Single cell RNA-seq reveals profound transcriptional similarity between Barrett's oesophagus and oesophageal submucosal glands

Citation for published version:
Owen, RP, White, MJ, Severson, DT, Braden, B, Bailey, A, Goldin, R, Wang, LM, Ruiz-Puig, C, Maynard, ND, Green, A, Piazza, P, Buck, D, Middleton, MR, Ponting, CP, Schuster-Böckler, B & Lu, X 2018, 'Single cell RNA-seq reveals profound transcriptional similarity between Barrett's oesophagus and oesophageal submucosal glands' Nature Communications, vol. 9, no. 1, pp. 4261. DOI: 10.1038/s41467-018-06796-9

Digital Object Identifier (DOI):
10.1038/s41467-018-06796-9

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Nature Communications

Barrett's oesophagus is associated by an increased risk of oseophageal cancer, but its cell of origin is unclear. Here the authors show, using single-cell RNA sequencing of biopsies from 6 patients and 2 unaffected subjects, that cells in Barrett's oesophagus show a transcriptional profile that is similar to that of cells in oesophageal submucosal glands.

Single cell RNA-seq reveals profound transcriptional similarity between Barrett's oesophagus and oesophageal submucosal glands

Authors

Richard Peter Owen[1†], Michael Joseph White[1†], David Tyler Severson[1†], Barbara Braden[2], Adam Bailey[2], Robert Goldin[3], Lai Mun Wang[4], Carlos Ruiz Puig[1], Nicholas David Maynard[5], Angie Green[6], Paolo Piazza[6^], David Buck[6], Mark Ross Middleton[7], Chris Paul Ponting[8], Benjamin Schuster-Böckler[1*] and Xin Lu[1*]

Affiliations

1. Ludwig Institute for Cancer Research, Nuffield Department of Medicine, University of Oxford, Oxford, UK. OX3 7DQ
2. Translational Gastroenterology Unit, Nuffield Department of Medicine, University of Oxford, Oxford, UK. OX3 9DU
3. Centre for Pathology, St Mary's Hospital, Imperial College, London, UK. W2 1NY
4. Department of Pathology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. OX3 9DU

5. Department of Upper GI Surgery, Oxford University Hospitals, Oxford, UK. OX3 7LE
6. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. OX3 7BN
7. Department of Oncology, Old Road Campus Research Building, Roosevelt Drive Oxford, UK. OX3 7DQ
8. MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Crewe Road, Edinburgh, UK. EH4 2XU

†These authors contributed equally to this work; ^Current address, Faculty of Medicine, Department of Medicine, Imperial College London, London, UK;
*Correspondence to: xin.lu@ludwig.ox.ac.uk and benjamin.schuster-boeckler@ludwig.ox.ac.uk.

Abstract

Barrett's oesophagus is a precursor of oesophageal adenocarcinoma. In this common condition, squamous epithelium in the oesophagus is replaced by columnar epithelium in response to acid reflux. Barrett's oesophagus is highly heterogeneous and its relationships to normal tissues are unclear. Here we investigate the cellular complexity of Barrett's oesophagus and the upper gastrointestinal tract using RNA-sequencing of single cells from multiple biopsies from six patients with Barrett's oesophagus and two patients without oesophageal pathology. We find that cell populations in Barrett's oesophagus, marked by *LEFTY1* and *OLFM4*, exhibit a profound transcriptional overlap with oesophageal submucosal gland cells, but not with gastric or duodenal cells. Additionally, SPINK4 and ITLN1 mark cells that precede morphologically identifiable goblet cells in colon and Barrett's oesophagus, potentially aiding the identification of metaplasia. Our findings reveal striking transcriptional relationships between normal tissue populations and cells in a premalignant condition, with implications for clinical practice.

