous chapter that pre-progressive and progressive Barrett's demonstrate and increasing mitotic age that can perhaps be used to stratify patients into cancer risk but we do not know if this extends to one particular gland phenotype or is common to all, implying that fundamentally the environment dictates gland cellular turnover and expansion within the Barrett's lesion.

Here I address these unresolved issues by demonstrating the distribution of gland phenotypes in a cross-sectional Barrett's patient cohort both at a fixed point in space close to the GOJ and throughout the Barrett's lesion. I then demonstrate the clonal relationship between glands that display more than one gland differentiation pattern within the same gland and that it is possible to order the evolution of one phenotype into another. Additionally, I use the high-resolution, allele-specific methylation sequencing (ASM-Seq) array to reveal if each phenotype has a unique expansion and

mitotic age dynamic. This may help with our understanding of the role gland phenotype plays in progression in Barrett's.

7.2 Brief overview of methods (see chapter 4 for detailed methods)

7.2.1 Patients

Patients were recruited from the surveillance BO endoscopic clinic at Barts Health NHS Trust and from the archives of both the Royal London Hospital and University College London Hospital approved under multicenter ethical approval from London research ethics committee (11/LO/1613 and 15/LO/2127). Snap frozen biopsies and formalin-fixed paraffin-embedded (FFPE) specimens were used in this study.

A series of 64 biopsies from 51 BO patients were collected from 1.0-2.0cm proximal of the GOJ and were FFPE-preserved and an additional biopsy was flash frozen using cryospray. All biopsies met the following inclusion criteria: 1) Biopsies taken at the same anatomical height within the esophagus, regardless of BO maximum length; 2) Taken from the BO lesion identified during endoscopy, and 3) Absence of dysplasia or cancer at the time of endoscopy or any previous history of dysplasia. The mean age of the patients within cohort 1 was 62.2 (range 27-89) years, the female to male ratio was 1:4.9 and the mean maximum BE segment length was 4.5 cm (range 1.5-14 cm, median = 4.0 cm). For 25 of these patients, we obtained further archival FFPE H&E sections from all biopsies taken at the same surveillance endoscopy. Progressor biopsies were taken at the time of endoscopy in patients that had progressed or eventually progressed to dysplasia and were taken by Professor Laurence Lovat of University College Hospital (UCH), London, UK under his ethical approval.

7.2.2 Gland phenotyping

At least two experienced pathologists determined gland phenotype in both FFPE and frozen sections (Dr Marnix Jansen, UCH and Prof Sir Nicholas Wright, Barts) by identifying the individual differentiated cells known to be present in each phenotype (such as parietal cells, Chief cells, goblet cells, foveolar cells or Paneth cells). Each case was also subjected to immunohistochemistry to confirm each glands phenotype.

7.2.3 Gland immunohistochemistry

Glands were principally phenotyped on serial sections to an H&E stained slide. The method is described in **section 4.2.3**, but briefly one section was stained with an antibody to goblet cells (MUC2) and one with an antibody to foveolar cells (MUC5AC). Together this staining protocol is able to distinguish between cardiac type and specialized type glands.

7.2.3 Laser capture microdissection

For frozen tissue sections several serial sections were cut and each gland of interest was identified in the sections immediately preceding LCM slides either stained for MUC2/MUC5AC or by H&E. This permitted selection of phenotyped glands for LCM. Progressive samples were identified as per Chapter 6.

7.2.4 Mitochondrial DNA sequencing

A nested PCR protocol was used as previously published[131]. Briefly, the mitochondrial genome from each microdissected area was amplified into nine, 2 kb fragments, which were subsequently re-amplified into 500 bp fragments. Primer sequences and

PCR conditions were used as previously described (Greaves et al.). The second round PCR primers contained an M13 sequence to facilitate sanger sequencing. PCR products were ExoSaP-treated according to manufacturer's protocol (GE Healthcare, UK) and Sanger sequenced by Eurofins Genomics (Ebersberg, Germany). Obtained sequences were viewed using 4Peaks software (**https://nucleobite.com**) and compared to the revised Cambridge reference sequence using online tools provided at **www.mitomap.com**. Polymorphisms and non-epithelial mutations were eliminated from analysis by comparison with sequences from a microdissected area of stroma. Each mutation was confirmed using the same PCR sequencing protocol repeated from extra DNA extracted from the original LCM section.

7.2.5 ASM-Seq

The results chapters 5 and 6 show a detailed methodology of the ASM-Seq technique and it is not necessary to cover this again. DNA was extracted only from glands that could be effectively phenotyped through all serial sections using H&E or immunohistochemistry.

7.3 Results

7.3.1 The distribution of gland phenotype adjacent to the GOJ and throughout the Barrett's lesion

Here, I provided a detailed analysis of gland phenotype from H&E sections from the cohort of 51 Barrett's patients biopsies taken by myself at the routine endoscopy Barrett's surveillance lists over the course of my PhD. **Figure 7.1a** demonstrates representative examples of each gland phenotype observed and confirmed by a pathologist. We observed five gland phenotypes (atrophic corpus, oxyntocardiac, cardiac, specialized and mature intestinal) in the cohort but it is clear that the predominant phenotypes were cardiac type and specialized type (**Fig. 7.1b**). The diagnostic biopsies for 25 of these patients were available for phenotyping and **figure 7.1c** shows all 217 Seattle biopsies taken. These patients are identified as black dots on **figure 7.1c**. It is important to note that there is a similar distribution of gland phenotype adjacent to the GOJ as there is throughout the Barrett's lesion. An important observation shows that while cardiac and specialized glands predominate, we often observe more than one gland type within each biopsy. Most display one phenotype (n=30, 46.9%) or two (n=30, 46.9%) however 3 biopsies showed three gland phenotypes (n=3, 4.7%). Evans et al., (to which this chapter contributed to) extended this observation and showed that diversity was increased in patients that progressed to dysplasia. Therefore, the three cases displaying three phenotypes should be monitored closely for signs of progression in future endoscopies. Furthermore, there was no association between the length of the Barrett's segment and the gland phenotype observed in this cohort and this is shown in the bar graphs above **figures 7.1b** and **7.1c**.

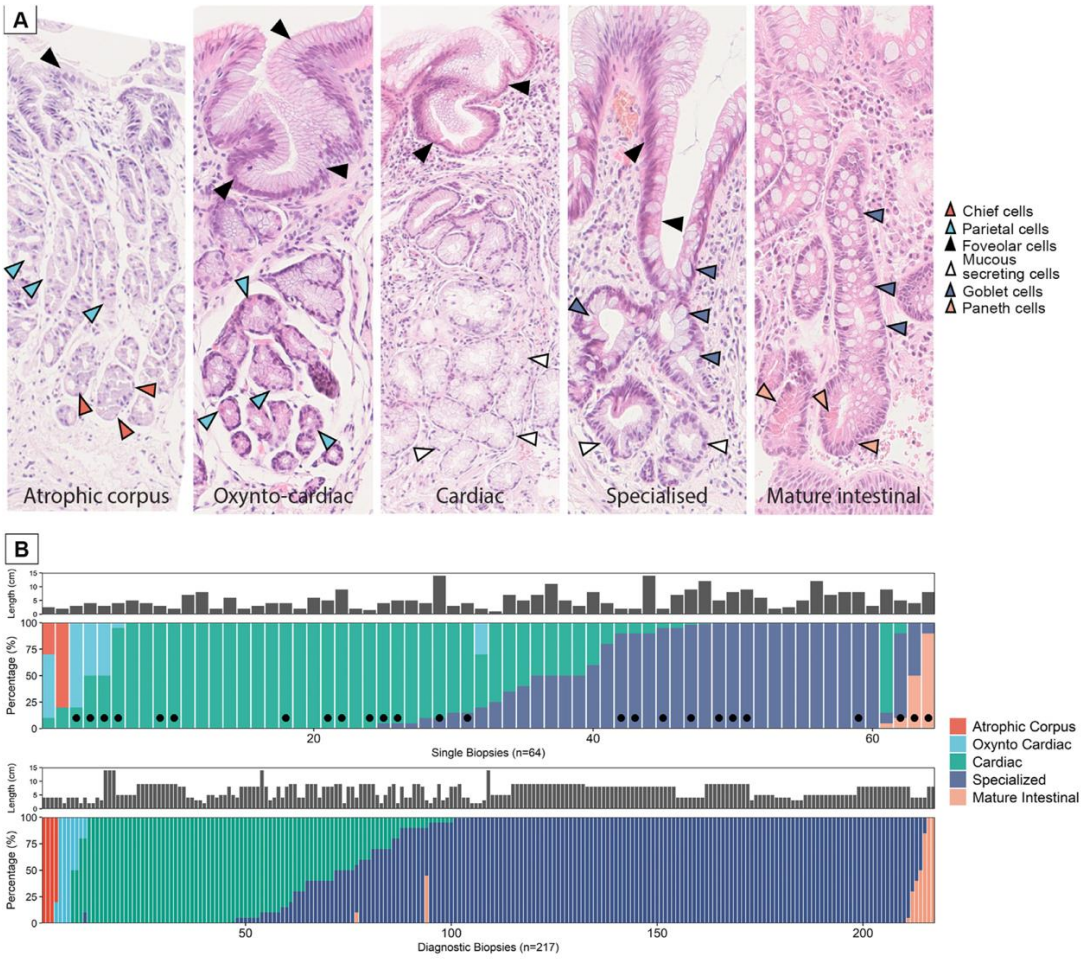

Figure 7.1: *The distribution of gland phenotypes in Barrett's oesophagus. (A) Representative H&E images of each gland phenotype observed. Left to right: Atrophic corpus, oxyntocardiac, cardiac, specialised and mature intestinal. (B) Phenotype distribution in biopsies taken adjacent to GOJ. Top, Barrett's length for each patient (biopsy). Bottom, the percentage of each gland phenotype within each biopsy. (C) For each patient marked (black circle) in (B) all Seattle biopsies were phenotyped. Top, Barrett's length for each patient (biopsy). Bottom, percentage of each phenotype per biopsy.*

7.3.2 Clonal ordering of gland phenotype within mixed glands

To demonstrate phenotypic gland evolution in BO, Dr Emanuella Carlotti and Dr James Evans determined if distinct gland phenotypes within biopsies share a common ancestor. I assisted in this experiment through collection and staining. This is added here purely for contextual sake and kind permission was given by my supervisor, Dr Stuart McDonald to include this data. My role was to investigate glands that displayed intragland mixing (section 7.4.3) and I performed these experiments.

The most frequent gland phenotypes observed in this cohort were cardiac and specialized glands. To determine if these show a common ancestor, we observed a biopsy that demonstrated a mixture of glands that were MUC2- MUC5AC+ (Cardiac glands) and those that were MUC2+ MUC5AC+ (specialised glands). **Figure 7.2a** shows an H&E with a cardiac gland labelled 1 and specialised gland labelled 2. This was confirmed by IHC for MUC2 (**Fig. 7.2b**) and MUC5AC (**Fig. 7.2c**). Each gland was microdissected (**Figs. 7.2d-e**). Interestingly a common mtDNA mutation (m.10492 T>C) was observed in both gland types indicating a common shared gland of origin) (**Figs. 7.2f-g**). The mutation was not observed in surrounding stroma and therefore a germline polymorphism was excluded.

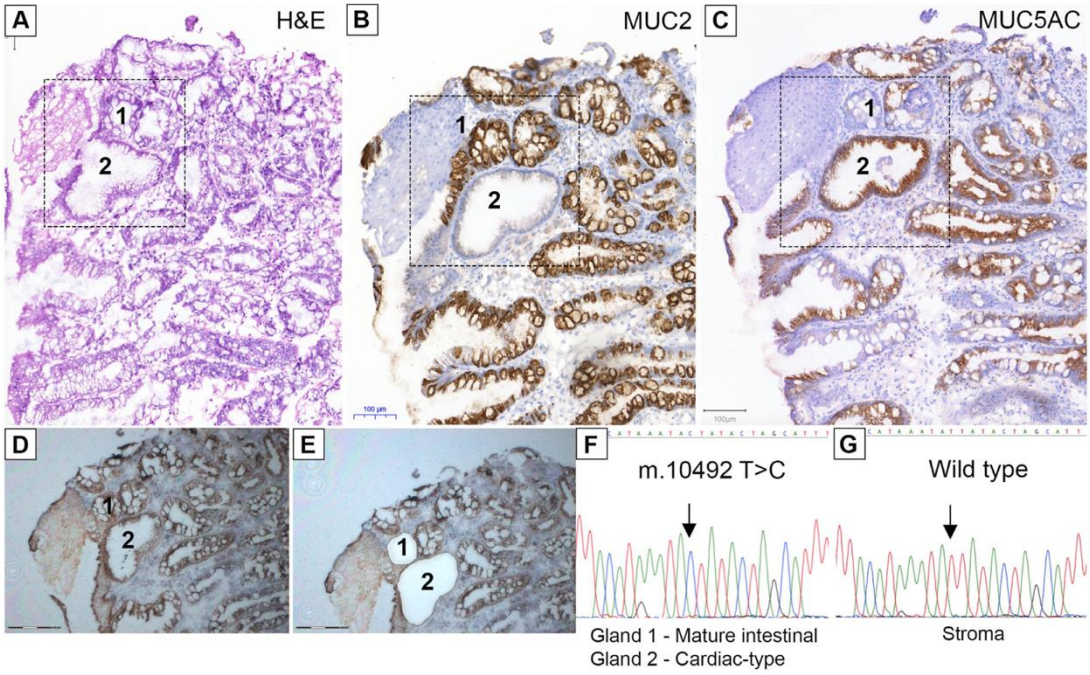

Figure 7.2: *Gland phenotype is an evolutionary process. (A) H&E indicating a specialised and cardiac gland (1 and 2 respectively) that were confirmed by IHC for MUC2 (B) and MUC5AC (C). Pre and post LCM figures are shown respectively (D) & (E). A common m.10492T>C mutation was observed in glands 1 and 2 (F) but not in adjacent stroma (G).*

7.3.3 Intragland phenotypic evolution

To determine the mechanism by which one gland phenotype converts to another we reviewed all biopsies from our non-dysplastic cohort and discovered 3 cases where there was clear mixing of phenotypes within individual glands. **Figure 7.3a** is an H&E section showing a clear cardiac (foveolar only) cell population in the surface portion of the gland with a goblet cell-only (specialized) isthmus portion of the gland. In a separate case we performed IHC for MUC2 and MUC5AC and showing expression is unique to distinct cell populations in some glands but not others (**Figs. 7.3b-c**). To confirm intragland phenotypic evolution we show a BE biopsy that contains entirely cardiac-type glands with the exception of a single gland that partially expressed both cardiac and specialized epithelium (**Fig. 7.4a**) identified by H&E and confirmed by an expert pathologist. Cardiac area is labelled blue and the specialised are labelled red respectively (**Fig. 7.4a**). Cells from each region were microdissected (**Figs. 7.4b-c**) and we detected a common heteroplasmic *m.3010 A>G* mutation in the MT-RNR2 region of the mitochondrial genome (**Figs. 7.4d-e**) that was not detected elsewhere in the biopsy. Interestingly, we also discovered a second heteroplasmic mutation, *m.2706 A>G* also located in the MT-RNR2 region that was detected only in the specialized cells of this gland (**Figs. 7.4d-e**). This data strongly indicates the presence of two clones within a single gland competing for clonal dominance, the process known as niche succession. The presence of an additional mutation in the specialized but not cardiac epithelium permits ordering of the timings of these mutations and is evidence that the specialized phenotype arose after the cardiac phenotype (**Fig. 7.4f**).

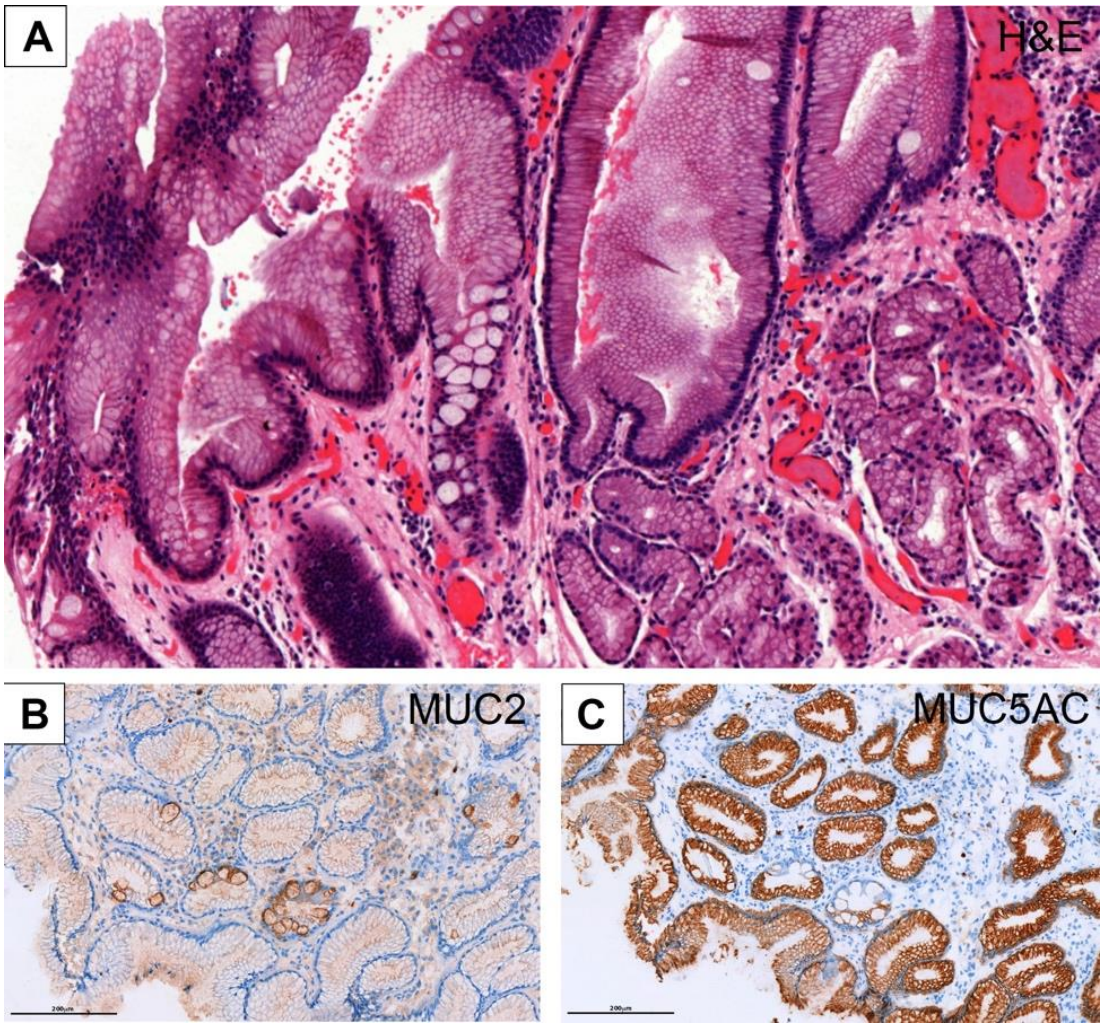

Figure 7.3: *Intragland phenotypic mixing in Barrett's oesophagus. (A) An H&E with a clear single gland only half full of goblet cells. (B) IHC for MUC2 and a serial stain for MUC5AC(C) show mixing of goblet cells and MUC5AC negative cells.*

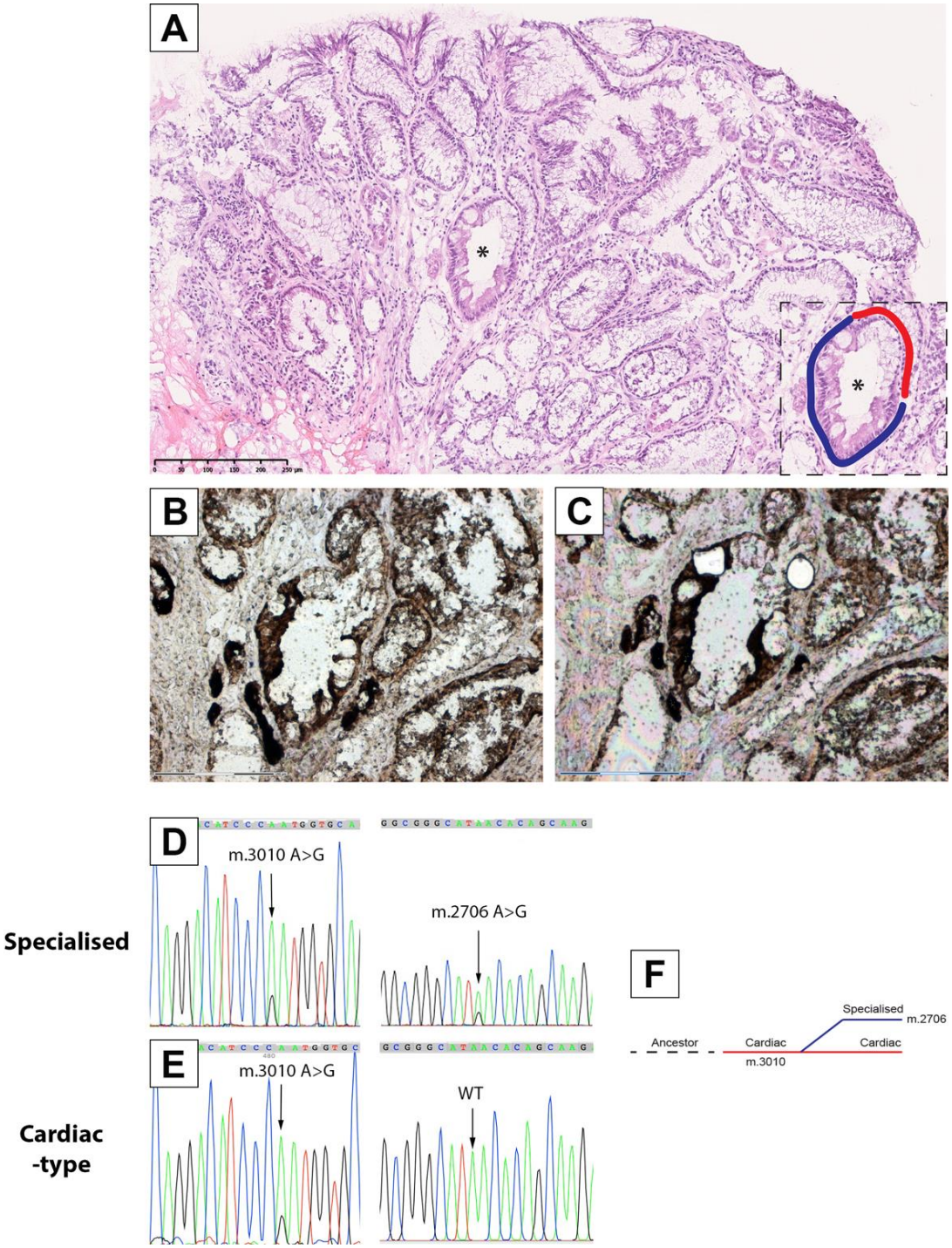

Figure 7.4: *Sequencing reveals clonal ordering of phenotypic evolution. (A) An H&E of a biopsy with entirely cardiac glands except a single gland that shows both cardiac (red) and specialised cells (blue)(Insert) within the same gland. (B) and (C) Pre- and post- LCM respectively from each area within the gland. (D) Both phenotypes shared a common m.3010A>G mutation but only the specialised gland displayed a m.2706A>G heteroplasmic mutation. (F) Clonal ordering reveals that specialised arose after the cardiac phenotype.*

7.3.4 ASM-Seq analysis reveals differential mitotic ages of cardiac and specialised glands.

Figure 7.5 shows the protocol for calling phenotypes in frozen sections. MUC2 and MUC5AC IHC was performed on serial sections and then depending on the outcome, a phenotype was called and the gland was microdissected over the subsequent 6 serial sections. Methylation is known to increase with patient age and therefore to determine if patients with either cardiac or specialised glands was as a result of age, we performed ASM-Seq on individual Barrett's glands that were laser capture microdissected.

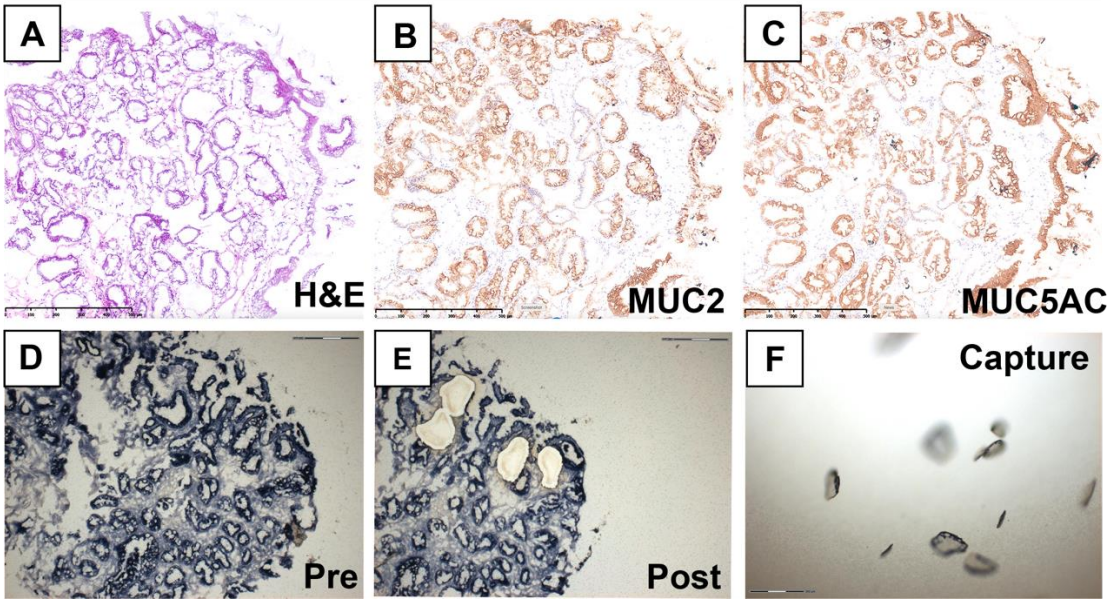

Figure 7.5: *A representative H&E (A) of a frozen biopsy section of non-dysplastic Barrett's. Serial IHC for MUC2 (B) and MUC5AC (C) revealed glands (in this case) that were all specialised. Multiple glands were microdissected, but ASM-Seq'd individually (D&E). All serial sections of individual glands were pooled and sequenced* (F).

To determine the role of patient age in pairwise difference (PWD) of methylation of our target genes, we plotted PWD of cardiac glands against patient age (**Fig. 7.6a**) and also for specialised glands against patient age (**Fig. 7.6b**). We found no correlation between age and phenotype. This is not completely unexpected due to the nature of Barrett's time of diagnosis being known yet the age at which it first developed is not.

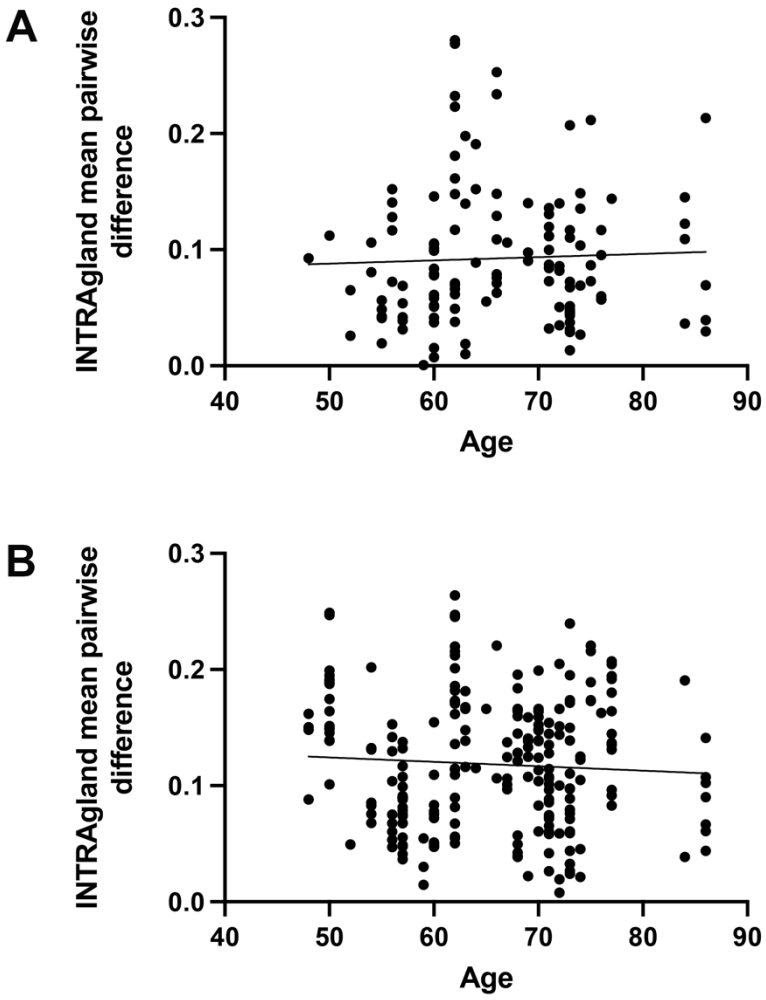

***Figure 7.6**. The relationship of patient age and phenotype of ASM-Seq pairwise differences. (A) All patients displaying cardiac glands were plotted against age at endoscopy. (B) All patients displaying specialised glands were plotted against age. Each dot represents an individual gland and its mean pairwise difference.*

Subsequent to age, we investigated to see if there were any pairwise differences between cardiac and specialised glands. Interestingly there was significantly increased PWD in specialised (n=229) compared to cardiac glands (n=125 glands) and based on a corrected mean PWD for each gland sequenced. This suggests that specialised glands are mitotically older compared to cardiac glands. A possible explanation for this is that cellular turnover is higher in specialised glands and therefore show greater clonal expansion potential. **Figure 7.7** shows the comparison of PWD in all phenotyped glands.

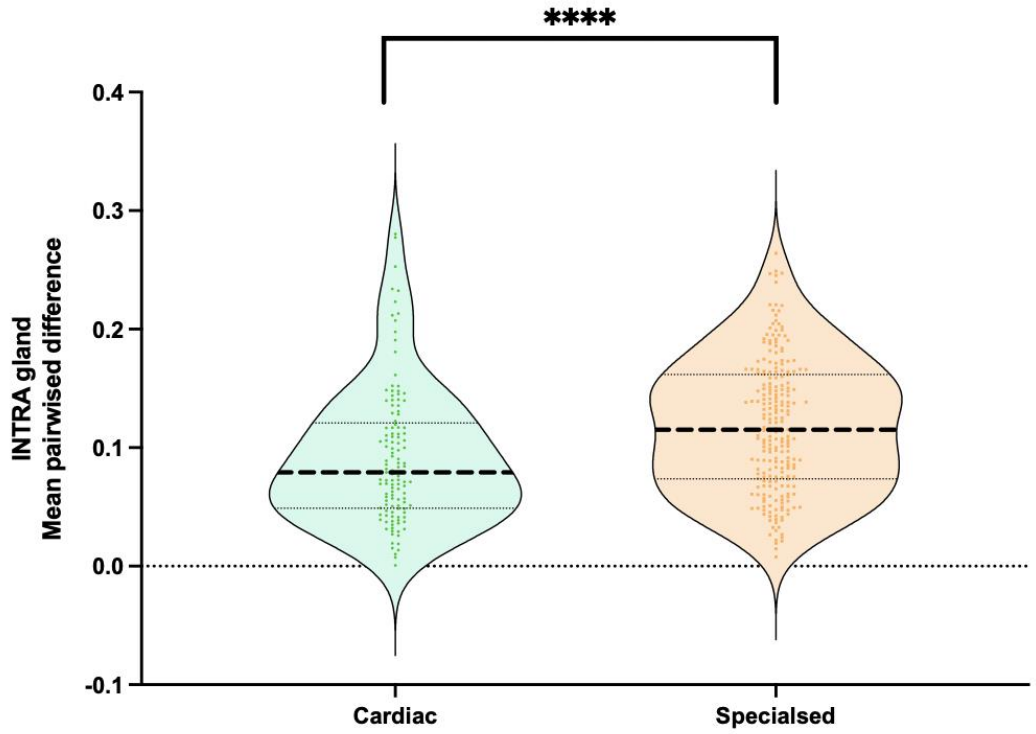

***Figure 7.7**: Mean PWD analysis reveals specialised glands are mitotically older than cardiac. This is based on intragland analysis and therefore each dot represents a single gland. ****=P<0.0001*

It has been hypothesised that specialised glands pose a greater risk of progression than other gland phenotypes. To determine if glands of each phenotype demonstrate a similar mitotic age to dysplastic glands, we microdissected pathologist confirmed dysplastic glands from frozen sections of patients with high grade dysplasia (n=48 glands) and compared their PWD with the PWD of both specialised (n=229) and cardiac glands (n=124). **Figure 7.8a** and **7.8b** both show that while individual phenotypes show no difference between dysplastic glands, together (from Chapter 6) dysplastic glands show a higher mitotic age than non-dysplastic glands.

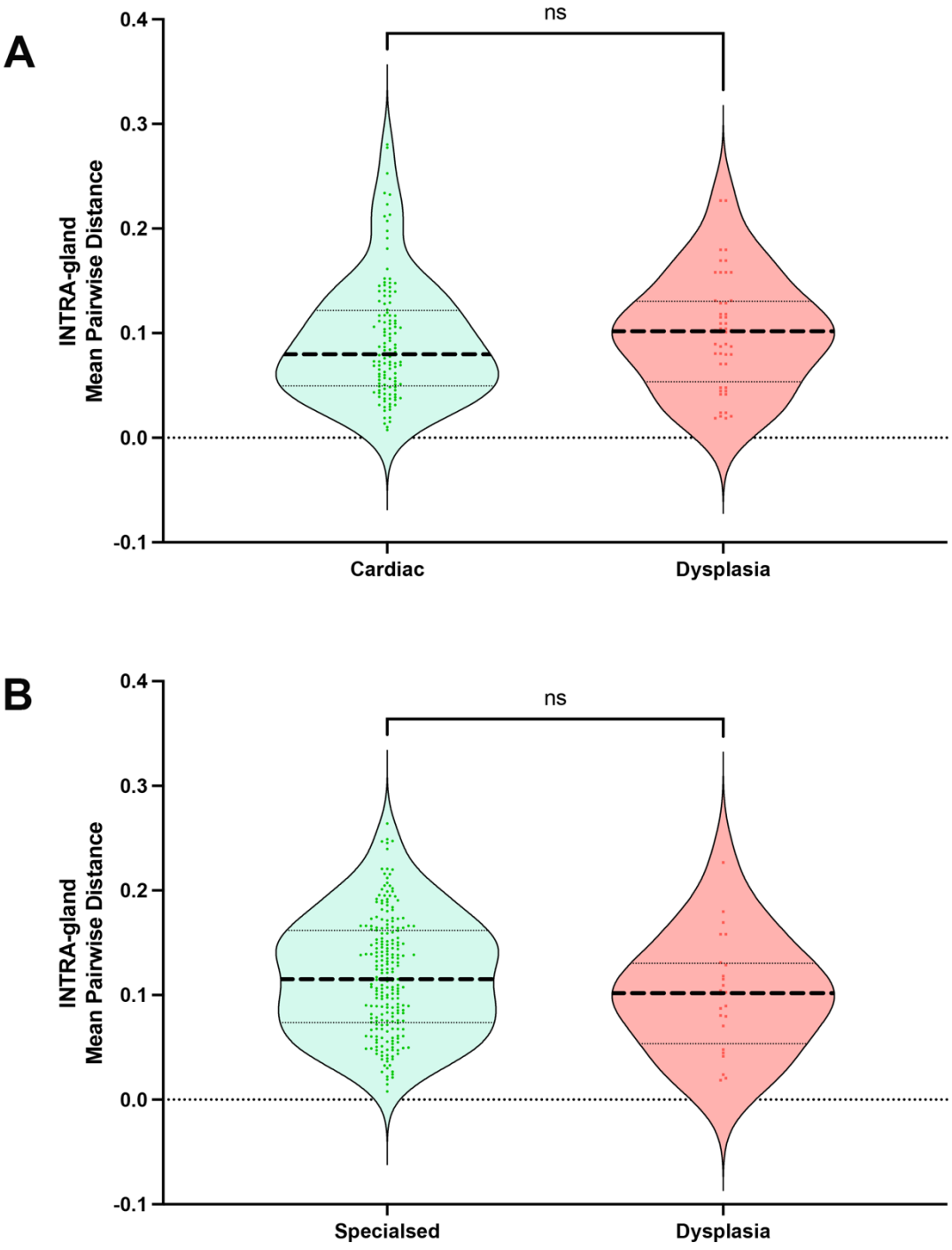

Figure 7.8: *Mean intragland PWD of gland phenotype compared with dysplastic glands. (A) Cardiac and (B) specialised glands showed no significant pairwise difference between dysplastic glands.*

7.3.5 The cellular dynamics of gland bases and gland tops

My host laboratory has previously demonstrated using label retaining methodology that Barrett's glands show bi-migration patterns from stem cells migrating up the pit of the gland and down to the gland base. Due to technological restrictions at the time, we now apply ASM-Seq to gland bases and gland tops to determine if there is any differential in cell migration dynamics between cell positions. We show in **figure 7.9** that indeed microdissected tops of glands show a significantly increased mean PWD compared to bases. This indicates a significantly higher turnover and therefore more methylation errors within cells that migrate to the tops of the gland.

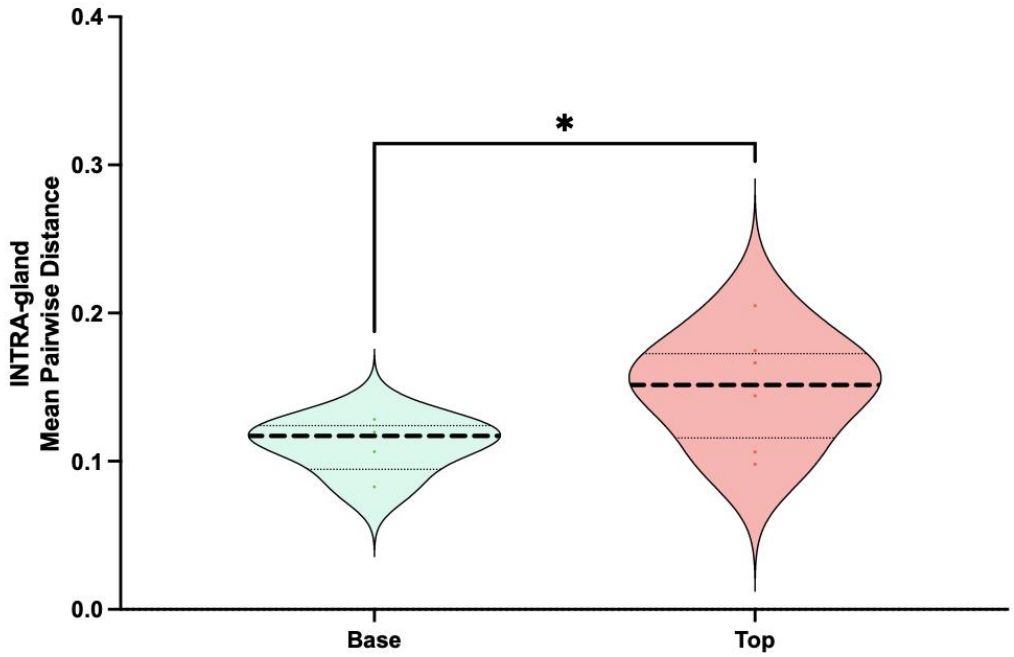

***Figure 7.9**. Comparison of cell location mean PWD in individual Barrett's (specialised) glands. Tops of glands show significantly increased mean PWD. *=p<0.05*

7.4 Discussion

The role of phenotype in the evolution of Barrett's and its role in progression to cancer is very poorly understood. We know that Barrett's displays a rich spectrum of gland phenotypes and we have shown here that within glands adjacent to the GOJ (a location that many believe is common to all Barrett's patients) display as rich a phenotypic diversity as do biopsies taken from throughout the Barrett's lesion (Figure 7.1). This is interesting, especially when we consider work by Evans et al., [305]that have shown that phenotypic diversity is increased in Barrett's patients prior to the onset of dysplasia. Previously, there has been a lot of research done into the genetic abnormalities of Barrett's where several groups have demonstrated an increase in copy number alterations (CNAs) in patients also prior to the development of dysplasia [185, 237, 238, 360]. While important, identifying such genetic abnormalities in patients in the general population is difficult and has not been trialled in a large cohort of patients. The Fitzgerald group in Cambridge has had some success in sequencing such genetic changes and are awaiting the outcome of a new, much larger trial[52]. It would be easier and far better for pathologists however, if they could use standardised IHC as a means to understand potential cancer risk in a very large population of patients. Additionally, when we consider that natural selection plays an important role in the progression of all cancers, it is important to remember that phenotype not genotype is the factor upon which all selection is based[99].

The data presented here adds greatly to our understanding of the distribution of gland phenotype and also the evolution from one phenotype to another. Data, of which I was involved but not the primary research, demonstrates clonal relationships between patches of Barrett's glands that display distinct phenotypes (Figure 7.2) however, the mechanism that appears to be at play is niche succession based on selection of specific phenotypes. Niche succession is the process by which a single stem cell within the stem cell niche (here the Barrett's stem cell niche is thought to be the neck of the gland[100]) with time, will divide sufficiently to take over the entire niche, expelling all competitors. It is known that niche succession is dependent on

natural selection based partly on its location within the stem cell niche and any oncogenic advantages it may have. We have been able to effectively demonstrate a similar process in the Barrett's gland. Using mtDNA Sanger sequencing of microdissected phenotypically distinct areas of glands that show more than one phenotypic lineage, we were able to detect differentially expressed mitochondrial DNA mutations where both lineages shared a common mutation but only the specialised lineage possessed a second mutation (**Fig. 7.4**). The odds of 2 glands with distinct epithelial phenotypes possessing the same mtDNA mutation independently is vanishingly small and has been estimated at less than $1:10^{-9}$ to 1 [131]. This means that the specialised lineage must have arisen at a point in time after the cardiac lineage. This is an important observation as it adds dynamics to our increasing understanding of the role of phenotype in the evolution of Barrett's and perhaps future research will show its role in developing cancer.

ASM-Seq has greatly added to our understanding of the mitotic age of Barrett's glands. Sequencing multiple CpG-rich sites within genes that are not expressed in Barrett's under any known circumstances, allows us to compare the mitotic history of glands when we know that after every cellular division there is a chance of an error being made changing the methylated CpG make up of any gene. The fact that these genes are not expressed in Barrett's means that changes in the local environment will not affect selection of cells that as this is neutral to this situation. As discussed in previous chapters, next generation methylation sequencing allows us to compare in a CpG site-specific manner how many methylation changes have occurred at hundreds of CpG sites. Knowing that methylation patterns are inherited from mother to daughter cells, any error at a known rate means we can calculate the number of changes that have occurred and the number is proportional to the number of cell divisions. With the advent of next generation sequencing, this number has vastly increased allowing greater accuracy on how many cell divisions and therefore how mitotically old a gland is compared with its neighbours or those with a different phenotype. Data presented here shows that specialised glands show greater diversity of methylation sequence and therefore we can infer that many more cell divisions have occurred in its natural history. This means that in the context of Barrett's,

specialised epithelium has a selective advantage over cardiac epithelium, allowing it to dominant the epithelial landscape[99, 305]. Interestingly, we did not observe a significant different between any individual gland phenotype and PWD compared with microdissected dysplastic glands. This suggests that the progression of dysplasia is no dependent of any individual phenotype but rather the mitotic age of the lesion itself. Chapter 6 demonstrates a robust relationship between non-dysplastic and dysplastic mitotic age, it is however not dependent on phenotype. The selection 'event' is therefore likely to be external to the epithelial cell where (for example) and inflammatory combined with a low pH environment is more important than a single epithelial phenotype. Evidence for this comes from a recent paper from my host laboratory where non-goblet columnar epithelium was shown to the be the gland-of-origin for adenocarcinoma.

Future research using ASM-Seq will be able to further refine and model the cellular and clonal expansions of gland phenotypes in combination with other, potential genetic markers. Given more time, and of course no COVID, we would have been able to generate such models that could be used to determine better the pathway to cancer in Barrett's oesophagus.

8 Discussion

8.1 Discussion

The entire concept of this thesis relied heavily on design and optimisation of the novel ASM-Seq technique described in full detail in results Chapter 5. A process that was envisaged to take a year took almost double that owing to the difficulties and challenges that have been described throughout this work and detailed further in the supplements. Nevertheless, the development of such a technique was ambitious but has demonstrated through the subsequent data presented in Chapters 6 and 7 its worth in obtaining a greater understanding of Barrett's oesophagus.

The ASM-Seq protocol is the first of its kind to successfully tag 5' DNA templates of bisulfite or enzymatically converted DNA with a UMI such that PCR and sequencing error and biases can be detected and reduced to the greatest degree possible through careful and considered bioinformatic processing. Furthermore it gives a first glimpse into the intricate detail that can be achieved in epigenetic sequencing on a DNA molecule by DNA molecule basis. While traditional techniques of methylation sequencing have their advantages they rely heavily on averaging and inferring. Whereas ASM-Seq delivers all the aggregate discrete epigenetic data of a particular biological system such that sifting and sorting through suddenly reveals many permutations of analysis. There are of course improvements and further optimisation that could be undertaken but are out of the scope of this thesis. In particular, the methylation plots are variable and need retesting on a regular basis to understand whether this is a one-off issue or more systemic. Sensitivity testing could also be bettered to pick up more of the template DNA. While a lot of the issue with this could be blamed on bisulfite it could also probably be improved with better primer design, for example shorter sequences and higher GC content. Finally there is real scope to expand the complexity and coverage of the technique with additional primer sets and targets of interests. It has been designed in such a way to be transferrable and as bespoke as any researcher would need.

The key findings that ASM-Seq permitted in Barrett's were the recognition that patients who progress to cancer have a mitotically older lesion as evidenced by greater diversity methylation patterns, a function of cell division. Additionally, there is the finding of rapid reduction in diversity patterns prior to the onset of malignancy that is in all probability characteristic of a neoplastic clonal expansion across the segment. This is coupled with intragland data that demonstrates this reduction in the progressor cohort with a lesion spatially spreading to homogenise the "near and far" patterns. This latter finding is not present in the equivalent non-progressor cohort who's segments remains in a state of indolence but continue to develop intraglandular stem cell diversity through the process evolutionary gradualism[356]. In phenotypic space we have demonstrated the clonal ordering and transition of a gastric phenotype to intestinal metaplasia mitochondrial sequencing and their mutations as a lineage trace.

Through greater understanding of the Barrett's lesion and its histogenesis there is the possibility of finally identifying a reliable and ubiquitous biomarker that can be used in the clinical sphere to reassure and discharge those patients who will never progress to cancer but instead confidently detect those with a lesion that is destined for malignancy. This Thesis has added to the rapidly expanding body of work in the field and presented in this thesis over the course of it chapters. Through continued analysis of this large dataset reported in this thesis further clonal relationships in the malignant and non-malignant phenotype I am in no doubt will be revealed.

8.2 Future work

This thesis presents novel work related to a new laboratory technique. While this technique is up and running further optimisation is possible to improve sensitivity and reliability. There is also the potential to reduce PCR cycle number and workflow clean-up steps that I did not have the time to re-trial. From a bioinformatics perspective, there is great scope to further enhance the output from this raw data. In particular drawing an extensive spatial map of all the cases aligned with their sequencing outputs. In addition, phylogenetic analysis could also be undertaken to

fully characterise the clonal dynamic in never before seen detail. Such a project requires significant bioinformatics and mathematical expertise that is beyond the scope of this thesis and is best completed as part of a wider collaboration. Furthermore, the case mix can be enhanced with more progressors over multiple timepoints to fully understand its clonal evolution to cancer and this proposed reduction in methylation diversity at the onset of transition. Obtaining and sequencing more "pre-progression" points would also be exciting to see. Finally, there remains the unanswered question of origins of Barrett's. Through taking further samples from the cardia and squamous and subjecting them to ASM-Seq this may reveal clonal relationships, especially of coupled with advanced bioinformatics.

9 References

1. Feldman M, Friedman LS, Brandt LJ. Sleisenger and Fordtran's Gastrointestinal and Liver Disease: Pathophysiology, Diagnosis, Management: Elsevier Health Sciences; 2020.

2. Mittal RK, Balaban DH. The esophagogastric junction. *N Engl J Med.* 1997;336:924-932.

3. Fitzgerald RC, di Pietro M, Ragunath K, Ang Y, Kang J-Y, Watson P, Trudgill N, Patel P, Kaye PV, Sanders S, O'Donovan M, Bird-Lieberman E, Bhandari P, Jankowski JA, Attwood S, Parsons SL, Loft D, Lagergren J, Moayyedi P, Lyratzopoulos G, de Caestecker J. British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. *Gut.* 2013;63:7-42.

4. Takubo K, Aida J, Sawabe M, Arai T, Kato H, Pech O, Arima M. The normal anatomy around the oesophagogastric junction: a histopathologic view and its correlation with endoscopy. *Best practice & research Clinical gastroenterology.* 2008;22:569-583.

5. Al Yassin TM, Toner PG. Fine structure of squamous epitheilum and submucosal glands of human oesophagus. *J Anat.* 1977;123:705-721.