Introduction

At least 80% of cancers arise from epithelial cells. In many tumours a change in cell type, referred to as metaplasia, is a key step in cancer initiation. Barrett's oesophagus (BO) is an example of metaplasia in the distal oesophagus and affects 1 in 50 people[1]. BO is defined as replacement of squamous epithelium by columnar epithelium, and it gives a 30-fold increased risk of developing oesophageal adenocarcinoma (OAC) which has a five year survival of only 15%[2-4]. BO is associated with gastro-oesophageal reflux disease, suggesting it occurs in response to a chronically inflamed environment[5]. Remarkably, several anatomically distant cell types are also identifiable in BO, most commonly intestinal goblet cells but also Paneth and pancreatic acinar cells, among others[6-8].

This apparent plasticity in BO has obscured its relationship with normal gastrointestinal (GI) tissues, as no normal GI tissue is as heterogeneous as BO. Several theories are proposed for the origin of BO. A widely held view is that BO originates from the stomach[9,10], and studies looking for similarities (e.g. in gene or protein expression and cellular appearance) between BO and selected normal tissues - including the intestine, gastric pylorus, gastric corpus and gastric cardia – have found some shared attributes[11,12]. There is also evidence suggesting BO may originate directly from native oesophageal squamous[13] or submucosal gland cells[14-17], from recruitment of circulating stem cells[18], or from reactivation of dormant p63$^-$/KRT7$^+$ residual embryonic cells (RECs) *in situ*[19]. In contrast to p63$^-$/KRT7$^+$ RECs, a recent study identified p63$^+$/KRT5$^+$/KRT7$^+$ cells derived from the squamocolumnar junction as the cells of origin of BO in a transgenic mouse model with ectopic expression of CDX2 in KRT5$^+$ epithelium[20]. Many of the proposed BO origin theories are based on transgenic mouse studies, and the submucosal gland cell theories are based on human histopathology studies. Unfortunately, submucosal gland theories cannot be tested in mice since mice and humans have key differences in their gastrointestinal anatomy, and rodents lack oesophageal glands[21]. These difficulties argue for an unbiased and systematic genetic approach to BO characterisation in humans with all relevant control cell types to better understand the origin of BO cell types.

Single cell RNA-sequencing (RNA-seq) combined with computational methods for functional clustering of cell types provides a less biased approach to understanding cellular heterogeneity. Given the highly heterogeneous nature of BO, we hypothesise that single cell RNA-seq might clarify the relationships between cells in normal tissues and BO, and indicate whether there are specialised cells in BO with similar functions to cells elsewhere in the gastrointestinal tract. Therefore we apply this approach to biopsies from BO, normal oesophagus, stomach and small intestine (duodenum). This reveals a cell population in BO that expresses the developmental gene (*LEFTY1)* and is distinct from intestinal or gastric cells, but has a highly similar RNA composition to columnar gene expressing cells from oesophageal submucosal glands in normal oesophagus.

Results

Single cell RNA-seq identifies subpopulations in normal upper GI epithelia

To characterise the cell populations in BO, samples were taken from 13 BO patients (A-D, I-Q) attending for routine endoscopic surveillance of non-dysplastic BO. From each patient, we took biopsies from BO, adjacent macroscopically normal oesophagus (20mm proximal to BO), stomach (20mm distal to the gastro-oesophageal junction) and duodenum (**Figure 1a**). Individual 2mm biopsies were divided to provide tissue for single cell RNA-seq, bulk tissue RNA-seq and histology in 4 out of 13 patients, and bulk tissue RNA-seq and histology alone in the remaining 9 patients (see **Methods**). Single cells and histology were also prepared from normal oesophageal biopsies from two patients with gastro-oesophageal reflux disease but no previous or current diagnosis of BO or any other oesophageal pathology. All sampled patients were taking regular acid suppression therapy and had no features of oesophageal dysplasia or malignancy (**Supplementary Table 1**).

Bulk RNA-sequencing followed by hierarchical clustering of differentially expressed genes in the duodenal, gastric, oesophageal and BO samples from 13 patients with BO showed a clear distinction between squamous (i.e. normal oesophagus) and non-squamous (i.e. gastric, duodenum and BO) epithelia (**Figure 1b**). BO samples from all 13 patients had some similarities to duodenal and gastric samples (**Figure 1b**). When a defined list of genes known to distinguish gastrointestinal epithelia[12] was used in hierarchical clustering, BO samples appeared most closely related to gastric tissue, consistent with previous studies[22] (**Figure 1c**).