6. Squier CA, Kremer MJ. Biology of oral mucosa and esophagus. *J Natl Cancer Inst Monogr.* 2001:7-15.

7. Wright NA, Alison M. The biology of epithelial cell populations. Vol 1. Oxford [Oxfordshire] : New York: Clarendon Press ; Oxford University Press; 1984.

8. Alcolea MP. Oesophageal Stem Cells and Cancer. *Advances in experimental medicine and biology.* 2017;1041:187-206.

9. Seery JP. Stem cells of the oesophageal epithelium. *Journal of cell science.* 2002;115:1783-1789.

10. Seery J, Watt F. Asymmetric stem-cell divisions define the architecture of human oesophageal epithelium. *Current biology : CB.* 2000;10:1447-1450.

11. Wang DH, Souza RF. Transcommitment: Paving the Way to Barrett's Metaplasia. *Advances in experimental medicine and biology.* 2016;908:183-212.

12. Wright NA. Black Bible Vol2. 1984:1-723.

13. Long JD, Orlando RC. Esophageal submucosal glands: structure and function. *Am J Gastroenterol.* 1999;94:2818-2824.

14. Kuo B, Urma D. Esophagus - anatomy and development. *GI Motility Online.* 2006;Part 1 Oral Cavity, pharynx and esophagus.

15. Naini BV, Souza RF, Odze RD. Barrett's Esophagus: A Comprehensive and Contemporary Review for Pathologists. *Am J Surg Pathol.* 2016;40:e45-66.

16. Peters Y, Al-Kaabi A, Shaheen NJ, Chak A, Blum A, Souza RF, Di Pietro M, Iyer PG, Pech O, Fitzgerald RC, Siersema PD. Barrett oesophagus. *Nat Rev Dis Primers.* 2019;5:35.

17. Slack JMW. Metaplasia and transdifferentiation: from pure biology to the clinic. *Nature Reviews Molecular Cell Biology.* 2007;8:369-378.

18. Tan WK, di Pietro M, Fitzgerald RC. Past, present and future of Barrett's oesophagus. *Eur J Surg Oncol.* 2017;43:1148-1160.

19. Shaheen NJ, Falk GW, Iyer PG, Gerson LB, American College of G. ACG Clinical Guideline: Diagnosis and Management of Barrett's Esophagus. *Am J Gastroenterol.* 2016;111:30-50; quiz 51.

20. Shaheen NJ, Falk GW, Iyer PG, Souza RF, Yadlapati RH, Sauer BG, Wani S. Diagnosis and Management of Barrett's Esophagus: An Updated ACG Guideline. *Am J Gastroenterol.* 2022;117:559-587.

21. di Pietro M, Fitzgerald RC, group BSGBsgw. Revised British Society of Gastroenterology recommendation on the diagnosis and management of Barrett's oesophagus with low-grade dysplasia. *Gut.* 2018;67:392-393.

22. Spechler SJ, Sharma P, Souza RF, Inadomi JM, Shaheen NJ. American Gastroenterological Association medical position statement on the management of Barrett's esophagus. *Gastroenterology.* 2011;140:1084-1091.

23. Spechler SJ, Sharma P, Souza RF, Inadomi JM, Shaheen NJ. American Gastroenterological Association technical review on the management of Barrett's esophagus. *Gastroenterology.* 2011;140:e18-52; quiz e13.

24. Quante M, Graham TA, Jansen M. Insights Into the Pathophysiology of Esophageal Adenocarcinoma. *Gastroenterology.* 2018;154:406-420.

25. Lagergren J, Bergström R, Lindgren A, Nyrén O. Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma. *N Engl J Med.* 1999;340:825-831.

26. Booth CL, Thompson KS. Barrett's esophagus: A review of diagnostic criteria, clinical surveillance practices and new developments. *J Gastrointest Oncol.* 2012;3:232-242.

27. Moawad FJ, Young PE, Gaddam S, Vennalaganti P, Thota PN, Vargo J, Cash BD, Falk GW, Sampliner RE, Lieberman D, Sharma P. Barrett's oesophagus length is established at the time of initial endoscopy and does not change over time: results from a large multicentre cohort. *Gut.* 2015;64:1874-1880.

28. Sharma P, Dent J, Armstrong D, Bergman JJ, Gossner L, Hoshihara Y, Jankowski JA, Junghard O, Lundell L, Tytgat GN, Vieth M. The development and validation of an endoscopic grading system for Barrett's esophagus: the Prague C & M criteria. *Gastroenterology.* 2006;131:1392-1399.

29. Pohl H, Pech O, Arash H, Stolte M, Manner H, May A, Kraywinkel K, Sonnenberg A, Ell C. Length of Barrett's oesophagus and cancer risk: implications from a large sample of patients with early oesophageal adenocarcinoma. *Gut.* 2016;65:196-201.

30. Iftikhar SY, James PD, Steele RJ, Hardcastle JD, Atkinson M. Length of Barrett's oesophagus: an important factor in the development of dysplasia and adenocarcinoma. *Gut.* 1992;33:1155-1158.

31. Avidan B, Sonnenberg A, Schnell TG, Chejfec G, Metz A, Sontag SJ. Hiatal hernia size, Barrett's length, and severity of acid reflux are all risk factors for esophageal adenocarcinoma. *Am J Gastroenterol.* 2002;97:1930-1936.

32. Desai TK, Krishnan K, Samala N, Singh J, Cluley J, Perla S, Howden CW. The incidence of oesophageal adenocarcinoma in non-dysplastic Barrett's oesophagus: a meta-analysis. *Gut.* 2012;61:970-976.

33. Vahabzadeh B, Seetharam AB, Cook MB, Wani S, Rastogi A, Bansal A, Early DS, Sharma P. Validation of the Prague C & M criteria for the endoscopic grading of Barrett's esophagus by gastroenterology trainees: a multicenter study. *Gastrointest Endosc.* 2012;75:236-241.

34. Levine DS, Blount PL, Rudolph RE, Reid BJ. Safety of a systematic endoscopic biopsy protocol in patients with Barrett's esophagus. *The American Journal of Gastroenterology.* 2000;95:1152-1157.

35. Tschanz ER. Do 40% of Patients Resected for Barrett Esophagus With High-Grade Dysplasia Have Unsuspected Adenocarcinoma? *Archives of pathology & laboratory medicine.* 2005;129:177-180.

36. Kandiah K, Chedgy FJQ, Subramaniam S, Longcroft-Wheaton G, Bassett P, Repici A, Sharma P, Pech O, Bhandari P. International development and validation of a classification system for the identification of Barrett's neoplasia using acetic acid chromoendoscopy: the Portsmouth acetic acid classification (PREDICT). *Gut.* 2017.

37. Abrams JA, Kapel RC, Lindberg GM, Saboorian MH, Genta RM, Neugut AI, Lightdale CJ. Adherence to biopsy guidelines for Barrett's esophagus surveillance in the community setting in the United States. *Clin Gastroenterol Hepatol.* 2009;7:736-742; quiz 710.

38. Kariv R, Plesec TP, Goldblum JR, Bronner M, Oldenburgh M, Rice TW, Falk GW. The Seattle protocol does not more reliably predict the detection of cancer at the time of esophagectomy than a less intensive surveillance protocol. *Clin Gastroenterol Hepatol.* 2009;7:653-658; quiz 606.

39. Schlemper RJ, Riddell RH, Kato Y, Borchard F, Cooper HS, Dawsey SM, Dixon MF, Fenoglio-Preiser CM, Flejou JF, Geboes K, Hattori T, Hirota T, Itabashi M, Iwafuchi M, Iwashita A, Kim YI, Kirchner T, Klimpfinger M, Koike M, Lauwers GY, Lewin KJ, Oberhuber G, Offner F, Price AB, Rubio CA, Shimizu M, Shimoda T, Sipponen P, Solcia E, Stolte M, Watanabe H, Yamabe H. The Vienna classification of gastrointestinal epithelial neoplasia. *Gut.* 2000;47:251-255.

40. De Angelis R, Sant M, Coleman MP, Francisci S, Baili P, Pierannunzio D, Trama A, Visser O, Brenner H, Ardanaz E, Bielska-Lasota M, Engholm G, Nennecke A, Siesling S, Berrino F, Capocaccia R. Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE--5-a population-based study. *The Lancet Oncology.* 2014;15:23-34.

41. Wong MCS, Hamilton W, Whiteman DC, Jiang JY, Qiao Y, Fung FDH, Wang HHX, Chiu PWY, Ng EKW, Wu JCY, Yu J, Chan FKL, Sung JJY. Global Incidence and mortality of oesophageal cancer and their correlation with socioeconomic indicators temporal patterns and trends in 41 countries. *Sci Rep.* 2018;8:4522.

42. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer.* 2010;127:2893-2917.

43. Smyth EC, Lagergren J, Fitzgerald RC, Lordick F, Shah MA, Lagergren P, Cunningham D. Oesophageal cancer. *Nat Rev Dis Primers.* 2017;3:17048.

44. Gatenby PA, Hainsworth A, Caygill C, Watson A, Winslet M. Projections for oesophageal cancer incidence in England to 2033. *Eur J Cancer Prev.* 2011;20:283-286.

45. Hur C, Miller M, Kong CY, Dowling EC, Nattinger KJ, Dunn M, Feuer EJ. Trends in esophageal adenocarcinoma incidence and mortality. *Cancer.* 2013;119:1149-1158.

46. Smittenaar CR, Petersen KA, Stewart K, Moitt N. Cancer incidence and mortality projections in the UK until 2035. *Br J Cancer.* 2016;115:1147-1155.

47. Vaughan TL, Fitzgerald RC. Precision prevention of oesophageal adenocarcinoma. *Nat Rev Gastroenterol Hepatol.* 2015;12:243-248.

48. Zeng Y, Ruan W, Liu J, Liang W, He J, Cui F, Pan H, He J. Esophageal cancer in patients under 50: a SEER analysis. *J Thorac Dis.* 2018;10:2542-2550.

49. Cancer Research UK Oesophageal Cancer Statistics. Accessed Online at **https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer**. *Cancer Research UK.* 2019.

50. Wenker TN, Tan MC, Liu Y, El-Serag HB, Thrift AP. Prior Diagnosis of Barrett's Esophagus Is Infrequent, but Associated with Improved Esophageal Adenocarcinoma Survival. *Dig Dis Sci.* 2018.

51. Runge TM, Abrams JA, Shaheen NJ. Epidemiology of Barrett's Esophagus and Esophageal Adenocarcinoma. *Gastroenterology clinics of North America.* 2015;44:203-231.

52. Fitzgerald RC, di Pietro M, O'Donovan M, Maroni R, Muldrew B, Debiram-Beecham I, Gehrung M, Offman J, Tripathi M, Smith SG, Aigret B, Walter FM, Rubin G, Sasieni P. Cytosponge-trefoil factor 3 versus usual care to identify Barrett's oesophagus in a primary care setting: a multicentre, pragmatic, randomised controlled trial. *Lancet.* 2020;396:333-344.

53. Lin EC, Holub J, Lieberman D, Hur C. Low Prevalence of Suspected Barrett's Esophagus in Patients With Gastroesophageal Reflux Disease Without Alarm Symptoms. *Clin Gastroenterol Hepatol.* 2019;17:857-863.

54. Hamade N, Weng G, Desai M, Chandrasekar VT, Dasari C, Kennedy K, Sharma P. Significant decline in the prevalence of Barrett's esophagus among patients with gastroesophageal reflux disease. *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus.* 2021;34.

55. Rex DK, Cummings OW, Shaw M, Cumings MD, Wong RK, Vasudeva RS, Dunne D, Rahmani EY, Helper DJ. Screening for Barrett's esophagus in colonoscopy patients with and without heartburn. *Gastroenterology.* 2003;125:1670-1677.

56. Ronkainen J, Aro P, Storskrubb T, Johansson SE, Lind T, Bolling-Sternevald E, Vieth M, Stolte M, Talley NJ, Agreus L. Prevalence of Barrett's esophagus in the general population: an endoscopic study. *Gastroenterology.* 2005;129:1825-1831.

57. Cameron AJ, Lomboy CT. Barrett's esophagus: Age, prevalence, and extent of columnar epithelium. *Gastroenterology.* 1992;103:1241-1245.

58. den Hoed CM, van Blankenstein M, Dees J, Kuipers EJ. The minimal incubation period from the onset of Barrett's oesophagus to symptomatic adenocarcinoma. *Br J Cancer.* 2011;105:200-205.

59. Hvid-Jensen F, Pedersen L, Drewes AM, Sorensen HT, Funch-Jensen P. Incidence of adenocarcinoma among patients with Barrett's esophagus. *N Engl J Med.* 2011;365:1375-1383.

60. Nguyen T, Thrift AP, Yu X, Duan Z, El-Serag HB. The Annual Risk of Esophageal Adenocarcinoma Does Not Decrease Over Time in Patients With Barrett's Esophagus. *Am J Gastroenterol.* 2017;112:1049-1055.

61. Gatenby P, Bhattacharjee S, Wall C, Caygill C, Watson A. Risk stratification for malignant progression in Barrett's esophagus: Gender, age, duration and year of surveillance. *World J Gastroenterol.* 2016;22:10592-10600.

62. Sikkema M, de Jonge PJ, Steyerberg EW, Kuipers EJ. Risk of esophageal adenocarcinoma and mortality in patients with Barrett's esophagus: a systematic review and meta-analysis. *Clin Gastroenterol Hepatol.* 2010;8:235-244; quiz e232.

63. Bhat S, Coleman HG, Yousef F, Johnston BT, McManus DT, Gavin AT, Murray LJ. Risk of malignant progression in Barrett's esophagus patients: results from a large population-based study. *J Natl Cancer Inst.* 2011;103:1049-1057.

64. Gatenby P, Caygill C, Wall C, Bhatacharjee S, Ramus J, Watson A, Winslet M. Lifetime risk of esophageal adenocarcinoma in patients with Barrett's esophagus. *World Journal of Gastroenterology : WJG.* 2014;20:9611-9617.

65. Curvers WL, ten Kate FJ, Krishnadath KK, Visser M, Elzer B, Baak LC, Bohmer C, Mallant-Hent RC, van Oijen A, Naber AH, Scholten P, Busch OR, Blaauwgeers HG, Meijer GA, Bergman JJ. Low-grade dysplasia in Barrett's

esophagus: overdiagnosed and underestimated. *Am J Gastroenterol.* 2010;105:1523-1530.

66. Duits LC, Phoa KN, Curvers WL, Ten Kate FJ, Meijer GA, Seldenrijk CA, Offerhaus GJ, Visser M, Meijer SL, Krishnadath KK, Tijssen JG, Mallant-Hent RC, Bergman JJ. Barrett's oesophagus patients with low-grade dysplasia can be accurately risk-stratified after histological review by an expert pathology panel. *Gut.* 2015;64:700-706.

67. Lim CH, Treanor D, Dixon MF, Axon AT. Low-grade dysplasia in Barrett's esophagus has a high risk of progression. *Endoscopy.* 2007;39:581-587.

68. Kestens C, Offerhaus GJ, van Baal JW, Siersema PD. Patients With Barrett's Esophagus and Persistent Low-grade Dysplasia Have an Increased Risk for High-grade Dysplasia and Cancer. *Clin Gastroenterol Hepatol.* 2016;14:956-962 e951.

69. Singh S, Manickam P, Amin AV, Samala N, Schouten LJ, Iyer PG, Desai TK. Incidence of esophageal adenocarcinoma in Barrett's esophagus with low-grade dysplasia: a systematic review and meta-analysis. *Gastrointest Endosc.* 2014;79:897-909 e894; quiz 983 e891, 983 e893.

70. Collard J-M. High-grade dysplasia in Barrett's esophagus: The case for esophagectomy. *Chest Surgery Clinics.* 2002;12:77-92.

71. Rastogi A, Puli S, El-Serag HB, Bansal A, Wani S, Sharma P. Incidence of esophageal adenocarcinoma in patients with Barrett's esophagus and high-grade dysplasia: a meta-analysis. *Gastrointest Endosc.* 2008;67:394-398.

72. Weston AP, Sharma P, Topalovski M, Richards R, Cherian R, Dixon A. Long-term follow-up of Barrett's high-grade dysplasia. *Am J Gastroenterol.* 2000;95:1888-1893.

73. Shaheen NJ, Provenzale D, Sandler RS. Upper endoscopy as a screening and surveillance tool in esophageal adenocarcinoma: a review of the evidence. *Am J Gastroenterol.* 2002;97:1319-1327.

74. Corley DA, Mehtani K, Quesenberry C, Zhao W, de Boer J, Weiss NS. Impact of endoscopic surveillance on mortality from Barrett's esophagus-associated esophageal adenocarcinomas. *Gastroenterology.* 2013;145:312-319 e311.

75. Conio M, Blanchi S, Lapertosa G, Ferraris R, Sablich R, Marchi S, D'Onofrio V, Lacchin T, Iaquinto G, Missale G, Ravelli P, Cestari R, Benedetti G, Macrì G, Fiocca R, Munizzi F, Filiberti R. Long-term endoscopic surveillance of patients with Barrett's esophagus. Incidence of dysplasia and adenocarcinoma: a prospective study. *Am J Gastroenterol.* 2003;98:1931-1939.

76. Macdonald CE, Wicks AC, Playford RJ. Final results from 10 year cohort of patients undergoing surveillance for Barrett's oesophagus: observational study. *Bmj.* 2000;321:1252-1255.

77. Qiao Y, Hyder A, Bae SJ, Zarin W, O'Neill TJ, Marcon NE, Stein L, Thein HH. Surveillance in Patients With Barrett's Esophagus for Early Detection of Esophageal Adenocarcinoma: A Systematic Review and Meta-Analysis. *Clinical and translational gastroenterology.* 2015;6:e131.

78. Verbeek RE, Leenders M, Ten Kate FJ, van Hillegersberg R, Vleggaar FP, van Baal JW, van Oijen MG, Siersema PD. Surveillance of Barrett's esophagus and mortality from esophageal adenocarcinoma: a population-based cohort study. *Am J Gastroenterol.* 2014;109:1215-1222.

79. Codipilly DC, Chandar AK, Singh S, Wani S, Shaheen NJ, Inadomi JM, Chak A, Iyer PG. The Effect of Endoscopic Surveillance in Patients With Barrett's Esophagus: A Systematic Review and Meta-analysis. *Gastroenterology.* 2018;154:2068-2086.e2065.

80. Hamade N, Kamboj AK, Krishnamoorthi R, Singh S, Hassett LC, Katzka DA, Kahi CJ, Fatima H, Iyer PG. Systematic review with meta-analysis: neoplasia detection rate and post-endoscopy Barrett's neoplasia in Barrett's oesophagus. *Alimentary pharmacology & therapeutics.* 2021;54:546-559.

81. Han S, Wani S. Quality Indicators in Barrett's Esophagus: Time to Change the Status Quo. *Clinical endoscopy.* 2018;51:344-351.

82. Network CGAR. Integrated genomic characterization of oesophageal carcinoma. *Nature.* 2017;541:169-175.

83. Sawas T, Killcoyne S, Iyer PG, Wang KK, Smyrk TC, Kisiel JB, Qin Y, Ahlquist DA, Rustgi AK, Costa RJ, Gerstung M, Fitzgerald RC, Katzka DA. Identification of Prognostic Phenotypes of Esophageal Adenocarcinoma in 2 Independent Cohorts. *Gastroenterology.* 2018;155:1720-1728.e1724.

84. Curtius K, Rubenstein JH, Chak A, Inadomi JM. Computational modelling suggests that Barrett's oesophagus may be the precursor of all oesophageal adenocarcinomas. *Gut.* 2020;70:1435-1440.

85. Nowicki-Osuch K, Zhuang L, Jammula S, Bleaney CW, Mahbubani KT, Devonshire G, Katz-Summercorn A, Eling N, Wilbrey-Clark A, Madissoon E, Gamble J, Di Pietro M, O'Donovan M, Meyer KB, Saeb-Parsy K, Sharrocks AD, Teichmann SA, Marioni JC, Fitzgerald RC. Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. *Science.* 2021;373:760-767.

86. Wang Z, Kambhampati S, Cheng Y, Ma K, Simsek C, Tieu AH, Abraham JM, Liu X, Prasath V, Duncan M, Stark A, Trick A, Tsai HL, Wang H, He Y, Khashab MA, Ngamruengphong S, Shin EJ, Wang TH, Meltzer SJ. Methylation Biomarker Panel Performance in EsophaCap Cytology Samples for Diagnosing Barrett's Esophagus: A Prospective Validation Study. *Clin Cancer Res.* 2019;25:2127-2135.

87. Moinova HR, LaFramboise T, Lutterbaugh JD, Chandar AK, Dumot J, Faulx A, Brock W, De la Cruz Cabrera O, Guda K, Barnholtz-Sloan JS, Iyer PG, Canto MI, Wang JS, Shaheen NJ, Thota PN, Willis JE, Chak A, Markowitz SD. Identifying DNA methylation biomarkers for non-endoscopic detection of Barrett's esophagus. *Sci Transl Med.* 2018;10.

88. Visrodia K, Singh S, Krishnamoorthi R, Ahlquist DA, Wang KK, Iyer PG, Katzka DA. Magnitude of Missed Esophageal Adenocarcinoma After Barrett's Esophagus Diagnosis: A Systematic Review and Meta-analysis. *Gastroenterology.* 2016;150:599-607.e597; quiz e514-595.

89. Vithayathil M, Modolell I, Ortiz-Fernandez-Sordo J, Oukrif D, Pappas A, Januszewicz W, O'Donovan M, Hadjinicolaou A, Bianchi M, Blasko A, White J, Kaye P, Novelli M, Wernisch L, Ragunath K, di Pietro M. Image-Enhanced Endoscopy and Molecular Biomarkers Vs Seattle Protocol to Diagnose Dysplasia in Barrett's Esophagus. *Clin Gastroenterol Hepatol.* 2022.

90. Greenberg SB, Shaheen NJ. Endoscopic Surveillance of Barrett's Esophagus: Using Old Principles and New Technology to Improve Care. *Am J Gastroenterol.* 2022;117:201-204.

91. Wolfson P, Ho KMA, Wilson A, McBain H, Hogan A, Lipman G, Dunn J, Haidry R, Novelli M, Olivo A, Lovat LB. Endoscopic eradication therapy for Barrett's esophagus-related neoplasia: a final 10-year report from the UK National HALO Radiofrequency Ablation Registry. *Gastrointest Endosc.* 2022.

92. Komanduri S, Muthusamy VR, Wani S. Controversies in Endoscopic Eradication Therapy for Barrett's Esophagus. *Gastroenterology.* 2018.

93. Beg S, Ragunath K, Wyman A, Banks M, Trudgill N, Pritchard DM, Riley S, Anderson J, Griffiths H, Bhandari P, Kaye P, Veitch A. Quality standards in upper gastrointestinal endoscopy: a position statement of the British Society of Gastroenterology (BSG) and Association of Upper Gastrointestinal Surgeons of Great Britain and Ireland (AUGIS). *Gut.* 2017;66:1886-1899.

94. Alzoubaidi D, Ragunath K, Wani S, Penman ID, Trudgill NJ, Jansen M, Banks M, Bhandari P, Morris AJ, Willert R, Boger P, Smart HL, Ravi N, Dunn J, Gordon C, Mannath J, Mainie I, di Pietro M, Veitch AM, Thorpe S, Magee C, Everson M,

Sami S, Bassett P, Graham D, Attwood S, Pech O, Sharma P, Lovat LB, Haidry R. Quality indicators for Barrett's endotherapy (QBET): UK consensus statements for patients undergoing endoscopic therapy for Barrett's neoplasia. *Frontline gastroenterology.* 2020;11:259-271.

95. Sharma P, Katzka DA, Gupta N, Ajani J, Buttar N, Chak A, Corley D, El-Serag H, Falk GW, Fitzgerald R, Goldblum J, Gress F, Ilson DH, Inadomi JM, Kuipers EJ, Lynch JP, McKeon F, Metz D, Pasricha PJ, Pech O, Peek R, Peters JH, Repici A, Seewald S, Shaheen NJ, Souza RF, Spechler SJ, Vennalaganti P, Wang K. Quality indicators for the management of Barrett's esophagus, dysplasia, and esophageal adenocarcinoma: international consensus recommendations from the American Gastroenterological Association Symposium. *Gastroenterology.* 2015;149:1599-1606.

96. Vennalaganti P, Kanakadandi V, Goldblum JR, Mathur SC, Patil DT, Offerhaus GJ, Meijer SL, Vieth M, Odze RD, Shreyas S, Parasa S, Gupta N, Repici A, Bansal A, Mohammad T, Sharma P. Discordance Among Pathologists in the United States and Europe in Diagnosis of Low-Grade Dysplasia for Patients With Barrett's Esophagus. *Gastroenterology.* 2017;152:564-570.e564.

97. Montgomery E, Bronner MP, Goldblum JR, Greenson JK, Haber MM, Hart J, Lamps LW, Lauwers GY, Lazenby AJ, Lewin DN, Robert ME, Toledano AY, Shyr Y, Washington K. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation. *Hum Pathol.* 2001;32:368-378.

98. Salomao MA, Lam-Himlin D, Pai RK. Substantial Interobserver Agreement in the Diagnosis of Dysplasia in Barrett Esophagus Upon Review of a Patient's Entire Set of Biopsies. *Am J Surg Pathol.* 2018;42:376-381.

99. McDonald SAC, Graham TA, Lavery DL, Wright NA, Jansen M. The Barrett's Gland in Phenotype Space. *JCMGH.* 2015;1:41-54.

100. Lavery DL, Nicholson AM, Poulsom R, Jeffery R, Hussain A, Gay LJ, Jankowski JA, Zeki SS, Barr H, Harrison R, Going J, Kadirkamanathan S, Davis P, Underwood T, Novelli MR, Rodriguez-Justo M, Shepherd N, Jansen M, Wright NA, McDonald SA. The stem cell organisation, and the proliferative and gene expression profile of Barrett's epithelium, replicates pyloric-type gastric glands. *Gut.* 2014;63:1854-1863.

101. Lavery DL, Martinez P, Gay LJ, Cereser B, Novelli MR, Rodriguez-Justo M, Meijer SL, Graham TA, McDonald SA, Wright NA, Jansen M. Evolution of oesophageal adenocarcinoma from metaplastic columnar epithelium without goblet cells in Barrett's oesophagus. *Gut.* 2016;65:907-913.

102. Takubo K, Aida J, Naomoto Y, Sawabe M, Arai T, Shiraishi H, Matsuura M, Ell C, May A, Pech O, Stolte M, Vieth M. Cardiac rather than intestinal-type

background in endoscopic resection specimens of minute Barrett adenocarcinoma. *Hum Pathol.* 2009;40:65-74.

103. Srivastava A, Golden KL, Sanchez CA, Liu K, Fong PY, Li X, Cowan DS, Rabinovitch PS, Reid BJ, Blount PL, Odze RD. High Goblet Cell Count Is Inversely Associated with Ploidy Abnormalities and Risk of Adenocarcinoma in Barrett's Esophagus. *PLoS One.* 2015;10:e0133403.

104. Kelty CJ, Gough MD, Van Wyk Q, Stephenson TJ, Ackroyd R. Barrett's oesophagus: intestinal metaplasia is not essential for cancer risk. *Scand J Gastroenterol.* 2007;42:1271-1274.

105. Gatenby PAC, Ramus JR, Caygill CPJ, Shepherd NA, Watson A. Relevance of the detection of intestinal metaplasia in non-dysplastic columnar-lined oesophagus. *Scandinavian journal of gastroenterology.* 2008;43:524-530.

106. Harrison R, Perry I, Haddadin W, McDonald S, Bryan R, Abrams K, Sampliner R, Talley NJ, Moayyedi P, Jankowski JA. Detection of intestinal metaplasia in Barrett's esophagus: an observational comparator study suggests the need for a minimum of eight biopsies. *Am J Gastroenterol.* 2007;102:1154-1161.

107. Paull A, Trier JS, Dalton MD, Camp RC, Loeb P, Goyal RK. The histologic spectrum of Barrett's esophagus. *N Engl J Med.* 1976;295:476-480.

108. Chandrasoma PT, Der R, Dalton P, Kobayashi G, Ma Y, Peters J, Demeester T. Distribution and significance of epithelial types in columnar-lined esophagus. *Am J Surg Pathol.* 2001;25:1188-1193.

109. Piazuelo MB, Haque S, Delgado A, Du JX, Rodriguez F, Correa P. Phenotypic differences between esophageal and gastric intestinal metaplasia. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc.* 2003;17:62-74.

110. Theodorou D, Ayazi S, DeMeester SR, Zehetner J, Peyre CG, Grant KS, Augustin F, Oh DS, Lipham JC, Chandrasoma PT, Hagen JA, DeMeester TR. Intraluminal pH and goblet cell density in Barrett's esophagus. *J Gastrointest Surg.* 2012;16:469-474.

111. Grondin JA, Kwon YH, Far PM, Haq S, Khan WI. Mucins in Intestinal Mucosal Defense and Inflammation: Learning From Clinical and Experimental Studies. *Frontiers in immunology.* 2020;11:2054.

112. Jass JR. Mucin core proteins as differentiation markers in the gastrointestinal tract. *Histopathology.* 2000;37:561-564.

113. Reis CA, David L, Correa P, Carneiro F, de Bolós C, Garcia E, Mandel U, Clausen H, Sobrinho-Simões M. Intestinal metaplasia of human stomach displays distinct patterns of mucin (MUC1, MUC2, MUC5AC, and MUC6) expression. *Cancer research.* 1999;59:1003-1007.

114. Dixon J, Strugala V, Griffin SM, Welfare MR, Dettmar PW, Allen A, Pearson JP. Esophageal mucin: an adherent mucus gel barrier is absent in the normal esophagus but present in columnar-lined Barrett's esophagus. *Am J Gastroenterol.* 2001;96:2575-2583.

115. Freund JN, Domon-Dell C, Kedinger M, Duluc I. The Cdx-1 and Cdx-2 homeobox genes in the intestine. *Biochemistry and cell biology = Biochimie et biologie cellulaire.* 1998;76:957-969.

116. Hahn HP, Blount PL, Ayub K, Das KM, Souza R, Spechler S, Odze RD. Intestinal differentiation in metaplastic, nongoblet columnar epithelium in the esophagus. *Am J Surg Pathol.* 2009;33:1006-1015.

117. Groisman GM, Amar M, Meir A. Expression of the intestinal marker Cdx2 in the columnar-lined esophagus with and without intestinal (Barrett's) metaplasia. *Mod Pathol.* 2004;17:1282-1288.

118. Phillips RW, Frierson HF, Jr., Moskaluk CA. Cdx2 as a marker of epithelial intestinal differentiation in the esophagus. *Am J Surg Pathol.* 2003;27:1442-1447.

119. Glickman JN, Blount PL, Sanchez CA, Cowan DS, Wongsurawat VJ, Reid BJ, Odze RD. Mucin core polypeptide expression in the progression of neoplasia in Barrett's esophagus. *Hum Pathol.* 2006;37:1304-1315.

120. Reed CC, Shaheen NJ. Durability of Endoscopic Treatment for Dysplastic Barrett's Esophagus. *Curr Treat Options Gastroenterol.* 2019;17:171-186.

121. Gupta M, Iyer PG, Lutzke L, Gorospe EC, Abrams JA, Falk GW, Ginsberg GG, Rustgi AK, Lightdale CJ, Wang TC, Fudman DI, Poneros JM, Wang KK. Recurrence of esophageal intestinal metaplasia after endoscopic mucosal resection and radiofrequency ablation of Barrett's esophagus: results from a US Multicenter Consortium. *Gastroenterology.* 2013;145:79-86 e71.

122. Titi M, Overhiser A, Ulusarac O, Falk GW, Chak A, Wang K, Sharma P. Development of subsquamous high-grade dysplasia and adenocarcinoma after successful radiofrequency ablation of Barrett's esophagus. *Gastroenterology.* 2012;143:564-566 e561.

123. Minacapelli CD, Bajpai M, Geng X, Cheng CL, Chouthai AA, Souza R, Spechler SJ, Das KM. Barrett's metaplasia develops from cellular reprograming of

esophageal squamous epithelium due to gastroesophageal reflux. *American journal of physiology Gastrointestinal and liver physiology.* 2017;312:G615-G622.

124. Wang DH. The Esophageal Squamous Epithelial Cell-Still a Reasonable Candidate for the Barrett's Esophagus Cell of Origin? *Cell Mol Gastroenterol Hepatol.* 2017;4:157-160.

125. Odze R, Spechler SJ, Podgaetz E, Nguyen A, Konda V, Souza RF. Histologic Study of the Esophagogastric Junction of Organ Donors Reveals Novel Glandular Structures in Normal Esophageal and Gastric Mucosae. *Clinical and translational gastroenterology.* 2021;12:e00346.

126. Goldenring JR. Pyloric metaplasia, pseudopyloric metaplasia, ulcer-associated cell lineage and spasmolytic polypeptide-expressing metaplasia: reparative lineages in the gastrointestinal mucosa. *J Pathol.* 2018;245:132-137.

127. Liber AF. Aberrant pyloric glands in regional ileitis. *AMA archives of pathology.* 1951;51:205-212.

128. Lee FD. Pyloric Metaplasia in the Small Intestine. *The Journal of pathology and bacteriology.* 1964;87:267-277.

129. Correa P. A human model of gastric carcinogenesis. *Cancer research.* 1988;48:3554-3560.

130. Graham DY. History of Helicobacter pylori, duodenal ulcer, gastric ulcer and gastric cancer. *World J Gastroenterol.* 2014;20:5191-5204.

131. Greaves LC, Preston SL, Tadrous PJ, Taylor RW, Barron MJ, Oukrif D, Leedham SJ, Deheragoda M, Sasieni P, Novelli MR, Jankowski JAZ, Turnbull DM, Wright NA, McDonald SAC. Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. *Proceedings of the National Academy of Sciences of the United States of America.* 2006;103:714-719.

132. Nicholson AM, Graham TA, Simpson A, Humphries A, Burch N, Rodriguez-Justo M, Novelli M, Harrison R, Wright NA, McDonald SA, Jankowski JA. Barrett's metaplasia glands are clonal, contain multiple stem cells and share a common squamous progenitor. *Gut.* 2012;61:1380-1389.

133. Robertson EV, Derakhshan MH, Wirz AA, Mitchell DR, Going JJ, Kelman AW, Ballantyne SA, McColl KEL. Hiatus hernia in healthy volunteers is associated with intrasphincteric reflux and cardiac mucosal lengthening without traditional reflux. *Gut.* 2017;66:1208-1215.

134. Robertson EV, Derakhshan MH, Wirz AA, Lee YY, Seenan JP, Ballantyne SA, Hanvey SL, Kelman AW, Going JJ, McColl KE. Central obesity in asymptomatic volunteers is associated with increased intrasphincteric acid reflux and lengthening of the cardiac mucosa. *Gastroenterology.* 2013;145:730-739.

135. Quante M, Bhagat G, Abrams JA, Marache F, Good P, Lee MD, Lee Y, Friedman R, Asfaha S, Dubeykovskaya Z, Mahmood U, Figueiredo J-L, Kitajewski J, Shawber C, Lightdale CJ, Rustgi AK, Wang TC. Bile acid and inflammation activate gastric cardia stem cells in a mouse model of Barrett-like metaplasia. *Cancer cell.* 2012;21:36-51.

136. Wang X, Ouyang H, Yamamoto Y, Kumar PA, Wei TS, Dagher R, Vincent M, Lu X, Bellizzi AM, Ho KY, Crum CP, Xian W, McKeon F. Residual embryonic cells as precursors of a Barrett's-like metaplasia. *Cell.* 2011;145:1023-1035.

137. Senoo M, Pinto F, Crum CP, McKeon F. p63 Is essential for the proliferative potential of stem cells in stratified epithelia. *Cell.* 2007;129:523-536.

138. Jiang M, Li H, Zhang Y, Yang Y, Lu R, Liu K, Lin S, Lan X, Wang H, Wu H, Zhu J, Zhou Z, Xu J, Lee DK, Zhang L, Lee YC, Yuan J, Abrams JA, Wang TC, Sepulveda AR, Wu Q, Chen H, Sun X, She J, Chen X, Que J. Transitional basal cells at the squamous-columnar junction generate Barrett's oesophagus. *Nature.* 2017;550:529-533.

139. Boch JA, Shields HM, Antonioli DA, Zwas F, Sawhney RA, Trier JS. Distribution of cytokeratin markers in Barrett's specialized columnar epithelium. *Gastroenterology.* 1997;112:760-765.

140. Shields HM, Zwas F, Antonioli DA, Doos WG, Kim S, Spechler SJ. Detection by scanning electron microscopy of a distinctive esophageal surface cell at the junction of squamous and Barrett's epithelium. *Dig Dis Sci.* 1993;38:97-108.

141. Shields HM, Rosenberg SJ, Zwas FR, Ransil BJ, Lembo AJ, Odze R. Prospective evaluation of multilayered epithelium in Barrett's esophagus. *Am J Gastroenterol.* 2001;96:3268-3273.

142. Dunn LJ, Burt AD, Hayes N, Griffin SM. Columnar Metaplasia in the Esophageal Remnant After Esophagectomy: A Common Occurrence and a Valuable Insight Into the Development of Barrett Esophagus. *Ann Surg.* 2016;264:1016-1021.

143. Que J, Garman KS, Souza RF, Spechler SJ. Pathogenesis and Cells of Origin of Barrett's Esophagus. *Gastroenterology.* 2019.

144. Willet SG, Lewis MA, Miao ZF, Liu D, Radyk MD, Cunningham RL, Burclaff J, Sibbel G, Lo HG, Blanc V, Davidson NO, Wang ZN, Mills JC. Regenerative

proliferation of differentiated cells by mTORC1-dependent paligenosis. *EMBO J.* 2018;37.

145. Huo X, Zhang X, Yu C, Cheng E, Zhang Q, Dunbar KB, Pham TH, Lynch JP, Wang DH, Bresalier RS, Spechler SJ, Souza RF. Aspirin prevents NF-kappaB activation and CDX2 expression stimulated by acid and bile salts in oesophageal squamous cells of patients with Barrett's oesophagus. *Gut.* 2018;67:606-615.

146. Souza RF. Reflux esophagitis and its role in the pathogenesis of Barrett's metaplasia. *J Gastroenterol.* 2017;52:767-776.

147. Nishimura K, Tanaka T, Tanaka Y, Matono S, Murata K, Shirouzu K, Fujita H. Reflux esophagitis and columnar-lined esophagus after cervical esophagogastrostomy (following esophagectomy). *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus.* 2010;23:94-99.

148. Carballo GB, Honorato JR, de Lopes GPF, Spohr T. A highlight on Sonic hedgehog pathway. *Cell communication and signaling : CCS.* 2018;16:11.

149. Vercauteren Drubbel A, Pirard S, Kin S, Dassy B, Lefort A, Libert F, Nomura S, Beck B. Reactivation of the Hedgehog pathway in esophageal progenitors turns on an embryonic-like program to initiate columnar metaplasia. *Cell Stem Cell.* 2021;28:1411-1427.e1417.

150. Wang DH, Clemons NJ, Miyashita T, Dupuy AJ, Zhang W, Szczepny A, Corcoran-Schwartz IM, Wilburn DL, Montgomery EA, Wang JS, Jenkins NA, Copeland NA, Harmon JW, Phillips WA, Watkins DN. Aberrant epithelial-mesenchymal Hedgehog signaling characterizes Barrett's metaplasia. *Gastroenterology.* 2010;138:1810-1822.

151. Katz JM MA, Basit H. . Embryology, Esophagus. [Updated 2021 Aug 11]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: **https://www.ncbi.nlm.nih.gov/books/NBK542304/**.

152. Litingtung Y, Lei L, Westphal H, Chiang C. Sonic hedgehog is essential to foregut development. *Nat Genet.* 1998;20:58-61.

153. Leedham SJ, Preston SL, McDonald SA, Elia G, Bhandari P, Poller D, Harrison R, Novelli MR, Jankowski JA, Wright NA. Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human Barrett's oesophagus. *Gut.* 2008;57:1041-1048.

154. Coad RA, Woodman AC, Warner PJ, Barr H, Wright NA, Shepherd NA. 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. *J Pathol.* 2005;206:388-394.

155. Biddlestone LR, Barham CP, Wilkinson SP, Barr H, Shepherd NA. The histopathology of treated Barrett's esophagus: squamous reepithelialization after acid suppression and laser and photodynamic therapy. *Am J Surg Pathol.* 1998;22:239-245.

156. Lorinc E, Oberg S. Submucosal glands in the columnar-lined oesophagus: evidence of an association with metaplasia and neosquamous epithelium. *Histopathology.* 2012;61:53-58.

157. Gonzalez G, Huang Q, Mashimo H. Characterization of oncocytes in deep esophageal glands. *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus.* 2016;29:670-680.

158. Paulson TG, Xu L, Sanchez C, Blount PL, Ayub K, Odze RD, Reid BJ. Neosquamous epithelium does not typically arise from Barrett's epithelium. *Clin Cancer Res.* 2006;12:1701-1706.

159. 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-Bockler B, Lu X. Single cell RNA-seq reveals profound transcriptional similarity between Barrett's oesophagus and oesophageal submucosal glands. *Nat Commun.* 2018;9:4261.

160. Hutchinson L, Stenstrom B, Chen D, Piperdi B, Levey S, Lyle S, Wang TC, Houghton J. 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 Dev.* 2011;20:11-17.

161. Sarosi G, Brown G, Jaiswal K, Feagins LA, Lee E, Crook TW, Souza RF, Zou YS, Shay JW, Spechler SJ. Bone marrow progenitor cells contribute to esophageal regeneration and metaplasia in a rat model of Barrett's esophagus. *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus.* 2008;21:43-50.

162. Souza RF, Spechler SJ. Mechanisms and pathophysiology of Barrett oesophagus. *Nature Reviews Gastroenterology & Hepatology.* 2022.

163. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, Mitchell TJ, Rubanova Y, Anur P, Yu K, Tarabichi M, Deshwar A, Wintersinger J, Kleinheinz K, Vázquez-García I, Haase K, Jerman L, Sengupta S, Macintyre G, Malikic S, Donmez N, Livitz DG, Cmero M, Demeulemeester J, Schumacher S, Fan Y, Yao X, Lee J, Schlesner M, Boutros PC, Bowtell DD, Zhu H, Getz G, Imielinski M, Beroukhim R, Sahinalp SC, Ji Y, Peifer M, Markowetz F,

Mustonen V, Yuan K, Wang W, Morris QD, Spellman PT, Wedge DC, Van Loo P. The evolutionary history of 2,658 cancers. *Nature.* 2020;578:122-128.

164. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg Å, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Imielinsk M, Jäger N, Jones DTW, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt ANJ, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Initiative APCG, Consortium IBC, Consortium IM-S, PedBrain I, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. *Nature.* 2013;500:415-421.

165. Ross-Innes CS, Becq J, Warren A, Cheetham RK, Northen H, O'Donovan M, Malhotra S, di Pietro M, Ivakhno S, He M, Weaver JMJ, Lynch AG, Kingsbury Z, Ross M, Humphray S, Bentley D, Fitzgerald RC. Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat Genet.* 2015;47:1038-1046.

166. Weaver JMJ, Ross-Innes CS, Shannon N, Lynch AG, Forshew T, Barbera M, Murtaza M, Ong C-AJ, Lao-Sirieix P, Dunning MJ, Smith L, Smith ML, Anderson CL, Carvalho B, O'Donovan M, Underwood TJ, May AP, Grehan N, Hardwick R, Davies J, Oloumi A, Aparicio S, Caldas C, Eldridge MD, Edwards PAW, Rosenfeld N, Tavaré S, Fitzgerald RC, Consortium O. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nature genetics.* 2014;46:837-843.

167. Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, Bandla S, Imamura Y, Schumacher SE, Shefler E, McKenna A, Carter SL, Cibulskis K, Sivachenko A, Saksena G, Voet D, Ramos AH, Auclair D, Thompson K, Sougnez C, Onofrio RC, Guiducci C, Beroukhim R, Zhou Z, Lin L, Lin J, Reddy R, Chang A, Landrenau R, Pennathur A, Ogino S, Luketich JD, Golub TR, Gabriel SB, Lander ES, Beer DG, Godfrey TE, Getz G, Bass AJ. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature Publishing Group.* 2013;45:478-486.

168. Stachler MD, Taylor-Weiner A, Peng S, McKenna A, Agoston AT, Odze RD, Davison JM, Nason KS, Loda M, Leshchiner I, Stewart C, Stojanov P, Seepo S, Lawrence MS, Ferrer-Torres D, Lin J, Chang AC, Gabriel SB, Lander ES, Beer DG, Getz G, Carter SL, Bass AJ. Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat Genet.* 2015;47:1047-1055.

169. Newell F, Patel K, Gartside M, Krause L, Brosda S, Aoude LG, Loffler KA, Bonazzi VF, Patch AM, Kazakoff SH, Holmes O, Xu Q, Wood S, Leonard C, Lampe G, Lord RV, Whiteman DC, Pearson JV, Nones K, Waddell N, Barbour AP. Complex structural rearrangements are present in high-grade dysplastic Barrett's oesophagus samples. *BMC medical genomics.* 2019;12:31.

170. Agrawal N, Jiao Y, Bettegowda C, Hutfless SM, Wang Y, David S, Cheng Y, Twaddell WS, Latt NL, Shin EJ, Wang L-D, Wang L, Yang W, Velculescu VE, Vogelstein B, Papadopoulos N, Kinzler KW, Meltzer SJ. Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer discovery.* 2012;2:899-905.

171. Massague J. TGFbeta signalling in context. *Nat Rev Mol Cell Biol.* 2012;13:616-630.

172. Streppel MM, Lata S, DelaBastide M, Montgomery EA, Wang JS, Canto MI, Macgregor-Das AM, Pai S, Morsink FH, Offerhaus GJ, Antoniou E, Maitra A, McCombie WR. Next-generation sequencing of endoscopic biopsies identifies ARID1A as a tumor-suppressor gene in Barrett's esophagus. *Oncogene.* 2014;33:347-357.

173. Barrett MT, Sanchez CA, Prevo LJ, Wong DJ, Galipeau PC, Paulson TG, Rabinovitch PS, Reid BJ. Evolution of neoplastic cell lineages in Barrett oesophagus. *Nature genetics.* 1999;22:106-109.

174. Galipeau PC, Prevo LJ, Sanchez CA, Longton GM, Reid BJ. Clonal expansion and loss of heterozygosity at chromosomes 9p and 17p in premalignant esophageal (Barrett's) tissue. *J Natl Cancer Inst.* 1999;91:2087-2095.

175. Galipeau PC, Cowan DS, Sanchez CA, Barrett MT, Emond MJ, Levine DS, Rabinovitch PS, Reid BJ. 17p (p53) allelic losses, 4N (G2/tetraploid) populations, and progression to aneuploidy in Barrett's esophagus. *Proc Natl Acad Sci U S A.* 1996;93:7081-7084.

176. Morales CP, Souza RF, Spechler SJ. Hallmarks of cancer progression in Barrett's oesophagus. *Lancet.* 2002;360:1587-1589.

177. Wong DJ, Paulson TG, Prevo LJ, Galipeau PC, Longton G, Blount PL, Reid BJ. p16(INK4a) lesions are common, early abnormalities that undergo clonal expansion in Barrett's metaplastic epithelium. *Cancer Res.* 2001;61:8284-8289.

178. Wong DJ, Barrett MT, Stöger R, Emond MJ, Reid BJ. p16INK4a promoter is hypermethylated at a high frequency in esophageal adenocarcinomas. *Cancer research.* 1997;57:2619-2622.

179. Bian YS, Osterheld MC, Fontolliet C, Bosman FT, Benhattar J. p16 inactivation by methylation of the CDKN2A promoter occurs early during neoplastic progression in Barrett's esophagus. *Gastroenterology.* 2002;122:1113-1121.

180. Stachler MD, Camarda ND, Deitrick C, Kim A, Agoston AT, Odze RD, Hornick JL, Nag A, Thorner AR, Ducar M, Noffsinger A, Lash RH, Redston M, Carter SL, Davison JM, Bass AJ. Detection of Mutations in Barrett's Esophagus Before Progression to High-grade Dysplasia or Adenocarcinoma. *Gastroenterology.* 2018.

181. Lai LA, Kostadinov R, Barrett MT, Peiffer DA, Pokholok D, Odze R, Sanchez CA, Maley CC, Reid BJ, Gunderson KL, Rabinovitch PS. Deletion at fragile sites is a common and early event in Barrett's esophagus. *Mol Cancer Res.* 2010;8:1084-1094.

182. Kuroki T, Tajima Y, Furui J, Kanematsu T. Common fragile genes and digestive tract cancers. *Surgery today.* 2006;36:1-5.

183. Michael D, Beer DG, Wilke CW, Miller DE, Glover TW. Frequent deletions of FHIT and FRA3B in Barrett's metaplasia and esophageal adenocarcinomas. *Oncogene.* 1997;15:1653-1659.

184. Martinez P, Mallo D, Paulson TG, Li X, Sanchez CA, Reid BJ, Graham TA, Kuhner MK, Maley CC. Evolution of Barrett's esophagus through space and time at single-crypt and whole-biopsy levels. *Nat Commun.* 2018;9:794.

185. Li X, Galipeau PC, Paulson TG, Sanchez CA, Arnaudo J, Liu K, Sather CL, Kostadinov RL, Odze RD, Kuhner MK, Maley CC, Self SG, Vaughan TL, Blount PL, Reid BJ. Temporal and spatial evolution of somatic chromosomal alterations: a case-cohort study of Barrett's esophagus. *Cancer prevention research (Philadelphia, Pa).* 2014;7:114-127.