For single cell RNA-seq, a total of 4237 cells were sequenced from 8 patients (**Supplementary Table 1**) in three batches. Due to known issues with batch effects in single cell experiments[23], analysis of cells from each batch has been kept separate where feasible and cells were permuted across plates and pooled prior to sequencing (see **Methods**). The first batch yielded 1040 cells (207 duodenum, 227 gastric, 371 BO and 235 oesophagus) suitable for analysis from four patients (A-D) with BO and intestinal metaplasia. A total of 214, 35, 66 and 56 BO cells were analysed from each BO patient, respectively. The second batch yielded 648 oesophagus cells suitable for analysis from two patients (E-F) with symptoms of gastro-oesophageal reflux but no identifiable oesophageal pathology. Finally, the third batch of cells yielded 194 cells (29 pylorus, 109 gastric, 32 BO and 24 oesophagus) suitable for analysis from two patients (G-H) with BO and intestinal metaplasia. Overall, there was a mean of $1.2x10^5$ gene counts per cell and a median of 3978 genes were detected per cell (with at least one count per gene).

First, we clustered the cells from each normal tissue type from the BO patients by gene expression (**Figure 1d**). The eleven clusters (D1-D4, G1-G3 and O1-O4, in duodenum, gastric and oesophageal samples, respectively) were then annotated on the basis of genes previously characterized as expressed in specific cell types (complete list in **Supplementary Data 1**). In the duodenum, these are: intestinal alkaline phosphatase (*ALPI*)-expressing enterocytes (D1); mucin 2 (*MUC2*)-expressing goblet cells (D2); olfactomedin 4 (*OLFM4*)-expressing crypt cells (D3); and some uncharacterized cells expressing Joining Chain Of Multimeric IgA And IgM (*JCHAIN*) (D4). In the gastric samples, these are: chromogranin (*CHGA*)-expressing enteroendocrine cells (G1); gastrokinin (*GKN1*)- and trefoil factor 1 (*TFF1*)-expressing foveolar cells (G2); and mucin 6 (*MUC6*)- and *TFF1*-expressing mucus neck cells (G3). Of note, the proton pump gene *ATP4A* and the intrinsic factor gene *GIF* were rarely detectable in gastric cells, indicating these are cardiac-type gastric samples **(Supplementary Fig. 1**).

Interestingly, four clusters were identified in the oesophageal samples. Two of these express expected squamous genes (*KRT5*, *KRT14*, *TP63*; clusters O1 and O2) and two express the columnar gene *TFF3* (clusters O3 and O4). The two squamous clusters can be distinguished by the presence (O1) or absence (O2) of acute phase response (*SAA1*) gene expression, presumably representing squamous cells in different states. The detection of *TFF3* in O3 and O4 is of great interest and is consistent with these cells being from the columnar epithelium of oesophageal submucosal glands (OSGs)[24], a structure in the normal human oesophagus. To validate this, we used samples of normal oesophagus taken from the proximal part of an oesophagectomy specimen following resection for a Siewert type III junctional tumour to illustrate the structure of OSGs, OSG ducts and squamous epithelium (**Figure 1e**). Since OSGs comprise different cell lineages, including squamous lineages, we detected cytokeratin 14 (KRT14, a squamous cell marker)-expressing cells in OSG ducts, demonstrating they are bona fide OSGs. Using the adjacent sections from the same OSG-containing specimen, we observed TFF3 and keratin 7 expression in OSG structures exclusively (**Figure 1f**). These results show that single cell transcriptomic analysis can identify gastrointestinal epithelial cell subpopulations, including cells from OSGs that cannot be distinguished by conventional bulk RNA-seq.