186. Rabinovitch PS, Reid BJ, Haggitt RC, Norwood TH, Rubin CE. Progression to cancer in Barrett's esophagus is associated with genomic instability. *Lab Invest.* 1989;60:65-71.

187. Nones K, Waddell N, Wayte N, Patch A-M, Bailey P, Newell F, Holmes O, Fink JL, Quinn MCJ, Tang YH, Lampe G, Quek K, Loffler KA, Manning S, Idrisoglu S, Miller D, Xu Q, Waddell N, Wilson PJ, Bruxner TJC, Christ AN, Harliwong I, Nourse C, Nourbakhsh E, Anderson M, Kazakoff S, Leonard C, Wood S, Simpson PT, Reid LE, Krause L, Hussey DJ, Watson DI, Lord RV, Nancarrow D, Phillips WA, Gotley D, Smithers BM, Whiteman DC, Hayward NK, Campbell PJ, Pearson JV, Grimmond SM, Barbour AP. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature communications.* 2014;5:1-9.

188. Killcoyne S, Gregson E, Wedge DC, Woodcock DJ, Eldridge MD, de la Rue R, Miremadi A, Abbas S, Blasko A, Kosmidou C, Januszewicz W, Jenkins AV, Gerstung M, Fitzgerald RC. Genomic copy number predicts esophageal cancer years before transformation. *Nature Medicine.* 2020;26:1726-1732.

189. Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, MacRae S, Grehan N, O'Donovan M, Miremadi A, Yang T-P, Bower L, Chettouh H, Crawte J, Galeano-Dalmau N, Grabowska A, Saunders J, Underwood T, Waddell N, Barbour AP, Nutzinger B, Achilleos A, Edwards PAW, Lynch AG, Tavaré S, Fitzgerald RC, Noorani A, Elliott RF, Weaver J, Ross-Innes C, Smith L, Abdullahi Z, de la Rue R, Cluroe A, Malhotra S, Hardwick R, Ford H, Smith ML, Davies J, Turkington R, Hayes SJ, Ang Y, Preston SR, Oakes S, Bagwan I, Save V, Skipworth RJE, Hupp TR, O'Neill JR, Tucker O, Taniere P, Noble F, Owsley J, Lovat L, Haidry R, Eneh V, Crichton C, Barr H, Shepherd N, Old O, Lagergren J, Gossage J, Davies A, Chang F, Zylstra J, Sanders G, Berrisford R, Harden C, Bunting D, Lewis M, Cheong E, Kumar B, Parsons SL, Soomro I, Kaye P, Collier P, Igali L, Welch I, Scott M, Sothi S, Suortamo S, Lishman S, Beardsmore D, Francies HE, Garnett MJ, Pearson JV, Nones K, Patch A-M, Grimmond SM. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature genetics.* 2016.

190. Lane DP. Cancer. p53, guardian of the genome. *Nature.* 1992;358:15-16.

191. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell.* 2011;144:646-674.

192. Muller PAJ, Vousden KH. p53 mutations in cancer. *Nature Cell Biology.* 2013;15:2-8.

193. Zilfou JT, Lowe SW. Tumor suppressive functions of p53. *Cold Spring Harb Perspect Biol.* 2009;1:a001883.

194. Paulson TG, Galipeau PC, Oman KM, Sanchez CA, Kuhner MK, Smith LP, Hadi K, Shah M, Arora K, Shelton J, Johnson M, Corvelo A, Maley CC, Yao X, Sanghvi R, Venturini E, Emde AK, Hubert B, Imielinski M, Robine N, Reid BJ, Li X. Somatic whole genome dynamics of precancer in Barrett's esophagus reveals features associated with disease progression. *Nat Commun.* 2022;13:2300.

195. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, Cagan A, Murai K, Mahbubani K, Stratton MR, Fitzgerald RC, Handford PA, Campbell PJ, Saeb-Parsy K, Jones PH. Somatic mutant clones colonize the human esophagus with age. *Science.* 2018;362:911-917.

196. Redston M, Noffsinger A, Kim A, Akarca FG, Rara M, Stapleton D, Nowden L, Lash R, Bass AJ, Stachler MD. Abnormal TP53 Predicts Risk of Progression in

Patients With Barrett's Esophagus Regardless of a Diagnosis of Dysplasia. *Gastroenterology.* 2022;162:468-481.

197. Januszewicz W, Pilonis ND, Sawas T, Phillips R, O'Donovan M, Miremadi A, Malhotra S, Tripathi M, Blasko A, Katzka DA, Fitzgerald RC, di Pietro M. The utility of P53 immunohistochemistry in the diagnosis of Barrett's oesophagus with indefinite for dysplasia. *Histopathology.* 2022;80:1081-1090.

198. Reid BJ, Prevo LJ, Galipeau PC, Sanchez CA, Longton G, Levine DS, Blount PL, Rabinovitch PS. Predictors of progression in Barrett's esophagus II: baseline 17p (p53) loss of heterozygosity identifies a patient subset at increased risk for neoplastic progression. *American Journal Of Gastroenterology.* 2001;96:2839.

199. Cortés-Ciriano I, Lee JJ, Xi R, Jain D, Jung YL, Yang L, Gordenin D, Klimczak LJ, Zhang CZ, Pellman DS, Park PJ. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet.* 2020;52:331-341.

200. Finley JC, Reid BJ, Odze RD, Sanchez CA, Galipeau P, Li X, Self SG, Gollahon KA, Blount PL, Rabinovitch PS. Chromosomal instability in Barrett's esophagus is related to telomere shortening. *Cancer Epidemiol Biomarkers Prev.* 2006;15:1451-1457.

201. Mathieu N, Pirzio L, Freulet-Marrière MA, Desmaze C, Sabatier L. Telomeres and chromosomal instability. *Cellular and molecular life sciences : CMLS.* 2004;61:641-656.

202. Network CGAR. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature.* 2014;513:202-209.

203. Eads C, Lord R, Kurumboor S, Wickramasinghe K, Skinner M, Long T, Peters J, DeMeester T, Danenberg K, Danenberg P, Laird P, Skinner K. Fields of aberrant CpG island hypermethylation in Barrett's esophagus and associated adenocarcinoma. Vol 60. 18 ed; 2000:5021-5026.

204. Clement G, Braunschweig R, Pasquier N, Bosman FT, Benhattar J. Alterations of the Wnt signaling pathway during the neoplastic progression of Barrett's esophagus. *Oncogene.* 2006;25:3084-3092.

205. Smith E, De Young NJ, Pavey SJ, Hayward NK, Nancarrow DJ, Whiteman DC, Smithers BM, Ruszkiewicz AR, Clouston AD, Gotley DC, Devitt PG, Jamieson GG, Drew PA. Similarity of aberrant DNA methylation in Barrett's esophagus and esophageal adenocarcinoma. *Mol Cancer.* 2008;7:75.

206. Wang B, Song H, Jiang H, Fu Y, Ding X, Zhou C. Early diagnostic potential of APC hypermethylation in esophageal cancer. *Cancer Manag Res.* 2018;10:181-198.

207. Wang JS, Guo M, Montgomery EA, Thompson RE, Cosby H, Hicks L, Wang S, Herman JG, Canto MI. DNA promoter hypermethylation of p16 and APC predicts neoplastic progression in Barrett's esophagus. *Am J Gastroenterol.* 2009;104:2153-2160.

208. Schulmann K, Sterian A, Berki A, Yin J, Sato F, Xu Y, Olaru A, Wang S, Mori Y, Deacu E, Hamilton J, Kan T, Krasna MJ, Beer DG, Pepe MS, Abraham JM, Feng Z, Schmiegel W, Greenwald BD, Meltzer SJ. Inactivation of p16, RUNX3, and HPP1 occurs early in Barrett's-associated neoplastic progression and predicts progression risk. *Oncogene.* 2005;24:4138-4148.

209. Pinto R, Hauge T, Jeanmougin M, Pharo HD, Kresse SH, Honne H, Winge SB, Five MB, Kumar T, Mala T, Hauge T, Johnson E, Lind GE. Targeted genetic and epigenetic profiling of esophageal adenocarcinomas and non-dysplastic Barrett's esophagus. *Clin Epigenetics.* 2022;14:77.

210. Xu E, Gu J, Hawk ET, Wang KK, Lai M, Huang M, Ajani J, Wu X. Genome-wide methylation analysis shows similar patterns in Barrett's esophagus and esophageal adenocarcinoma. *Carcinogenesis.* 2013;34:2750-2756.

211. Krause L, Nones K, Loffler KA, Nancarrow D, Oey H, Tang YH, Wayte NJ, Patch AM, Patel K, Brosda S, Manning S, Lampe G, Clouston A, Thomas J, Stoye J, Hussey DJ, Watson DI, Lord RV, Phillips WA, Gotley D, Smithers BM, Whiteman DC, Hayward NK, Grimmond SM, Waddell N, Barbour AP. Identification of the CIMP-like subtype and aberrant methylation of members of the chromosomal segregation and spindle assembly pathways in esophageal adenocarcinoma. *Carcinogenesis.* 2016;37:356-365.

212. Wu W, Bhagat TD, Yang X, Song JH, Cheng Y, Agarwal R, Abraham JM, Ibrahim S, Bartenstein M, Hussain Z, Suzuki M, Yu Y, Chen W, Eng C, Greally J, Verma A, Meltzer SJ. Hypomethylation of noncoding DNA regions and overexpression of the long noncoding RNA, AFAP1-AS1, in Barrett's esophagus and esophageal adenocarcinoma. *Gastroenterology.* 2013;144:956-966 e954.

213. Kaz AM, Wong CJ, Luo Y, Virgin JB, Washington MK, Willis JE, Leidner RS, Chak A, Grady WM. DNA methylation profiling in Barrett's esophagus and esophageal adenocarcinoma reveals unique methylation signatures and molecular subclasses. *Epigenetics.* 2011;6:1403-1412.

214. Kaz AM, Wong CJ, Varadan V, Willis JE, Chak A, Grady WM. Global DNA methylation patterns in Barrett's esophagus, dysplastic Barrett's, and

esophageal adenocarcinoma are associated with BMI, gender, and tobacco use. *Clin Epigenetics.* 2016;8:111.

215. Jammula S, Katz-Summercorn AC, Li X, Linossi C, Smyth E, Killcoyne S, Biasci D, Subash VV, Abbas S, Blasko A, Devonshire G, Grantham A, Wronowski F, O'Donovan M, Grehan N, Eldridge MD, Tavaré S, Fitzgerald RC. Identification of Subtypes of Barrett's Esophagus and Esophageal Adenocarcinoma Based on DNA Methylation Profiles and Integration of Transcriptome and Genome Data. *Gastroenterology.* 2020;158:1682-1697.e1681.

216. Eads CA, Lord RV, Wickramasinghe K, Long TI, Kurumboor SK, Bernstein L, Peters JH, DeMeester SR, DeMeester TR, Skinner KA, Laird PW. Epigenetic patterns in the progression of esophageal adenocarcinoma. *Cancer research.* 2001;61:3410-3418.

217. Clement G, Braunschweig R, Pasquier N, Bosman FT, Benhattar J. Methylation of APC, TIMP3, and TERT: a new predictive marker to distinguish Barrett's oesophagus patients at risk for malignant transformation. *J Pathol.* 2006;208:100-107.

218. Jin Z, Cheng Y, Gu W, Zheng Y, Sato F, Mori Y, Olaru AV, Paun BC, Yang J, Kan T, Ito T, Hamilton JP, Selaru FM, Agarwal R, David S, Abraham JM, Wolfsen HC, Wallace MB, Shaheen NJ, Washington K, Wang J, Canto MI, Bhattacharyya A, Nelson MA, Wagner PD, Romero Y, Wang KK, Feng Z, Sampliner RE, Meltzer SJ. A multicenter, double-blinded validation study of methylation biomarkers for progression prediction in Barrett's esophagus. *Cancer Res.* 2009;69:4112-4115.

219. Alvi MA, Liu X, O'Donovan M, Newton R, Wernisch L, Shannon NB, Shariff K, di Pietro M, Bergman JJ, Ragunath K, Fitzgerald RC. DNA methylation as an adjunct to histopathology to detect prevalent, inconspicuous dysplasia and early-stage neoplasia in Barrett's esophagus. *Clin Cancer Res.* 2013;19:878-888.

220. Iyer PG, Taylor WR, Johnson ML, Lansing RL, Maixner KA, Yab TC, Simonson JA, Devens ME, Slettedahl SW, Mahoney DW, Berger CK, Foote PH, Smyrk TC, Wang KK, Wolfsen HC, Ahlquist DA. Highly Discriminant Methylated DNA Markers for the Non-endoscopic Detection of Barrett's Esophagus. *Am J Gastroenterol.* 2018;113:1156-1166.

221. Chettouh H, Mowforth O, Galeano-Dalmau N, Bezawada N, Ross-Innes C, MacRae S, Debiram-Beecham I, O'Donovan M, Fitzgerald RC. Methylation panel is a diagnostic biomarker for Barrett's oesophagus in endoscopic biopsies and non-endoscopic cytology specimens. *Gut.* 2018;67:1942-1949.

222. Januszewicz W, Subhash VV, Waldock W, Fernando DI, Bartalucci G, Chettouh H, Miremadi A, O'Donovan M, Fitzgerald RC, di Pietro M. The utility of a methylation panel in the assessment of clinical response to radiofrequency ablation for Barrett's esophagus. *EBioMedicine.* 2020;58:102877.

223. Wright NA. Is Barrett's-Associated Esophageal Adenocarcinoma a Clonal Disease? *Dig Dis Sci.* 2018;63:2022-2027.

224. McDonald SAC, Greaves LC, Gutierrez-Gonzalez L, Rodriguez-Justo M, Deheragoda M, Leedham SJ, Taylor RW, Lee CY, Preston SL, Lovell M, Hunt T, Elia G, Oukrif D, Harrison R, Novelli MR, Mitchell I, Stoker DL, Turnbull DM, Jankowski JAZ, Wright NA. Mechanisms of field cancerization in the human stomach: the expansion and spread of mutated gastric stem cells. *Gastroenterology.* 2008;134:500-510.

225. Cummins AG, Catto-Smith AG, Cameron DJ, Couper RT, Davidson GP, Day AS, Hammond PD, Moore DJ, Thompson FM. Crypt fission peaks early during infancy and crypt hyperplasia broadly peaks during infancy and childhood in the small intestine of humans. *Journal of pediatric gastroenterology and nutrition.* 2008;47:153-157.

226. Cheng H, Bjerknes M, Amar J, Gardiner G. Crypt production in normal and diseased human colonic epithelium. *The Anatomical record.* 1986;216:44-48.

227. Galandiuk S, Rodriguez-Justo M, Jeffery R, Nicholson AM, Cheng Y, Oukrif D, Elia G, Leedham SJ, McDonald SA, Wright NA, Graham TA. Field cancerization in the intestinal epithelium of patients with Crohn's ileocolitis. *Gastroenterology.* 2012;142:855-864 e858.

228. Wasan HS, Park HS, Liu KC, Mandir NK, Winnett A, Sasieni P, Bodmer WF, Goodlad RA, Wright NA. APC in the regulation of intestinal crypt fission. *The Journal of pathology.* 1998;185:246-255.

229. Baker AM, Gabbutt C, Williams MJ, Cereser B, Jawad N, Rodriguez-Justo M, Jansen M, Barnes CP, Simons BD, McDonald SA, Graham TA, Wright NA. Crypt fusion as a homeostatic mechanism in the human colon. *Gut.* 2019;68:1986-1993.

230. Blount PL, Meltzer SJ, Yin J, Huang Y, Krasna MJ, Reid BJ. Clonal ordering of 17p and 5q allelic losses in Barrett dysplasia and adenocarcinoma. *Proceedings of the National Academy of Sciences of the United States of America.* 1993;90:3221-3225.

231. Maley CC, Galipeau PC, Li X, Sanchez CA, Paulson TG, Reid BJ. Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. *Cancer Res.* 2004;64:3414-3427.

232. Maley CC. Multistage carcinogenesis in Barrett's esophagus. *Cancer Lett.* 2007;245:22-32.

233. Maley CC, Reid BJ. Natural selection in neoplastic progression of Barrett's esophagus. *Semin Cancer Biol.* 2005;15:474-483.

234. Kostadinov RL, Kuhner MK, Li X, Sanchez CA, Galipeau PC, Paulson TG, Sather CL, Srivastava A, Odze RD, Blount PL, Vaughan TL, Reid BJ, Maley CC. NSAIDs modulate clonal evolution in Barrett's esophagus. *PLoS Genet.* 2013;9:e1003553.

235. Nowell PC. The clonal evolution of tumor cell populations. *Science (New York, NY).* 1976;194:23-28.

236. Smith LP, Yamato JA, Galipeau PC, Paulson TG, Li X, Sanchez CA, Reid BJ, Kuhner MK. Within-patient phylogenetic reconstruction reveals early events in Barrett's Esophagus. *Evol Appl.* 2021;14:399-415.

237. Martinez P, Timmer MR, Lau CT, Calpe S, Sancho-Serra Mdel C, Straub D, Baker AM, Meijer SL, Kate FJ, Mallant-Hent RC, Naber AH, van Oijen AH, Baak LC, Scholten P, Bohmer CJ, Fockens P, Bergman JJ, Maley CC, Graham TA, Krishnadath KK. Dynamic clonal equilibrium and predetermined cancer risk in Barrett's oesophagus. *Nat Commun.* 2016;7:12158.

238. Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, Paulson TG, Blount PL, Risques RA, Rabinovitch PS, Reid BJ. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature genetics.* 2006;38:468-473.

239. Maley CC, Galipeau PC, Li X, Sanchez CA, Paulson TG, Blount PL, Reid BJ. The combination of genetic instability and clonal expansion predicts progression to esophageal adenocarcinoma. *Cancer research.* 2004;64:7629-7633.

240. Magurran AE. Measuring Biological Diversity; 2004.

241. Merlo LMF, Merlo LMF, Pepper JW, Pepper JW, Reid BJ, Reid BJ, Maley CC, Maley CC. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer.* 2006;6:924-935.

242. Zeki SS, McDonald SA, Graham TA. Field cancerization in Barrett's esophagus. *Discovery medicine.* 2011;12:371-379.

243. Katz-Summercorn AC, Jammula S, Frangou A, Peneva I, O'Donovan M, Tripathi M, Malhotra S, di Pietro M, Abbas S, Devonshire G, Januszewicz W, Blasko A, Nowicki-Osuch K, MacRae S, Northrop A, Redmond AM, Wedge DC, Fitzgerald

RC. Multi-omic cross-sectional cohort study of pre-malignant Barrett's esophagus reveals early structural variation and retrotransposon activity. *Nat Commun.* 2022;13:1407.

244. Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer.* 1953;6:963-968.

245. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nat Rev Cancer.* 2004;4:143-153.

246. Meaburn E, Schulz R. Next generation sequencing in epigenetics: insights and challenges. *Semin Cell Dev Biol.* 2012;23:192-199.

247. Feinberg AP. The epigenetics of cancer etiology. *Semin Cancer Biol.* 2004;14:427-432.

248. Bohn M, Heermann DW. Diffusion-driven looping provides a consistent framework for chromatin organization. *PLoS One.* 2010;5:e12218.

249. Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. *Science (New York, NY).* 2001;293:1068-1070.

250. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. *Journal of Molecular Biology.* 1987;196:261-282.

251. Illingworth RS, Bird AP. CpG islands--'a rough guide'. *FEBS Lett.* 2009;583:1713-1720.

252. Law J, Jacobsen S. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews Genetics.* 2010;11:204-220.

253. Bird A. DNA methylation patterns and epigenetic memory. *Genes & development.* 2002;16:6-21.

254. Yatabe Y, Tavaré S, Shibata D. Investigating stem cells in human colon by using methylation patterns. *Proceedings of the National Academy of Sciences of the United States of America.* 2001;98:10839-10844.

255. Issa J-P. Aging and epigenetic drift: a vicious cycle. *The Journal of clinical investigation.* 2014;124:24-29.

256. Rose JA, Yates PA, Simpson J, Tischfield JA, Stambrook PJ, Turker MS. Biallelic Methylation and Silencing of Mouse *Aprt* in Normal Kidney Cells. *Cancer Research.* 2000;60:3404.

257. Wigler M, Levy D, Perucho M. The somatic replication of DNA methylation. *Cell.* 1981;24:33-40.

258. Riggs AD, Xiong Z, Wang L, LeBon JM. Methylation dynamics, epigenetic fidelity and X chromosome structure. *Novartis Foundation symposium.* 1998;214:214-225; discussion 225-232.

259. Riggs AD, Xiong Z. Methylation and epigenetic fidelity. *Proc Natl Acad Sci U S A.* 2004;101:4-5.

260. Sontag LB, Lorincz MC, Georg Luebeck E. Dynamics, stability and inheritance of somatic DNA methylation imprints. *Journal of theoretical biology.* 2006;242:890-899.

261. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. *Nat Commun.* 2017;8:15183.

262. Werner B, Case J, Williams MJ, Chkhaidze K, Temko D, Fernández-Mateos J, Cresswell GD, Nichol D, Cross W, Spiteri I, Huang W, Tomlinson IPM, Barnes CP, Graham TA, Sottoriva A. Measuring single cell divisions in human tissues from multi-region sequencing data. *Nat Commun.* 2020;11:1035.

263. Sanz LA, Kota SK, Feil R. Genome-wide DNA demethylation in mammals. *Genome Biol.* 2010;11:110.

264. Cantone I, Fisher AG. Epigenetic programming and reprogramming during development. *Nature structural & molecular biology.* 2013;20:282-289.

265. Avner P, Heard E. X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet.* 2001;2:59-67.

266. Li Y, Tollefsbol TO. Age-related epigenetic drift and phenotypic plasticity loss: implications in prevention of age-related human diseases. *Epigenomics.* 2016;8:1637-1651.

267. Shibata D. Inferring human stem cell behaviour from epigenetic drift. *The Journal of pathology.* 2009;217:199-205.

268. Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. *Aging Cell.* 2015;14:924-932.

269. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013;14:R115.

270. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan J-B, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell.* 2013;49:359-367.

271. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell.* 2013;153:1194-1217.

272. Luebeck GE, Hazelton WD, Curtius K, Maden SK, Yu M, Carter KT, Burke W, Lampe PD, Li CI, Ulrich CM, Newcomb PA, Westerhoff M, Kaz AM, Luo Y, Inadomi JM, Grady WM. Implications of Epigenetic Drift in Colorectal Neoplasia. *Cancer Res.* 2019;79:495-504.

273. Shibata D. Mutation and epigenetic molecular clocks in cancer. *Carcinogenesis.* 2011;32:123-128.

274. Shibata D, Tavare S. Counting divisions in a human somatic cell tree: how, what and why? *Cell Cycle.* 2006;5:610-614.

275. Kim K-M, Shibata D. Tracing ancestry with methylation patterns: most crypts appear distantly related in normal adult human colon. *BMC gastroenterology.* 2004;4:8.

276. Humphries A, Cereser B, Gay LJ, Miller DSJ, Das B, Gutteridge A, Elia G, Nye E, Jeffery R, Poulsom R, Novelli MR, Rodriguez-Justo M, McDonald SAC, Wright NA, Graham TA. Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. *Proceedings of the National Academy of Sciences of the United States of America.* 2013;110:E2490-2499.

277. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? *Hum Mol Genet.* 2013;22:R7-R15.

278. Graham TA, Humphries A, Sanders T, Rodriguez-Justo M, Tadrous PJ, Preston SL, Novelli MR, Leedham SJ, McDonald SAC, Wright NA. Use of methylation patterns to determine expansion of stem cell clones in human colon tissue. *Gastroenterology.* 2011;140:1241-1250.e1241-1249.

279. Siegmund KD, Marjoram P, Tavare S, Shibata D. Many colorectal cancers are "flat" clonal expansions. *Cell Cycle.* 2009;8:2187-2193.

280. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, Marjoram P, Siegmund K, Press MF, Shibata D, Curtis C. A Big Bang model of human colorectal tumor growth. *Nature genetics.* 2015.

281. Cross W, Kovac M, Mustonen V, Temko D, Davis H, Baker AM, Biswas S, Arnold R, Chegwidden L, Gatenbee C, Anderson AR, Koelzer VH, Martinez P, Jiang X, Domingo E, Woodcock DJ, Feng Y, Kovacova M, Maughan T, Consortium SC, Jansen M, Rodriguez-Justo M, Ashraf S, Guy R, Cunningham C, East JE, Wedge DC, Wang LM, Palles C, Heinimann K, Sottoriva A, Leedham SJ, Graham TA, Tomlinson IPM. The evolutionary landscape of colorectal tumorigenesis. *Nat Ecol Evol.* 2018;2:1661-1672.

282. Siegmund K, Marjoram P, Woo Y, Tavaré S, Shibata D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America.* 2009.

283. Shibata D. Visualizing Human Colorectal Cancer Intratumor Heterogeneity with Phylogeography. *iScience.* 2020;23:101304.

284. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell.* 1990;61:759-767.

285. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL. Genetic alterations during colorectal-tumor development. *N Engl J Med.* 1988;319:525-532.

286. Adachi, Yasuda, Kakisako, Sato, Shiraishi, Kitano. Histopathologic characteristics of advanced colorectal cancer smaller than 2 cm in size. *Colorectal Dis.* 1999;1:19-22.

287. Woo Y-J, Siegmund KD, Tavaré S, Shibata D. Older individuals appear to acquire mitotically older colorectal cancers. *The Journal of pathology.* 2009;217:483-488.

288. Leedham SJ, Wright NA. Expansion of a mutated clone: from stem cell to tumour. *Journal of clinical pathology.* 2008;61:164-171.

289. Schofield R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood cells.* 1978;4:7-25.

290. Spradling A, Drummond-Barbosa D, Kai T. Stem cells find their niche. *Nature.* 2001;414:98-104.

291. Lopez-Garcia C, Klein AM, Simons BD, Winton DJ. Intestinal stem cell replacement follows a pattern of neutral drift. *Science (New York, NY).* 2010;330:822-825.

292. Shibata D. Clonal diversity in tumor progression. *Nat Genet.* 2006;38:402-403.

293. Kim K-M, Shibata D. Methylation reveals a niche: stem cell succession in human colon crypts. *Oncogene.* 2002;21:5441-5449.

294. Calabrese P, Calabrese P, Tavaré S, Tavaré S, Shibata D, Shibata D. Pretumor progression: clonal evolution of human stem cell populations. *The American journal of pathology.* 2004;164:1337-1346.

295. Kim K-M, Calabrese P, Tavaré S, Shibata D. Enhanced stem cell survival in familial adenomatous polyposis. *The American journal of pathology.* 2004;164:1369-1377.

296. Gabbutt C, Schenck RO, Weisenberger DJ, Kimberley C, Berner A, Househam J, Lakatos E, Robertson-Tessi M, Martin I, Patel R, Clark SK, Latchford A, Barnes CP, Leedham SJ, Anderson ARA, Graham TA, Shibata D. Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues. *Nat Biotechnol.* 2022;40:720-730.

297. Curtius K, Wong CJ, Hazelton WD, Kaz AM, Chak A, Willis JE, Grady WM, Luebeck EG. A Molecular Clock Infers Heterogeneous Tissue Age Among Patients with Barrett's Esophagus. *PLoS Comput Biol.* 2016;12:e1004919.

298. Killcoyne S, Fitzgerald RC. Evolution and progression of Barrett's oesophagus to oesophageal cancer. *Nat Rev Cancer.* 2021;21:731-741.

299. Walther V, Alison MR. Cell lineage tracing in human epithelial tissues using mitochondrial DNA mutations as clonal markers. *Wiley Interdisciplinary Reviews: Developmental Biology.* 2015:n/a-n/a.

300. Gabbutt C, Wright NA, Baker AM, Shibata D, Graham TA. Lineage tracing in human tissues. *J Pathol.* 2022;257:501-512.

301. Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. *Nature reviews Genetics.* 2005;6:389-402.

302. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. Sequence and organization of the human mitochondrial genome. *Nature.* 1981;290:457-465.

303. Taylor RW, Barron MJ, Borthwick GM, Gospel A, Chinnery PF, Samuels DC, Taylor GA, Plusa SM, Needham SJ, Greaves LC, Kirkwood TBL, Turnbull DM. Mitochondrial DNA mutations in human colonic crypt stem cells. *Journal of Clinical Investigation.* 2003;112:1351-1360.

304. Zeki S, Graham TA, McDonald SA. Utilizing DNA mutations to trace epithelial cell lineages in human tissues. *Methods Mol Biol.* 2012;916:289-301.

305. Evans JA, Carlotti E, Lin ML, Hackett RJ, Haughey MJ, Passman AM, Dunn L, Elia G, Porter RJ, McLean MH, Hughes F, ChinAleong J, Woodland P, Preston SL, Griffin SM, Lovat L, Rodriguez-Justo M, Huang W, Wright NA, Jansen M, McDonald SAC. Clonal Transitions and Phenotypic Evolution in Barrett's Esophagus. *Gastroenterology.* 2022;162:1197-1209.e1113.

306. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics.* 1999;23:147.

307. The Genotype-Tissue Expression (GTEx) Project. **https://gtexportal.org/home/**. Accessed October 2017.

308. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, Guan P, Korzeniewski GE, Lockhart NC, Rabiner CA, Rao AK, Robinson KL, Roche NV, Sawyer SJ, Segre AV, Shive CE, Smith AM, Sobin LH, Undale AH, Valentino KM, Vaught J, Young TR, Moore HM, Consortium GT. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank.* 2015;13:311-319.

309. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics.* 2010;26:493-500.

310. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 2019;47:W636-W641.

311. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Ponten F. Proteomics. Tissue-based map of the human proteome. *Science.* 2015;347:1260419.

312. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan-Golan Y, Kohn A, Rappaport N, Safran M, Lancet D. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics.* 2016;54:1 30 31-31 30 33.

313. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman

V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. Ensembl 2018. *Nucleic Acids Res.* 2018;46:D754-D761.

314. Kuo HC, Lin PY, Chung TC, Chao CM, Lai LC, Tsai MH, Chuang EY. DBCAT: database of CpG islands and analytical tools for identifying comprehensive methylation profiles in cancer cells. *J Comput Biol.* 2011;18:1013-1017.

315. Bisulfite Primer Seeker 12S; **www.zymoresearch.com/pages/bisulfite-primer-seeker**; Accessed October 2017.

316. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3--new capabilities and interfaces. *Nucleic Acids Res.* 2012;40:e115.

317. Li LC, Dahiya R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics.* 2002;18:1427-1431.

318. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics.* 2012;13:134.

319. Aranyi T, Varadi A, Simon I, Tusnady GE. The BiSearch web server. *BMC Bioinformatics.* 2006;7:431.

320. Illumina. Whole-genome Bisulfite Sequencing for Methylation Analysis Preparing Samples for the Illumina Sequencing Platform. Protocol available from **https://support.illumina.com**.

321. Nair SS, Luu PL, Qu W, Maddugoda M, Huschtscha L, Reddel R, Chenevix-Trench G, Toso M, Kench JG, Horvath LG, Hayes VM, Stricker PD, Hughes TP, White DL, Rasko JEJ, Wong JJ, Clark SJ. Guidelines for whole genome bisulphite sequencing of intact and FFPET DNA on the Illumina HiSeq X Ten. *Epigenetics Chromatin.* 2018;11:24.

322. Marzese DM, Hoon DS. Emerging technologies for studying DNA methylation for the molecular diagnosis of cancer. *Expert Rev Mol Diagn.* 2015;15:647-664.

323. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. *Genome Biol.* 2013;14:R51.

324. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, Reik W. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.* 2018;19:33.

325. Stahlberg A, Krzyzanowski PM, Jackson JB, Egyud M, Stein L, Godfrey TE. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic Acids Res.* 2016;44:e105.

326. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America.* 2011;108:9530-9535.

327. Sena JA, Galotto G, Devitt NP, Connick MC, Jacobi JL, Umale PE, Vidali L, Bell CJ. Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Sci Rep.* 2018;8:13121.

328. Shagin DA, Turchaninova MA, Shagina IA, Shugay M, Zaretsky AR, Zueva OI, Bolotin DA, Lukyanov S, Chudakov DM. Application of nonsense-mediated primer exclusion (NOPE) for preparation of unique molecular barcoded libraries. *BMC Genomics.* 2017;18:440.

329. Peng Q, Ecker JR. Detection of allele-specific methylation through a generalized heterogeneous epigenome model. *Bioinformatics.* 2012;28:i163-i171.

330. Fang F, Hodges E, Molaro A, Dean M, Hannon GJ, Smith AD. Genomic landscape of human allele-specific DNA methylation. *Proceedings of the National Academy of Sciences.* 2012;109:7332.

331. Herman JG, Graff JR, Myöhänen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences of the United States of America.* 1996;93:9821-9826.

332. Wojdacz TK, Hansen LL. Reversal of PCR bias for improved sensitivity of the DNA methylation melting curve assay. *Biotechniques.* 2006;41:274, 276, 278.

333. Neumann HP. Progress in DNA methylation research. New York: Nova Biomedical Books; 2007.

334. Yang Y, Sebra R, Pullman BS, Qiao W, Peter I, Desnick RJ, Geyer CR, DeCoteau JF, Scott SA. Quantitative and multiplexed DNA methylation analysis using

long-read single-molecule real-time bisulfite sequencing (SMRT-BS). *BMC Genomics.* 2015;16:350.

335. Fogg MJ, Pearl LH, Connolly BA. Structural basis for uracil recognition by archaeal family B DNA polymerases. *Nature structural biology.* 2002;9:922-927.

336. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, Quail MA. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics.* 2012;13:1.

337. Kou R, Lam H, Duan H, Ye L, Jongkam N, Chen W, Zhang S, Li S. Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLOS ONE.* 2016;11:e0146638.

338. Stahlberg A, Krzyzanowski PM, Egyud M, Filges S, Stein L, Godfrey TE. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. *Nat Protoc.* 2017;12:664-682.

339. Foley JW, Zhu C, Jolivet P, Zhu SX, Lu P, Meaney MJ, West RB. Gene-expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. 2019.

340. Bhang H-eC, Ruddy DA, Krishnamurthy Radhakrishna V, Caushi JX, Zhao R, Hims MM, Singh AP, Kao I, Rakiec D, Shaw P, Balak M, Raza A, Ackley E, Keen N, Schlabach MR, Palmer M, Leary RJ, Chiang DY, Sellers WR, Michor F, Cooke VG, Korn JM, Stegmeier F. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature Medicine.* 2015;21:440-448.

341. Zhao L, Liu Z, Levy SF, Wu S. Bartender: a fast and accurate clustering algorithm to count barcode reads. *Bioinformatics.* 2018;34:739-747.

342. Ji L, Sasaki T, Sun X, Ma P, Lewis ZA, Schmitz RJ. Methylated DNA is over-represented in whole-genome bisulfite sequencing data. *Front Genet.* 2014;5:341.

343. Vaisvila R, Ponnaluri VKC, Sun Z, Langhorst BW, Saleh L, Guan S, Dai N, Campbell MA, Sexton BS, Marks K, Samaranayake M, Samuelson JC, Church HE, Tamanaha E, Corrêa IR, Pradhan S, Dimalanta ET, Evans TC, Williams L, Davis TB. EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of DNA. 2020.

344. Butcher S KT, Zalewski L. Apocrita - High Performance Computing Cluster for Queen Mary University of London. 2017.

345. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at: **http://www.bioinformatics.babraham.ac.uk/projects/fastqc**. 2010.

346. Vander Heiden JA* YG, Uduman M, Stern JNH, O'Connor KC, Hafler DA, Vigneault F, Kleinstein SH. . pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. . *Bioinformatics.* 2014.

347. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* 2017;27:491-499.

348. Tsagiopoulou M, Maniou MC, Pechlivanis N, Togkousidis A, Kotrová M, Hutzenlaub T, Kappas I, Chatzidimitriou A, Psomopoulos F. UMIc: A Preprocessing Method for UMI Deduplication and Reads Correction. *Front Genet.* 2021;12:660366.

349. FgBio Packages available from **http://fulcrumgenomics.github.io/fgbio/**.

350. Zorita E, Cusco P, Filion GJ. Starcode: sequence clustering based on all-pairs search. *Bioinformatics.* 2015;31:1913-1919.

351. Rauluseviciute I, Drabløs F, Rye MB. DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. *Clin Epigenetics.* 2019;11:193.

352. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics.* 2011;27:1571-1572.

353. Kint S, De Spiegelaere W, De Kesel J, Vandekerckhove L, Van Criekinge W. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. *PLoS One.* 2018;13:e0199091.

354. Warnecke PM, Stirzaker C, Melki JR, Millar DS, Paul CL, Clark SJ. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res.* 1997;25:4422-4426.

355. Moskalev EA, Zavgorodnij MG, Majorova SP, Vorobjev IA, Jandaghi P, Bure IV, Hoheisel JD. Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. *Nucleic Acids Res.* 2011;39:e77.

356. Schmidt M, Hackett RJ, Baker A-M, McDonald SAC, Quante M, Graham TA. Evolutionary dynamics in Barrett oesophagus: implications for surveillance,

risk stratification and therapy. *Nature Reviews Gastroenterology & Hepatology.* 2022;19:95-111.

357. Sottoriva A, Barnes CP, Graham TA. Catch my drift? Making sense of genomic intra-tumour heterogeneity. *Biochim Biophys Acta.* 2017;1867:95-100.

358. Kim JY, Tavare S, Shibata D. Human hair genealogies and stem cell latency. *BMC Biol.* 2006;4:2.

359. McDonald SAC, Lavery D, Wright NA, Jansen M. Barrett oesophagus: lessons on its origins from the lesion itself. *Nature Publishing Group.* 2014:1-11.

360. Reid BJ, Paulson TG, Li X. Genetic Insights in Barrett's Esophagus and Esophageal Adenocarcinoma. *Gastroenterology.* 2015:1-28.

10 Appendix

10.1 Supplementary tables and figures

Name	First round primer sequences 5’ to 3’	Position (mtDNA genome)
A (F)	GCTCACATCACCCCATAAAC	627-646
A (R)	GATTACTCCGGTCTGAACTC	3087-3068
B (F)	ACCAACAAGTCATTATTACCC	2395-2415
B (R)	TGAGGAAATACTTGATGGCAG	4653-4633
C (F)	CCGTCATCTACTCTACCATC	4489-4508
C (R)	GGACGGATCAGACGAAGAG	6468-6450
D (F)	AATACCCATCATAATCGGAGG	6113-6133
D (R)	GGTGATGAGGAATAGTGTAAG	8437-8417
E (F)	AACCACTTTCACCGCTACAC	8128-8147
E (R)	AGTGAGATGGTAAATGCTAG	10516-10487
F (F)	ACTTCACGTCATTATTGGCTC	9821-9841
F (R)	ATAGGAGGAGAATGGGGGATAG	12101-12080
G (F)	ACCCCCCACTATTAACCTACTG	11866-11887
G (R)	GGTAGAATCCGAGTATGTTGG	13924-13904
H (F)	TATTCGCAGGATTTCTCATTAC	13721-13742
H (R)	AGCTTTGGGTGCTAATGGTG	15997-15978
I (F)	CCCATCCTCCATATATCCAAAC	15659-15680
I (R)	GGTTAGTATAGCTTAGTTAAAC	868-847

***Table S1**: Sequencing primers for nested mitochondrial DNA polymerase chain reaction.*

Gene	Target Sequence Forward	Target Sequence Reverse	Introns Spanned	RNA (cDNA) Product Size (bp)	Genomic DNA Product Size (bp)
NKX2-5	AAGTGTGCGTCTGCCTTTCC	CTGCGTGGACGTGAGTTTCA	Intron 1	303	1843
MYOD1	CAATCCAAACCAGCGGTTGC	ACTTCAGTTCTCCCGCCTCT	Introns 1 & 2	691	1454
TNNI3	CTGCAGATTGCAAAGCAAGAG	TCCGTGATGTTCTTGGTGACTT	Intron 5	212	1582
CSRP3	AGGAAGGCTCTTGACAGCAC	TCGGACTCTCCAAACTTCGC	Intron 4	233	2020
NPPB	TCAGCCTCGGACTTGGAAAC	AGGGTTGAGGAAAAAGCCCC	Introns 1 & 2	367	1141
MYO18B	AGAAAGGCCTCGGATACGGA	TGGACATGCTCCTCATCCAC	Introns 4 & 5	264	1640
PXDNL	TAAACAAGCTGGAGGCACGC	TTCTCTGGGGAATCACTTGGC	Intron 22	238	999
CAMK2B	GCAGACTTCGGCCTAGCTATC	ACACTCCACAGTCTCCTGTCT	Introns 7, 8, 9 & 10	405	1744
ANKRD2	GGGGCTGACATGATGACCAA	ATTCTTAGGACCCTCCGGCT	Intron 8	242	1115
SCN5A	ACCGAGGAGAAGGAAAAGCG	CAGTGATGTGTGGTGGCTCT	Introns 10 & 11	405	2092
SPTB	ATGAGATTCTGGGCCATACGC	TTGATGTTCCGGCCGTAGTC	Intron 23	349	946
TNNT2	GAGGAGGAGCTCGTTTCTCT	ATGTAACCCCCAAAATGCATCA	Introns 9 & 10	215	1921
MYLK3	AGGAGTGACGACAATGACCAC	ACCTCGTAACCCGCAGAGA	Intron 3	251	1760
SBK2	CGTGGCTTCCTGTACGAGTT	TTCAGGTCCCGGTACACCA	Intron 3	254	1073
SBK3	GGGACCAGTACCACCTCATC	TCTCAGGACCAAATCCCGAC	Intron 2	125	621

***Table S2**: Primer designs for reverse transcription polymerase chain reactions (RT-PCR). Primers were designed to span at least one intronic sequence thereby making them capable of amplifying both a copy DNA (cDNA) and genomic DNA (gDNA) template. Individual primer sequences, introns spanned and expected product sizes are detailed above.*

Target Name	Primer Set	Amplicon Size (bp)	CpGs in Target	MP #12	MP Pool #1	MP Pool #2	MP Pool #3	MP Pool #4	MP Pool #5	MP Pool #6
SCN5A	1	573	72							
CAMK2B	2	564	78							
1-NKX2-5	1	563	42							
NPPB	1	550	45							
1-TNNI3	1	540	26							
LOC-L	Long	532	28							
2-MYOD1	2	516	59							
SBK2	2	489	24							
BGN-L	Long	489	24							
PXDNL	2	487	16							
MYO18B	2	475	25							
2-NKX2-5	2	470	47							
2-SBK3	2	462	19							
1-SBK3	2	461	27							
CSRP3	2	451	25							
2-TNNI3	2	435	25							
ANKRD2	2	426	29							
TNNT2	2	425	11							
1-MYOD1	2	367	40							
Total CpG in MP Pool ->			662	460	461	253	504	532	583	662
Number of Targets in MP Pool ->			16	10	10	5	12	13	15	19

***Table S3**: Range of multiplex pools trialled during optimisation of multiplex allele specific methylation sequencing protocol. Green fill signifies presence of primer set within the stated multiplex pool.*

Illumina® Compatible P5 3rd Round Primers			
Primer Name	**Bases forming Index (i5)**	**i5 bases for Sample Sheet**	**Primer sequence**
501	CTCTCTAT	CTCTCTAT	AATGATACGGCGACCACCGAGATCTACAC**CTCTCTAT**ACACTCTTTCCCTACACGACG CTCTTCCGAT*C*T
502	TATCCTCT	TATCCTCT	AATGATACGGCGACCACCGAGATCTACAC**TATCCTCT**ACACTCTTTCCCTACACGACG CTCTTCCGAT*C*T
503	GTAAGGAG	GTAAGGAG	AATGATACGGCGACCACCGAGATCTACAC**GTAAGGAG**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
504	ACTGCATA	ACTGCATA	AATGATACGGCGACCACCGAGATCTACAC**ACTGCATA**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
505	AAGGAGTA	AAGGAGTA	AATGATACGGCGACCACCGAGATCTACAC**AAGGAGTA**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
506	CTAAGCCT	CTAAGCCT	AATGATACGGCGACCACCGAGATCTACAC**CTAAGCCT**ACACTCTTTCCCTACACGACG CTCTTCCGAT*C*T
507	CGTCTAAT	CGTCTAAT	AATGATACGGCGACCACCGAGATCTACAC**CGTCTAAT**ACACTCTTTCCCTACACGACG CTCTTCCGAT*C*T
508	TCTCTCCG	TCTCTCCG	AATGATACGGCGACCACCGAGATCTACAC**TCTCTCCG**ACACTCTTTCCCTACACGACG CTCTTCCGAT*C*T
509	TCGACTAG	TCGACTAG	AATGATACGGCGACCACCGAGATCTACAC**TCGACTAG**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
510	TTCTAGCT	TTCTAGCT	AATGATACGGCGACCACCGAGATCTACAC**TTCTAGCT**ACACTCTTTCCCTACACGACG CTCTTCCGAT*C*T
511	CCTAGAGT	CCTAGAGT	AATGATACGGCGACCACCGAGATCTACAC**CCTAGAGT**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
512	GCGTAAGA	GCGTAAGA	AATGATACGGCGACCACCGAGATCTACAC**GCGTAAGA**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
513	CTATTAAG	CTATTAAG	AATGATACGGCGACCACCGAGATCTACAC**CTATTAAG**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
514	AAGGCTAT	AAGGCTAT	AATGATACGGCGACCACCGAGATCTACAC**AAGGCTAT**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
515	GAGCCTTA	GAGCCTTA	AATGATACGGCGACCACCGAGATCTACAC**GAGCCTTA**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T
516	TTATGCGA	TTATGCGA	AATGATACGGCGACCACCGAGATCTACAC**TTATGCGA**ACACTCTTTCCCTACACGAC GCTCTTCCGAT*C*T

Illumina Compatible P7 3rd Round Primers			
Primer Name	**Bases forming Index (i7)**	**i7 bases for Sample Sheet**	**Primer Sequence**
701	TCGCCTTA	TAAGGCGA	CAAGCAGAAGACGGCATACGAGAT**TCGCCTTA**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
702	CTAGTACG	CGTACTAG	CAAGCAGAAGACGGCATACGAGAT**CTAGTACG**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
703	TTCTGCCT	AGGCAGAA	CAAGCAGAAGACGGCATACGAGAT**TTCTGCCT**GTGACTGGAGTTCAGACGTGTGCTC TTCCGAT*C*T
704	GCTCAGGA	TCCTGAGC	CAAGCAGAAGACGGCATACGAGAT**GCTCAGGA**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
705	AGGAGTCC	GGACTCCT	CAAGCAGAAGACGGCATACGAGAT**AGGAGTCC**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
706	CATGCCTA	TAGGCATG	CAAGCAGAAGACGGCATACGAGAT**CATGCCTA**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
707	GTAGAGAG	CTCTCTAC	CAAGCAGAAGACGGCATACGAGAT**GTAGAGAG**GTGACTGGAGTTCAGACGTGTGC TCTTCCGAT*C*T
708	CAGCCTCG	CGAGGCTG	CAAGCAGAAGACGGCATACGAGAT**CAGCCTCG**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
709	TGCCTCTT	AAGAGGCA	CAAGCAGAAGACGGCATACGAGAT**TGCCTCTT**GTGACTGGAGTTCAGACGTGTGCTC TTCCGAT*C*T
710	TCCTCTAC	GTAGAGGA	CAAGCAGAAGACGGCATACGAGAT**TCCTCTAC**GTGACTGGAGTTCAGACGTGTGCTC TTCCGAT*C*T
711	TCATGAGC	GCTCATGA	CAAGCAGAAGACGGCATACGAGAT**TCATGAGC**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
712	CCTGAGAT	ATCTCAGG	CAAGCAGAAGACGGCATACGAGAT**CCTGAGAT**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T

713	TAGCGAGT	ACTCGCTA	CAAGCAGAAGACGGCATACGAGAT**TAGCGAGT**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
714	GTAGCTCC	GGAGCTAC	CAAGCAGAAGACGGCATACGAGAT**GTAGCTCC**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
715	TACTACGC	GCGTAGTA	CAAGCAGAAGACGGCATACGAGAT**TACTACGC**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
716	AGGCTCCG	CGGAGCCT	CAAGCAGAAGACGGCATACGAGAT**AGGCTCCG**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
717	GCAGCGTA	TACGCTGC	CAAGCAGAAGACGGCATACGAGAT**GCAGCGTA**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
718	CTGCGCAT	ATGCGCAG	CAAGCAGAAGACGGCATACGAGAT**CTGCGCAT**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
719	GAGCGCTA	TAGCGCTC	CAAGCAGAAGACGGCATACGAGAT**GAGCGCTA**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
720	CGCTCAGT	ACTGAGCG	CAAGCAGAAGACGGCATACGAGAT**CGCTCAGT**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
721	GTCTTAGG	CCTAAGAC	CAAGCAGAAGACGGCATACGAGAT**GTCTTAGG**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
722	ACTGATCG	CGATCAGT	CAAGCAGAAGACGGCATACGAGAT**ACTGATCG**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
723	TAGCTGCA	TGCAGCTA	CAAGCAGAAGACGGCATACGAGAT**TAGCTGCA**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T
724	GACGTCGA	TCGACGTC	CAAGCAGAAGACGGCATACGAGAT**GACGTCGA**GTGACTGGAGTTCAGACGTGTGCT CTTCCGAT*C*T

***Table S4**: Library preparation primers Used in the allele specific methylation sequencing protocol. Primers are designed to be compatible with the Illumina® system flow-cell by virtue of incorporating the P5 or P7 sequence respectively. There are sixteen 500 series primers and twenty-four 700 series primers, the indices and details for the demultiplexing sample sheet are detailed above.*

Stage	Component	Actions Taken	Outcome
Set-up	Primer Design	a) Tested both shorter and longer amplicons for same and different target genes	Length of target amplicon has not significantly affected efficacy of the protocol, although an arbitrary limit of 600 bp has been applied for next generation sequencing purposes and bisulfite induced fragmentation of DNA.
		b) Length of universal sequence (US) overhang in 1st round primers	Shorter overhang length improves target enrichment and reduces non-specific product and primer dimerisation
		c) Length of unique molecular identifiers (UID)	UID length has ranged from 10-21 bp, this has not caused a noticeable change to the protocol efficiency or success
		d) Length of universal primers	Short universal primers with no added i5/i7 and P5/P7 overhang sequences results in a successful protocol. Prior attempts with a P5-i5-Adapter/US 2nd round polymerase chain reaction (PCR) primer fails to give a product due to excessive primer dimerisation and concatemer formation
		e) GC Content of all primers	Primers with a more balanced >40% GC content have shown greater efficacy in PCR and permitted higher annealing temperatures to reduce formation of non-specific product
	Wet-lab techniques	a) Set-up on ice	Performing all steps on ice reduces primer dimerisation and results in greater proportion of product enriched to the target – prevents premature action of the polymerase.
		b) Set-up in UV hood	No effect on presence or absence of contaminants in final PCR reaction as primers are directed to a template of bisulfite converted DNA (bDNA) rather than genomic DNA (gDNA).
Reagents	Polymerase	Multiple polymerases were trialled including Phusion U Hot Start (Thermofisher Scientific, UK); KAPA HiFi HotStart Uracil+ (Roche, UK); TaKaRa EpiTaq HS (Takara, Shiga, Japan); PyroMark Kit (HotStarTaq) (Qiagen, UK); Q5 Ultra II Polymerase (NEB, Massachusetts).	Phusion U polymerase proved the most robust, non-fastidious, easy to set-up and reliable in testing. Q5 Ultra II enzyme has proved reliable for the final library preparation step (3rd Round PCR).
	Primers	Primer concentration gradients	A middling concentration, ~200nM has proved the most efficacious in both singleplex and multiplex reactions in reducing primer dimers but retaining the ability to amplify a target.
	Magnesium	Magnesium concentration gradients	Additional magnesium (from [0 – 3mM]) above the commercial master mix has provided no additional benefit in amplification efficiency and product concentrations
	Dimethyl Sulfoxide (DMSO)	DMSO concentration gradients	Additional DMSO (from 0-10%) above the commercial master mix had provided no additional benefit
	Q® Solution	Addition of 5X Q® solution (Qiagen, UK)	No additional benefit in this protocol has been observed
	Template	a) Concentration	Higher concentrations and quantity of bDNA template improves target enrichment, specificity and PCR product yields, however sensitivity down to starting template of 5ng has been demonstrated.
		b) Template of pre-enriched target i.e. from bisulfite specific PCR	Efficient and reliable amplification using the protocol and this template, however defeats the objective of the protocol for allele specific methylation.
Thermocycling: - 1st Round PCR	Hot Start	Duration times	No significant effect – Manufacturer's recommendation is followed
	Denaturing	Duration times	No significant effect – Manufacturer's recommendation is followed
	Annealing	a) Temperature gradients	Temperature is firstly dependent on good primer design with similar melting temperatures (T_m). Gradient PCR shows best efficacy ~58-60°C for my primer sets. <58°C results in excessive non-specific product, >60°C results in reduced non-specific product but diminishing returns of enriched yield.
		b) Slow ramping vs normal ramp speeds	Slow ramping improves target specific product
		c) Duration times	Time in the order of minutes (vs seconds or hours) appears to improve target yield and specificity. 5 minutes has proved the most appropriate time duration to maximise this effect balanced against the total experimental time.
	Extension	a) Temperatures	Temperatures of 68°C and 72°C tested with no significant difference observed therefore reverted to manufacturer's recommendation
		b) Duration times	No significant effect with these short amplicons – Manufacturer's recommendation is followed

	Cycle number	2, 3, 5 or 10 cycles	Appears to have no significant effect on product yield but reduces the allele specificity of the protocol (see relevant section in text)
Thermocycling: - 2nd Round PCR	Annealing	Temperature gradients	A temperature of 60°C has proved most successful in amplification of the target enriched product
	Other steps	Hot Start, Denaturing, Extension	All are most efficacious following the manufacturer's protocol.
Clean-up Steps	Exonuclease I	With and without	Necessary to ensure digestion of 1st round primers. Use of Exonuclease I does not appear to alter efficacy or yield of reaction
	AMPure XP Beads	With and without	Use of AMPure XP beads between PCR rounds significantly improves target amplification in the 2nd and 3rd round PCR by reduction of dimers and concatemers which are removed by size selection
	Gel Extraction	Extraction of 1st round products from a gel	Attempts at extracting all products between 350-750bp (guided by DNA Hyperladder™) did not result in a successful protocol likely due to loss of low concentration target enriched products in the column based system

***Table S5**: Gives examples of different permutations attempted in the design and optimisation of my allele specific methylation protocol. The actions taken and outcomes for each variable are detailed above. Digital images of agarose gels exist and are available on request for each variable described.*

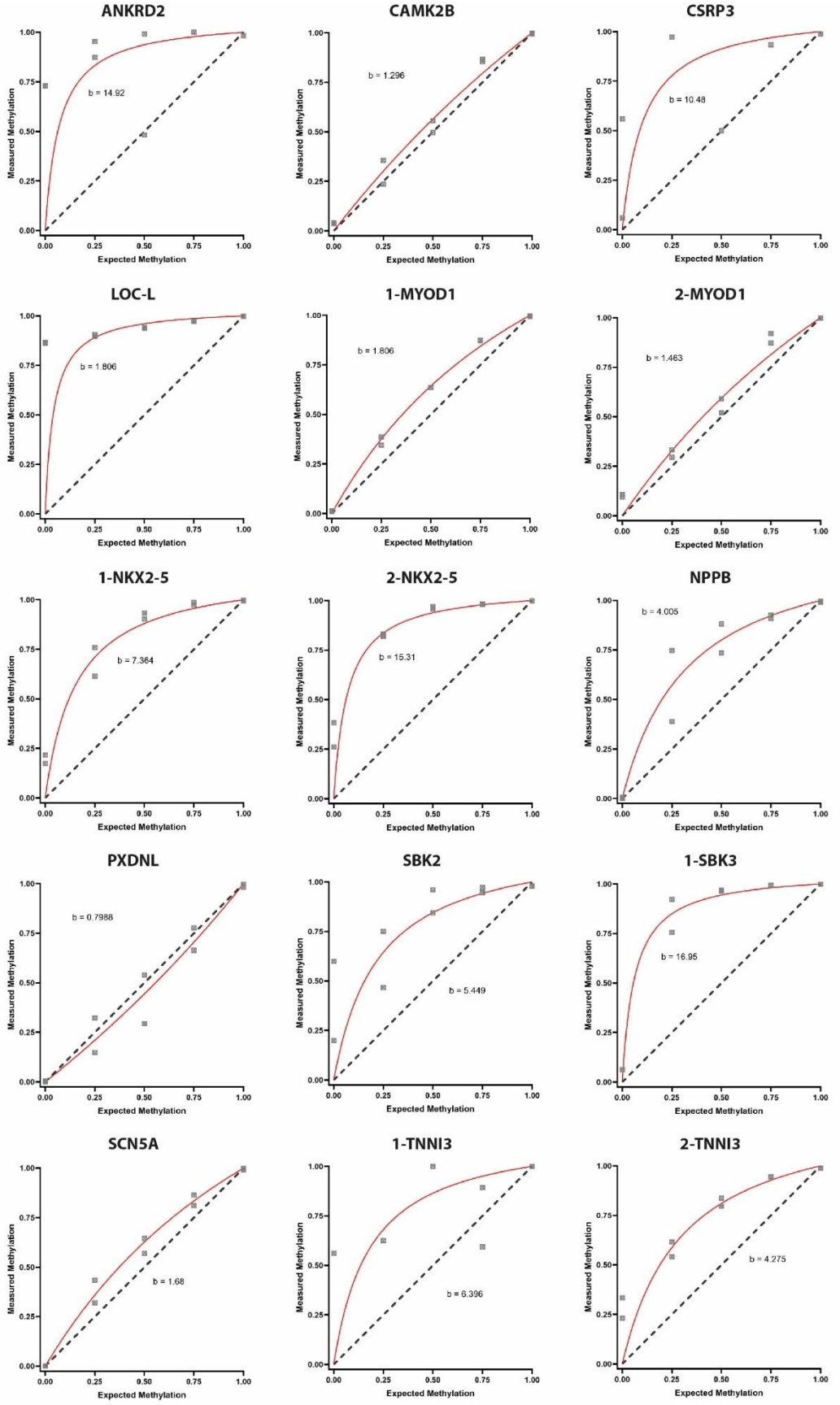

***Figure S1**: Methylation gradient plots for each gene target within the multiplex pool #5 run alongside and constructed from the Barrett's glands sequencing runs.*

10.2 Barrett's glands sample demographic table

The following pages display tables with the demographic details for the samples analysed in this thesis. The full table legend is found on page 285.

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
118	2	B364	55	M	16/05/2018	np	C9M9	Barrett's 34cm	Distal	cardiac
118	2	B365	55	M	16/05/2018	np	C9M9	Barrett's 34cm	Distal	cardiac
118	2	B366	55	M	16/05/2018	np	C9M9	Barrett's 34cm	Distal	cardiac
118	2	B367	55	M	16/05/2018	np	C9M9	Barrett's 34cm	Distal	cardiac
118	2	B368	55	M	16/05/2018	np	C9M9	Barrett's 34cm	Distal	cardiac
118	3	B196	57	M	07/10/2020	np	C9M9	Barrett's 33cm	Distal	specialised
118	3	B197	57	M	07/10/2020	np	C9M9	Barrett's 33cm	Distal	specialised
118	3	B198	57	M	07/10/2020	np	C9M9	Barrett's 33cm	Distal	specialised
118	3	B199	57	M	07/10/2020	np	C9M9	Barrett's 33cm	Distal	specialised
118	3	B200	57	M	07/10/2020	np	C9M9	Barrett's 33cm	Distal	specialised
118	3	B201	57	M	07/10/2020	np	C9M9	Barrett's 33cm	Distal	specialised
118	3	B202	57	M	07/10/2020	np	C9M9	Barrett's 31cm	Middle	specialised
118	3	B203	57	M	07/10/2020	np	C9M9	Barrett's 31cm	Middle	specialised
118	3	B204	57	M	07/10/2020	np	C9M9	Barrett's 31cm	Middle	specialised
118	3	B206	57	M	07/10/2020	np	C9M9	Barrett's 31cm	Middle	specialised
118	3	B207	57	M	07/10/2020	np	C9M9	Barrett's 29cm	Mid-Proximal	specialised
118	3	B209	57	M	07/10/2020	np	C9M9	Barrett's 29cm	Mid-Proximal	specialised
118	3	B211	57	M	07/10/2020	np	C9M9	Barrett's 28cm	Proximal	specialised
118	3	B212	57	M	07/10/2020	np	C9M9	Barrett's 28cm	Proximal	specialised
118	3	B213	57	M	07/10/2020	np	C9M9	Barrett's 28cm	Proximal	specialised
118	3	B214	57	M	07/10/2020	np	C9M9	Barrett's 28cm	Proximal	specialised
118	3	B215	57	M	07/10/2020	np	C9M9	Barrett's 28cm	Proximal	specialised
119	1	B132	62	M	10/06/2015	p	C3M7	Barrett's 38cm	Distal	specialised
119	1	B133	62	M	10/06/2015	p	C3M7	Barrett's 38cm	Distal	cardiac

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
119	1	B134	62	M	10/06/2015	p	C3M7	Barrett's 38cm	Distal	specialised
119	1	B135	62	M	10/06/2015	p	C3M7	Barrett's 38cm	Distal	specialised
119	1	B136	62	M	10/06/2015	p	C3M7	Barrett's 38cm	Distal	specialised
119	1	B137	62	M	10/06/2015	p	C3M7	Barrett's 38cm	Distal	specialised
119	2	B116	62	M	30/09/2015	p	C3M7	BO proximal	Proximal	cardiac
119	2	B117	62	M	30/09/2015	p	C3M7	BO proximal	Proximal	cardiac
119	2	B118	62	M	30/09/2015	p	C3M7	BO proximal	Proximal	cardiac
119	2	B119	62	M	30/09/2015	p	C3M7	BO proximal	Proximal	specialised
119	2	B120	62	M	30/09/2015	p	C3M7	BO proximal	Proximal	specialised
119	2	B121	62	M	30/09/2015	p	C3M7	BO proximal	Proximal	specialised
119	2	B122	62	M	30/09/2015	p	C3M7	BO proximal	Proximal	cardiac
119	2	B123	62	M	30/09/2015	p	C3M7	BO Distal	Distal	cardiac
119	2	B125	62	M	30/09/2015	p	C3M7	BO Distal	Distal	cardiac
119	2	B126	62	M	30/09/2015	p	C3M7	BO Distal	Distal	cardiac
119	2	B127	62	M	30/09/2015	p	C3M7	BO Distal	Distal	specialised
119	2	B128	62	M	30/09/2015	p	C3M7	BO Distal	Distal	specialised
119	2	B129	62	M	30/09/2015	p	C3M7	BO Distal	Distal	specialised
119	2	B130	62	M	30/09/2015	p	C3M7	BO Distal	Distal	specialised
119	2	B131	62	M	30/09/2015	p	C3M7	BO Distal	Distal	specialised
128	1	B023	68	M	16/09/2015	p	C3M4	Barrett's	Proximal	specialised
128	1	B026	68	M	16/09/2015	p	C3M4	Barrett's	Proximal	specialised
128	1	B027	68	M	16/09/2015	p	C3M4	Barrett's	Proximal	specialised
128	1	B028	68	M	16/09/2015	p	C3M4	Barrett's	Proximal	specialised
128	1	B031	68	M	16/09/2015	p	C3M4	Barrett's	Distal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
128	1	B032	68	M	16/09/2015	p	C3M4	Barrett's	Distal	specialised
128	1	B033	68	M	16/09/2015	p	C3M4	Barrett's	Distal	specialised
128	1	B034	68	M	16/09/2015	p	C3M4	Barrett's	Distal	specialised
128	2	B037	71	M	24/01/2018	p	C3M4	Barrett's 37cm	Distal	specialised
128	2	B369	71	M	24/01/2018	p	C3M4	Barrett's 37cm	Distal	cardiac
128	2	B370	71	M	24/01/2018	p	C3M4	Barrett's 37cm	Distal	cardiac
128	2	B371	71	M	24/01/2018	p	C3M4	Barrett's 37cm	Distal	cardiac
128	2	B372	71	M	24/01/2018	p	C3M4	Barrett's 37cm	Distal	cardiac
128	2	B373	71	M	24/01/2018	p	C3M4	Barrett's 37cm	Distal	cardiac
128	2	B374	71	M	24/01/2018	p	C3M4	Barrett's 37cm	Distal	cardiac
128	2	B375	71	M	24/01/2018	p	C3M4	Barrett's 35cm	Proximal	specialised
128	2	B376	71	M	24/01/2018	p	C3M4	Barrett's 35cm	Proximal	specialised
128	2	B377	71	M	24/01/2018	p	C3M4	Barrett's 35cm	Proximal	specialised
128	2	B378	71	M	24/01/2018	p	C3M4	Barrett's 35cm	Proximal	specialised
128	2	B379	71	M	24/01/2018	p	C3M4	Barrett's 35cm	Proximal	specialised
128	3	B319	71	M	16/05/2018	p	C3M4	Barrett's 36cm	Distal	specialised
128	3	B320	71	M	16/05/2018	p	C3M4	Barrett's 36cm	Distal	specialised
128	3	B321	71	M	16/05/2018	p	C3M4	Barrett's 36cm	Distal	specialised
128	3	B322	71	M	16/05/2018	p	C3M4	Barrett's 35cm	Mid-Distal	specialised
128	3	B323	71	M	16/05/2018	p	C3M4	Barrett's 35cm	Mid-Distal	specialised
128	3	B324	71	M	16/05/2018	p	C3M4	Barrett's 35cm	Mid-Distal	cardiac
128	3	B325	71	M	16/05/2018	p	C3M4	Barrett's 33cm	Proximal	specialised
128	3	B326	71	M	16/05/2018	p	C3M4	Barrett's 33cm	Proximal	specialised
128	3	B327	71	M	16/05/2018	p	C3M4	Barrett's 33cm	Proximal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
139	1	B021	57	M	21/10/2015	p	C8M8	Cardia	Cardia	cardiac
139	1	B040	57	M	21/10/2015	p	C8M8	Barrett's	Distal	cardiac
139	1	B041	57	M	21/10/2015	p	C8M8	Barrett's	Distal	cardiac
139	1	B042	57	M	21/10/2015	p	C8M8	Barrett's	Distal	cardiac
139	1	B043	57	M	21/10/2015	p	C8M8	Barrett's	Proximal	specialised
139	1	B044	57	M	21/10/2015	p	C8M8	Barrett's	Proximal	specialised
139	1	B045	57	M	21/10/2015	p	C8M8	Barrett's	Proximal	specialised
139	1	B046	57	M	21/10/2015	p	C8M8	Barrett's	Proximal	specialised
139	1	B048	57	M	21/10/2015	p	C8M8	Barrett's	Distal	cardiac
139	2	B050	60	M	24/01/2018	p	C8M8	Barrett's 37cm	Distal	cardiac
139	2	B051	60	M	24/01/2018	p	C8M8	Barrett's 37cm	Distal	specialised
139	2	B053	60	M	24/01/2018	p	C8M8	Barrett's 37cm	Distal	cardiac
139	2	B056	60	M	24/01/2018	p	C8M8	Barrett's 37cm	Distal	cardiac
139	2	B057	60	M	24/01/2018	p	C8M8	Barrett's 37cm	Distal	cardiac
139	2	B058	60	M	24/01/2018	p	C8M8	Barrett's 32 nodule	Mid-Proximal	specialised
139	2	B059	60	M	24/01/2018	p	C8M8	Barrett's 32 nodule	Mid-Proximal	specialised
139	2	B061	60	M	24/01/2018	p	C8M8	Barrett's 32 nodule	Mid-Proximal	specialised
139	2	B064	60	M	24/01/2018	p	C8M8	Barrett's 32 nodule	Mid-Proximal	dysplasia
139	2	B066	60	M	24/01/2018	p	C8M8	Barrett's 32 nodule	Mid-Proximal	dysplasia
139	2	B067	60	M	24/01/2018	p	C8M8	Barrett's 32 nodule	Mid-Proximal	dysplasia
139	2	B068	60	M	24/01/2018	p	C8M8	Barrett's 32 nodule	Mid-Proximal	dysplasia
139	2	B071	60	M	24/01/2018	p	C8M8	Barrett's 35cm	Mid-Distal	cardiac
139	2	B073	60	M	24/01/2018	p	C8M8	Barrett's 35cm	Mid-Distal	cardiac
139	2	B074	60	M	24/01/2018	p	C8M8	Barrett's 35cm	Mid-Distal	cardiac

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
139	2	B076	60	M	24/01/2018	p	C8M8	Barrett's 35cm	Mid-Distal	cardiac
139	2	B077	60	M	24/01/2018	p	C8M8	Barrett's 35cm	Mid-Distal	cardiac
145	1	B362	62	F	10/02/2016	np	C2M4	Barrett's Distal	Distal	cardiac
145	1	B363	62	F	10/02/2016	np	C2M4	Barrett's Distal	Distal	cardiac
145	2	B267	63	F	03/05/2017	np	C2M4	Barrett's	Middle	cardiac
145	2	B270	63	F	03/05/2017	np	C2M4	Barrett's	Middle	cardiac
145	3	B013	65	F	28/08/2019	np	C2M4	Barrett's 38cm	Distal	specialised
145	3	B017	65	F	28/08/2019	np	C2M4	Barrett's 38cm	Distal	cardiac
148	1	B407	68	M	02/03/2016	np	C0M4	BO proximal	Proximal	specialised
148	1	B408	68	M	02/03/2016	np	C0M4	BO proximal	Proximal	specialised
148	1	B409	68	M	02/03/2016	np	C0M4	BO proximal	Proximal	specialised
148	1	B410	68	M	02/03/2016	np	C0M4	BO proximal	Proximal	specialised
148	1	B411	68	M	02/03/2016	np	C0M4	Barrett's 30cm	Distal	specialised
148	1	B412	68	M	02/03/2016	np	C0M4	Barrett's 30cm	Distal	specialised
148	1	B413	68	M	02/03/2016	np	C0M4	Barrett's 30cm	Distal	specialised
148	1	B414	68	M	02/03/2016	np	C0M4	Barrett's 30cm	Distal	specialised
148	2	B415	70	M	07/03/2018	np	C0M4	Barrett's 31cm	Distal	specialised
148	2	B416	70	M	07/03/2018	np	C0M4	Barrett's 31cm	Distal	specialised
148	2	B417	70	M	07/03/2018	np	C0M4	Barrett's 31cm	Distal	specialised
148	2	B418	70	M	07/03/2018	np	C0M4	Barrett's 31cm	Distal	specialised
148	2	B419	70	M	07/03/2018	np	C0M4	Barrett's 29cm	Proximal	specialised
148	2	B420	70	M	07/03/2018	np	C0M4	Barrett's 29cm	Proximal	specialised
148	2	B421	70	M	07/03/2018	np	C0M4	Barrett's 29cm	Proximal	specialised
148	2	B422	70	M	07/03/2018	np	C0M4	Barrett's 29cm	Proximal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
148	3	B423	70	M	07/03/2018	np	C0M4	Barrett's 37cm	Distal	specialised
148	3	B424	70	M	07/03/2018	np	C0M4	Barrett's 37cm	Distal	specialised
148	3	B425	70	M	07/03/2018	np	C0M4	Barrett's 37cm	Distal	specialised
148	3	B426	70	M	07/03/2018	np	C0M4	Barrett's 37cm	Distal	specialised
148	3	B427	70	M	07/03/2018	np	C0M4	Barrett's 33cm	Proximal	specialised
148	3	B428	70	M	07/03/2018	np	C0M4	Barrett's 33cm	Proximal	specialised
148	3	B429	70	M	07/03/2018	np	C0M4	Barrett's 33cm	Proximal	specialised
153	1	B139	84	M	08/06/2016	p	C5M8	Barrett's 36cm	Proximal	specialised
153	1	B140	84	M	08/06/2016	p	C5M8	Barrett's 36cm	Proximal	cardiac
153	1	B141	84	M	08/06/2016	p	C5M8	Barrett's 36cm	Proximal	specialised
153	1	B144	84	M	08/06/2016	p	C5M8	Barrett's 38cm	Distal	cardiac
153	1	B145	84	M	08/06/2016	p	C5M8	Barrett's 38cm	Distal	cardiac
153	1	B146	84	M	08/06/2016	p	C5M8	Barrett's 38cm	Distal	cardiac
158	1	B431	64	M	15/06/2015	np	C0M3	BO proximal	Proximal	specialised
158	1	B432	64	M	15/06/2015	np	C0M3	BO proximal	Proximal	cardiac
158	1	B433	64	M	15/06/2015	np	C0M3	BO proximal	Proximal	cardiac
158	1	B434	64	M	15/06/2015	np	C0M3	BO proximal	Proximal	cardiac
158	2	B435	67	M	16/05/2018	np	C0M3	Barrett's 40cm	Mid-Distal	cardiac
158	2	B436	67	M	16/05/2018	np	C0M3	Barrett's 40cm	Mid-Distal	specialised
158	2	B437	67	M	16/05/2018	np	C0M3	Barrett's 40cm	Mid-Distal	specialised
158	2	B439	67	M	16/05/2018	np	C0M3	Barrett's 38cm	Proximal	specialised
158	2	B440	67	M	16/05/2018	np	C0M3	Barrett's 38cm	Proximal	specialised
158	2	B441	67	M	16/05/2018	np	C0M3	Barrett's 38cm	Proximal	specialised
158	3	B442	69	M	09/09/2020	np	C0M3	Barrett's 38cm	Distal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
158	3	B443	69	M	09/09/2020	np	C0M3	Barrett's 38cm	Distal	specialised
158	3	B444	69	M	09/09/2020	np	C0M3	Barrett's 38cm	Distal	cardiac
158	3	B445	69	M	09/09/2020	np	C0M3	Barrett's 37cm	Proximal	specialised
158	3	B446	69	M	09/09/2020	np	C0M3	Barrett's 37cm	Proximal	specialised
158	3	B447	69	M	09/09/2020	np	C0M3	Barrett's 37cm	Proximal	specialised
158	3	B448	69	M	09/09/2020	np	C0M3	Barrett's 37cm	Proximal	specialised
164	1	B147	63	M	03/08/2016	p	C0M2	OAC	data not available	dysplasia
164	1	B148	63	M	03/08/2016	p	C0M2	OAC	data not available	dysplasia
164	1	B149	63	M	03/08/2016	p	C0M2	OAC	data not available	dysplasia
164	1	B150	63	M	03/08/2016	p	C0M2	OAC	data not available	dysplasia
168	1	B449	48	M	31/08/2016	np	C3M7	Barrett's	Middle	specialised
168	1	B450	48	M	31/08/2016	np	C3M7	Barrett's	Middle	specialised
168	1	B451	48	M	31/08/2016	np	C3M7	Barrett's	Middle	specialised
168	1	B452	48	M	31/08/2016	np	C3M7	Barrett's	Middle	cardiac
168	1	B453	48	M	31/08/2016	np	C3M7	Barrett's	Middle	specialised
168	2	B454	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	gland base
168	2	B455	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	gland base
168	2	B456	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	gland base
168	2	B457	50	M	05/09/2018	np	C3M7	Barrett's 36cm	Mid-Distal	cardiac
168	2	B459	50	M	05/09/2018	np	C3M7	Barrett's 36cm	Mid-Distal	specialised
168	2	B460	50	M	05/09/2018	np	C3M7	Barrett's 33cm	Proximal	cardiac

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
168	2	B461	50	M	05/09/2018	np	C3M7	Barrett's 33cm	Proximal	specialised
168	2	B462	50	M	05/09/2018	np	C3M7	Barrett's 33cm	Proximal	specialised
168	2	B463	50	M	05/09/2018	np	C3M7	Barrett's 33cm	Proximal	cardiac
168	3	B465	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	specialised
168	3	B466	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	specialised
168	3	B467	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	specialised
168	3	B468	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	specialised
168	3	B469	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	specialised
168	3	B470	50	M	05/09/2018	np	C3M7	Barrett's 38cm	Distal	specialised
168	3	B471	50	M	05/09/2018	np	C3M7	Barrett's 35cm	Middle	specialised
168	3	B472	50	M	05/09/2018	np	C3M7	Barrett's 35cm	Middle	specialised
168	3	B473	50	M	05/09/2018	np	C3M7	Barrett's 35cm	Middle	specialised
168	3	B474	50	M	05/09/2018	np	C3M7	Barrett's 35cm	Middle	specialised
168	3	B475	50	M	05/09/2018	np	C3M7	Barrett's 35cm	Middle	cardiac
168	3	B476	50	M	05/09/2018	np	C3M7	Barrett's 33cm	Proximal	specialised
168	3	B477	50	M	05/09/2018	np	C3M7	Barrett's 33cm	Proximal	specialised
168	3	B478	50	M	05/09/2018	np	C3M7	Barrett's 33cm	Proximal	specialised
168	3	B479	50	M	05/09/2018	np	C3M7	Barrett's 33cm	Proximal	specialised
173	1	B243	52	M	05/10/2016	lgd	C8M9	BO Distal	Distal	cardiac
173	1	B244	52	M	05/10/2016	lgd	C8M9	BO Distal	Distal	specialised
173	1	B246	52	M	05/10/2016	lgd	C8M9	BO Distal	Distal	cardiac
174	1	B380	54	M	05/10/2016	np	C0M4	BO proximal	Proximal	specialised
174	1	B381	54	M	05/10/2016	np	C0M4	BO proximal	Proximal	specialised
174	1	B382	54	M	05/10/2016	np	C0M4	BO proximal	Proximal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
174	1	B383	54	M	05/10/2016	np	C0M4	BO proximal	Proximal	specialised
174	1	B384	54	M	05/10/2016	np	C0M4	BO proximal	Proximal	specialised
174	1	B385	54	M	05/10/2016	np	C0M4	BO Distal	Distal	cardiac
174	1	B386	54	M	05/10/2016	np	C0M4	BO Distal	Distal	specialised
174	1	B387	54	M	05/10/2016	np	C0M4	BO Distal	Distal	cardiac
174	1	B388	54	M	05/10/2016	np	C0M4	BO Distal	Distal	specialised
174	2	B114	56	M	25/07/2018	np	C0M4	Barrett's 37cm	Distal	specialised
174	2	B115	56	M	25/07/2018	np	C0M4	Barrett's 37cm	Distal	specialised
174	2	B315	56	M	25/07/2018	np	C0M4	Barrett's 35cm	Proximal	cardiac
174	2	B316	56	M	25/07/2018	np	C0M4	Barrett's 35cm	Proximal	cardiac
174	2	B317	56	M	25/07/2018	np	C0M4	Barrett's 35cm	Proximal	cardiac
174	2	B318	56	M	25/07/2018	np	C0M4	Barrett's 35cm	Proximal	specialised
174	2	B389	56	M	25/07/2018	np	C0M4	Barrett's 37cm	Distal	specialised
174	2	B390	56	M	25/07/2018	np	C0M4	Barrett's 37cm	Distal	specialised
174	2	B391	56	M	25/07/2018	np	C0M4	Barrett's 37cm	Distal	specialised
174	2	B392	56	M	25/07/2018	np	C0M4	Barrett's 37cm	Distal	specialised
174	2	B393	56	M	25/07/2018	np	C0M4	Barrett's 36cm	Middle	specialised
174	2	B394	56	M	25/07/2018	np	C0M4	Barrett's 36cm	Middle	specialised
174	2	B395	56	M	25/07/2018	np	C0M4	Barrett's 36cm	Middle	cardiac
174	2	B396	56	M	25/07/2018	np	C0M4	Barrett's 36cm	Middle	cardiac
177	1	B265	73	M	12/10/2016	lgd	C11M12	Barrett's	Distal	specialised
177	1	B266	73	M	12/10/2016	lgd	C11M12	Barrett's	Distal	specialised
177	1	B295	73	M	12/10/2016	lgd	C11M12	Barrett's Distal	Distal	specialised
177	1	B296	73	M	12/10/2016	lgd	C11M12	Barrett's Distal	Distal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
177	1	B297	73	M	12/10/2016	lgd	C11M12	Barrett's Distal	Distal	specialised
177	1	B298	73	M	12/10/2016	lgd	C11M12	Barrett's Distal	Distal	specialised
177	1	B299	73	M	12/10/2016	lgd	C11M12	Barrett's Proximal	Proximal	cardiac
177	1	B300	73	M	12/10/2016	lgd	C11M12	Barrett's Proximal	Proximal	cardiac
177	1	B301	73	M	12/10/2016	lgd	C11M12	Barrett's Proximal	Proximal	cardiac
177	1	B302	73	M	12/10/2016	lgd	C11M12	Barrett's Proximal	Proximal	cardiac
177	2	B261	73	M	14/12/2016	lgd	C11M12	BO proximal	Proximal	specialised
177	2	B262	73	M	14/12/2016	lgd	C11M12	BO proximal	Proximal	cardiac
177	2	B263	73	M	14/12/2016	lgd	C11M12	BO proximal	Proximal	cardiac
177	2	B264	73	M	14/12/2016	lgd	C11M12	BO proximal	Proximal	specialised
177	2	B397	73	M	14/12/2016	lgd	C11M12	BO Distal	Distal	cardiac
177	2	B398	73	M	14/12/2016	lgd	C11M12	BO Distal	Distal	cardiac
177	2	B399	73	M	14/12/2016	lgd	C11M12	BO Distal	Distal	cardiac
177	3	B105	75	M	17/01/2018	lgd	C11M12	Barrett's 34cm	Distal	cardiac
177	3	B106	75	M	17/01/2018	lgd	C11M12	Barrett's 34cm	Distal	cardiac
177	3	B107	75	M	17/01/2018	lgd	C11M12	Barrett's 34cm	Distal	cardiac
177	3	B109	75	M	17/01/2018	lgd	C11M12	Barrett's 26cm	Proximal	specialised
177	3	B110	75	M	17/01/2018	lgd	C11M12	Barrett's 26cm	Proximal	specialised
177	3	B111	75	M	17/01/2018	lgd	C11M12	Barrett's 26cm	Proximal	specialised
177	3	B112	75	M	17/01/2018	lgd	C11M12	Barrett's 26cm	Proximal	specialised
177	3	B113	75	M	17/01/2018	lgd	C11M12	Barrett's 26cm	Proximal	specialised
177	5	B216	76	M	18/09/2019	lgd	C11M12	Barrett's 31cm	Distal	cardiac
177	5	B217	76	M	18/09/2019	lgd	C11M12	Barrett's 31cm	Distal	cardiac
177	5	B218	76	M	18/09/2019	lgd	C11M12	Barrett's 31cm	Distal	cardiac

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
177	5	B219	76	M	18/09/2019	lgd	C11M12	Barrett's 31cm	Distal	cardiac
177	5	B220	76	M	18/09/2019	lgd	C11M12	Barrett's 31cm	Distal	cardiac
177	5	B223	76	M	18/09/2019	lgd	C11M12	Barrett's 23cm	Proximal	specialised
177	6	B400	77	M	02/12/2020	lgd	C11M12	Barrett's 30cm	Distal	specialised
177	6	B402	77	M	02/12/2020	lgd	C11M12	Barrett's 30cm	Distal	cardiac
177	6	B403	77	M	02/12/2020	lgd	C11M12	Barrett's 30cm	Distal	specialised
177	6	B404	77	M	02/12/2020	lgd	C11M12	Barrett's 22cm	Proximal	specialised
177	6	B405	77	M	02/12/2020	lgd	C11M12	Barrett's 22cm	Proximal	specialised
177	6	B406	77	M	02/12/2020	lgd	C11M12	Barrett's 22cm	Proximal	specialised
181	1	B337	69	M	16/11/2016	np	C4M5	BO proximal	Proximal	specialised
181	1	B338	69	M	16/11/2016	np	C4M5	BO proximal	Proximal	cardiac
181	1	B339	69	M	16/11/2016	np	C4M5	BO proximal	Proximal	cardiac
181	2	B273	72	M	30/01/2019	np	C4M5	Barrett's 31cm	Proximal	specialised
181	2	B274	72	M	30/01/2019	np	C4M5	Barrett's 31cm	Proximal	specialised
181	2	B275	72	M	30/01/2019	np	C4M5	Barrett's 31cm	Proximal	cardiac
181	2	B276	72	M	30/01/2019	np	C4M5	Barrett's 31cm	Proximal	specialised
181	2	B340	72	M	30/01/2019	np	C4M5	Barrett's 33cm	Distal	cardiac
181	2	B341	72	M	30/01/2019	np	C4M5	Barrett's 33cm	Distal	cardiac
181	2	B342	72	M	30/01/2019	np	C4M5	Barrett's 33cm	Distal	cardiac
181	2	B343	72	M	30/01/2019	np	C4M5	Barrett's 32cm	Mid-Distal	specialised
181	2	B344	72	M	30/01/2019	np	C4M5	Barrett's 32cm	Mid-Distal	gland base
181	2	B345	72	M	30/01/2019	np	C4M5	Barrett's 32cm	Mid-Distal	gland base
181	2	B346	72	M	30/01/2019	np	C4M5	Barrett's 30cm	Proximal	specialised
181	2	B347	72	M	30/01/2019	np	C4M5	Barrett's 30cm	Proximal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
181	2	B348	72	M	30/01/2019	np	C4M5	Barrett's 30cm	Proximal	specialised
181	2	B349	72	M	30/01/2019	np	C4M5	Barrett's 30cm	Proximal	specialised
181	2	B350	72	M	30/01/2019	np	C4M5	Barrett's 30cm	Proximal	cardiac
181	3	B351	74	M	05/05/2021	np	C4M5	Barrett's 31cm	Mid-Proximal	specialised
181	3	B352	74	M	05/05/2021	np	C4M5	Barrett's 31cm	Mid-Proximal	cardiac
181	3	B354	74	M	05/05/2021	np	C4M5	Barrett's 31cm	Mid-Proximal	cardiac
181	3	B355	74	M	05/05/2021	np	C4M5	Barrett's 31cm	Mid-Proximal	cardiac
181	3	B356	74	M	05/05/2021	np	C4M5	Barrett's 30cm	Proximal	specialised
181	3	B357	74	M	05/05/2021	np	C4M5	Barrett's 30cm	Proximal	specialised
181	3	B358	74	M	05/05/2021	np	C4M5	Barrett's 30cm	Proximal	cardiac
181	3	B359	74	M	05/05/2021	np	C4M5	Barrett's 30cm	Proximal	specialised
181	3	B360	74	M	05/05/2021	np	C4M5	Barrett's 30cm	Proximal	specialised
181	3	B361	74	M	05/05/2021	np	C4M5	Barrett's 30cm	Proximal	cardiac
191	1	B278	59	M	25/01/2017	np	C10M11	BO proximal	Proximal	specialised
191	1	B283	59	M	25/01/2017	np	C10M11	BO Distal	Distal	specialised
191	1	B286	59	M	25/01/2017	np	C10M11	BO Distal	Distal	specialised
191	1	B287	59	M	25/01/2017	np	C10M11	BO Distal	Distal	cardiac
191	2	B288	60	M	07/02/2018	np	C10M11	Barrett's 38cm	Distal	cardiac
191	2	B289	60	M	07/02/2018	np	C10M11	Barrett's 38cm	Distal	cardiac
191	2	B290	60	M	07/02/2018	np	C10M11	Barrett's 38cm	Distal	cardiac
191	2	B293	60	M	07/02/2018	np	C10M11	Barrett's 36cm	Mid-Distal	cardiac
191	2	B303	60	M	07/02/2018	np	C10M11	Barrett's 36cm	Mid-Distal	cardiac
191	2	B304	60	M	07/02/2018	np	C10M11	Barrett's 36cm	Mid-Distal	cardiac
191	2	B305	60	M	07/02/2018	np	C10M11	Barrett's 34cm	Mid-Proximal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
191	2	B306	60	M	07/02/2018	np	C10M11	Barrett's 34cm	Mid-Proximal	specialised
191	2	B307	60	M	07/02/2018	np	C10M11	Barrett's 34cm	Mid-Proximal	specialised
191	2	B309	60	M	07/02/2018	np	C10M11	Barrett's 34cm	Mid-Proximal	cardiac
191	2	B310	60	M	07/02/2018	np	C10M11	Barrett's 32cm	Proximal	specialised
191	2	B312	60	M	07/02/2018	np	C10M11	Barrett's 32cm	Proximal	specialised
191	2	B313	60	M	07/02/2018	np	C10M11	Barrett's 32cm	Proximal	specialised
191	2	B314	60	M	07/02/2018	np	C10M11	Barrett's 32cm	Proximal	specialised
191	3	B226	62	M	11/03/2020	np	C10M11	Barrett's 40cm	Distal	cardiac
191	3	B228	62	M	11/03/2020	np	C10M11	Barrett's 40cm	Distal	cardiac
191	3	B229	62	M	11/03/2020	np	C10M11	Barrett's 40cm	Distal	cardiac
191	3	B230	62	M	11/03/2020	np	C10M11	Barrett's 37cm	Mid-Distal	specialised
191	3	B232	62	M	11/03/2020	np	C10M11	Barrett's 37cm	Mid-Distal	specialised
191	3	B233	62	M	11/03/2020	np	C10M11	Barrett's 37cm	Mid-Distal	specialised
191	3	B234	62	M	11/03/2020	np	C10M11	Barrett's 34cm	Mid-Proximal	specialised
191	3	B235	62	M	11/03/2020	np	C10M11	Barrett's 34cm	Mid-Proximal	specialised
191	3	B236	62	M	11/03/2020	np	C10M11	Barrett's 34cm	Mid-Proximal	specialised
191	3	B237	62	M	11/03/2020	np	C10M11	Barrett's 34cm	Mid-Proximal	cardiac
191	3	B238	62	M	11/03/2020	np	C10M11	Barrett's 32cm	Proximal	specialised
191	3	B239	62	M	11/03/2020	np	C10M11	Barrett's 32cm	Proximal	specialised
191	3	B241	62	M	11/03/2020	np	C10M11	Barrett's 32cm	Proximal	specialised
191	3	B242	62	M	11/03/2020	np	C10M11	Barrett's 32cm	Proximal	specialised
199	1	B151	86	M	05/04/2017	p	C2M4	BO Dysplasia	data not available	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
199	1	B152	86	M	05/04/2017	p	C2M4	BO Dysplasia	data not available	specialised
199	1	B154	86	M	05/04/2017	p	C2M4	BO Dysplasia	data not available	specialised
199	1	B155	86	M	05/04/2017	p	C2M4	BO Dysplasia	data not available	specialised
199	1	B156	86	M	05/04/2017	p	C2M4	BO Dysplasia	data not available	specialised
199	1	B157	86	M	05/04/2017	p	C2M4	BO Dysplasia	data not available	specialised
199	1	B158	86	M	05/04/2017	p	C2M4	BO Dysplasia	data not available	specialised
199	1	B159	86	M	05/04/2017	p	C2M4	OAC	data not available	dysplasia
199	1	B160	86	M	05/04/2017	p	C2M4	OAC	data not available	dysplasia
199	1	B161	86	M	05/04/2017	p	C2M4	OAC	data not available	dysplasia
199	1	B162	86	M	05/04/2017	p	C2M4	OAC	data not available	dysplasia
199	1	B163	86	M	05/04/2017	p	C2M4	Barrett’s	data not available	cardiac
199	1	B164	86	M	05/04/2017	p	C2M4	Barrett’s	data not available	cardiac
199	1	B165	86	M	05/04/2017	p	C2M4	Barrett’s	data not available	cardiac

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
199	1	B167	86	M	05/04/2017	p	C2M4	Barrett's	data not available	cardiac
201	1	B169	73	M	06/04/2017	p	C2M5	OAC	data not available	dysplasia
201	1	B171	73	M	06/04/2017	p	C2M5	OAC	data not available	dysplasia
201	1	B172	73	M	06/04/2017	p	C2M5	OAC	data not available	dysplasia
201	1	B173	73	M	06/04/2017	p	C2M5	BO proximal	Proximal	specialised
201	1	B174	73	M	06/04/2017	p	C2M5	BO proximal	Proximal	specialised
201	1	B175	73	M	06/04/2017	p	C2M5	BO proximal	Proximal	specialised
201	1	B176	73	M	06/04/2017	p	C2M5	BO proximal	Proximal	cardiac
201	1	B177	73	M	06/04/2017	p	C2M5	BO Distal	Distal	specialised
201	1	B178	73	M	06/04/2017	p	C2M5	BO Distal	Distal	specialised
220	1	B181	71	M	02/03/2018	p	C2M4	EMR 35cm	Middle	dysplasia
220	1	B182	71	M	02/03/2018	p	C2M4	EMR 35cm	Middle	dysplasia
220	1	B185	71	M	02/03/2018	p	C2M4	EMR 35cm	Middle	dysplasia
220	1	B328	71	M	02/03/2018	p	C2M4	Barrett's 35cm	Middle	cardiac
220	1	B331	71	M	02/03/2018	p	C2M4	Barrett's 35cm	Middle	cardiac
220	1	B332	71	M	02/03/2018	p	C2M4	Barrett's 35cm	Middle	cardiac
221	1	B247	73	M	11/04/2018	lgd	C0M7	Barrett's 39cm	Distal	specialised
221	1	B249	73	M	11/04/2018	lgd	C0M7	Barrett's 39cm	Distal	dysplasia
221	1	B250	73	M	11/04/2018	lgd	C0M7	Barrett's 39cm	Distal	cardiac
221	1	B251	73	M	11/04/2018	lgd	C0M7	Barrett's 39cm	Distal	cardiac
221	1	B252	73	M	11/04/2018	lgd	C0M7	Barrett's 37cm	Middle	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
221	1	B253	73	M	11/04/2018	lgd	C0M7	Barrett's 37cm	Middle	specialised
221	1	B254	73	M	11/04/2018	lgd	C0M7	Barrett's 37cm	Middle	specialised
221	1	B255	73	M	11/04/2018	lgd	C0M7	Barrett's 37cm	Middle	specialised
221	1	B256	73	M	11/04/2018	lgd	C0M7	Barrett's 37cm	Middle	specialised
221	1	B257	73	M	11/04/2018	lgd	C0M7	Barrett's 35cm	Proximal	specialised
221	1	B258	73	M	11/04/2018	lgd	C0M7	Barrett's 35cm	Proximal	cardiac
221	1	B259	73	M	11/04/2018	lgd	C0M7	Barrett's 35cm	Proximal	cardiac
221	1	B260	73	M	11/04/2018	lgd	C0M7	Barrett's 35cm	Proximal	cardiac
231	1	B186	71	F	10/10/2018	p	C3M5	Barrett's 34cm	Distal	dysplasia
231	1	B187	71	F	10/10/2018	p	C3M5	Barrett's 34cm	Distal	dysplasia
231	1	B188	71	F	10/10/2018	p	C3M5	Barrett's 34cm	Distal	dysplasia
231	1	B189	71	F	10/10/2018	p	C3M5	Barrett's 34cm	Distal	dysplasia
231	1	B190	71	F	10/10/2018	p	C3M5	Barrett's 34cm	Distal	dysplasia
231	1	B191	71	F	10/10/2018	p	C3M5	Barrett's 31cm	Proximal	specialised
231	1	B192	71	F	10/10/2018	p	C3M5	Barrett's 31cm	Proximal	specialised
231	1	B193	71	F	10/10/2018	p	C3M5	Barrett's 31cm	Proximal	specialised
231	1	B194	71	F	10/10/2018	p	C3M5	Barrett's 31cm	Proximal	specialised
231	1	B195	71	F	10/10/2018	p	C3M5	Barrett's 31cm	Proximal	specialised
UF001	1	B078	63	F	11/03/2002	p	M7	Distal Barrett's	Distal	specialised
UF001	1	B079	63	F	11/03/2002	p	M7	Distal Barrett's	Distal	specialised
UF001	1	B080	63	F	11/03/2002	p	M7	Distal Barrett's	Distal	specialised
UF001	1	B081	63	F	11/03/2002	p	M7	Distal Barrett's	Distal	cardiac
UF001	1	B082	63	F	11/03/2002	p	M7	Proximal Barrett's	Proximal	specialised
UF001	1	B083	63	F	11/03/2002	p	M7	Proximal Barrett's	Proximal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
UF001	1	B084	63	F	11/03/2002	p	M7	Proximal Barrett's	Proximal	specialised
UF001	1	B085	63	F	11/03/2002	p	M7	Proximal Barrett's	Proximal	cardiac
UF002	1	B086	77	M	07/04/2003	p	M6	Distal Barrett's 37cm	Distal	specialised
UF002	1	B088	77	M	07/04/2003	p	M6	Distal Barrett's 37cm	Distal	specialised
UF002	1	B089	77	M	07/04/2003	p	M6	Distal Barrett's 37cm	Distal	specialised
UF002	1	B090	77	M	07/04/2003	p	M6	Distal Barrett's 37cm	Distal	specialised
UF002	1	B091	77	M	07/04/2003	p	M6	Distal Barrett's 37cm	Distal	specialised
UF002	1	B093	77	M	07/04/2003	p	M6	Proximal Barrett's 32cm	Proximal	specialised
UF002	1	B094	77	M	07/04/2003	p	M6	Proximal Barrett's 32cm	Proximal	specialised
UF002	1	B095	77	M	07/04/2003	p	M6	Proximal Barrett's 32cm	Proximal	specialised
UF004	1	B096	66	M	21/07/2003	p	data not available	Distal BO	Distal	cardiac
UF004	1	B097	66	M	21/07/2003	p	data not available	Distal BO	Distal	cardiac
UF004	1	B098	66	M	21/07/2003	p	data not available	Distal BO	Distal	specialised
UF004	1	B099	66	M	21/07/2003	p	data not available	Distal BO	Distal	cardiac
UF004	1	B100	66	M	21/07/2003	p	data not available	Distal BO	Distal	cardiac
UF004	1	B101	66	M	21/07/2003	p	data not available	Distal BO	Distal	specialised

Patient	TP	Gland	Age	Sex	Biopsy Date	Cohort	Prague Criteria	Biopsy Description	Segment Location	Phenotype
UF004	1	B102	66	M	21/07/2003	p	data not available	Distal BO	Distal	cardiac
UF004	1	B103	66	M	21/07/2003	p	data not available	Distal BO	Distal	cardiac
UF004	1	B334	66	M	21/07/2003	p	data not available	BO proximal	Proximal	cardiac
UF004	1	B335	66	M	21/07/2003	p	data not available	BO proximal	Proximal	cardiac
UF004	1	B336	66	M	21/07/2003	p	data not available	BO proximal	Proximal	cardiac

Table S6: *Demographic data for each laser capture microdissected Barrett's gland (notation of B******) organised by patient code (far left column) and TP (timepoint). Table details the age of patient at biopsy date, Prague score at that timepoint and biopsy description including free text description, location and histological glandular phenotype. Cohort column represents whether the particular gland/patient is classed as a non-progressor (np), progressor (p) or low-grade dysplasia (lgd).*