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

To identify genes characteristic of distinct BO cell populations we clustered all the BO cells by gene expression (**Figure 2a,** also see **Supplementary Data 1**). The clusters (B1-B4) can be distinguished by expression of *MUC2* (B1; goblet cells, 19% of BO cells); *LEFTY1* (B2 and B3, 71% of BO cells); and *CHGA* (B4; enteroendocrine cells, 9.7% of BO cells). Since all patients had intestinal metaplasia, goblet cells made up 22%, 2.9%, 29% and 7.1% of cells in patient A-D, respectively. *KRT7* is expressed similarly across all 4 clusters, consistent with it being a marker of BO[25,26]. The *LEFTY1*-expressing cells (B2 and B3; **Figure 2a**) are divided into a larger, low proliferating (*MKI67* (Ki67) negative) cluster (B2) and a smaller, high proliferating (*MKI67* positive) cluster (B3). LEFTY1, a secreted protein and transforming growth factor beta (TGF-β) superfamily member, is normally expressed in development, where it has roles in left-right asymmetry determination[27], but little is known about its potential roles in adult tissues and it has not previously been associated with BO.

To confirm the above finding and to further characterise LEFTY1 expression, we first examined MUC2, LEFTY1 and CHGA expression in sections generated from the same BO resection specimen. LEFTY1 expression was detected in BO epithelial cells (**Supplementary Data 2**). Interestingly, morphologically identifiable goblet cells are positive for MUC2 but not LEFTY1 or CHGA (**Figure 2b**).

To further characterise LEFTY1 expression, we stained 140 BO samples from 80 patients, 78 endoscopic biopsies from control sites (oesophagus, gastric fundus and duodenum) in 26 BO patients, and additionally five endoscopic samples from the pylorus, five resected samples of normal colon and five samples of normal oesophagus taken from the proximal part of an oesophagectomy specimen resected for junctional tumours (**Supplementary Data 2**). Overall there are two different LEFTY1 staining patterns: intensely positive cytoplasmic staining and moderate cytoplasmic staining. Moderate LEFTY1 staining only, was seen in the Brunner's gland of the duodenum and in the lower portion of the glands in the gastric fundus. In the colon there are a few, intensely positively LEFTY1 staining cells. Both moderate and intensely expressing LEFTY1 cells are present in the gastric pylorus and BO (**Supplementary Fig. 2**). Immunohistochemical staining of oesophageal samples showed that the squamous epithelium was negative for LEFTY1 staining, as were the OSGs in oesophagectomy samples from non-BO patients. All three OSGs from the 140 oesophageal samples showed moderate cytoplasmic staining throughout the OSG (**Figure 2c).** These expression patterns explain why the more superficial mucosal biopsies obtained for single cell RNA-seq show dramatic differences in *LEFTY1* expression between tissues.

OSGs share an RNA composition profile with Barrett's oesophagus

Taking all cells from BO patients together (A-D), the normal tissue cells separate clearly from the BO cells based on their gene expression, with the exception of specialised cell types such as goblet or enteroendocrine cells, but the majority of BO cells overlap with a sub-set of oesophageal cells, as seen in a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot (**Figure 3a**)**.** Clustering by gene expression (by the same method as in **Figure 1d**) assigned cells to 7 clusters (with brain controls in a separate cluster) (**Figure 3b, c,** also see **Supplementary Fig. 3a**). Most of these clusters are similar to those identified in the analysis of normal tissue alone (**Figure 1d**) and they can be related to known cell types based on expression of previously characterised genes (**Supplementary Fig. 3b,** also see **Supplementary Data 3** for complete list). The majority of duodenal cells fall in the cluster categorised as 'enterocytes' (similar to D1), gastric as 'mucus neck' (similar to G3), and a substantial proportion of oesophageal cells are in the 'squamous' cluster (similar to O1/O2) (**Figure 3c**). Some oesophageal cells, BO cells and a few duodenal cells fall into a 'goblet' cluster, and some gastric cells cluster with a few BO cells in the 'enteroendocrine' cluster. The group described as 'non-epithelial' contains some endothelial cells and *CD45*-low immune cells (**Supplementary Fig. 4**). Notably, the majority of BO cells (63%) are in the cluster labelled as 'Barrett's-type' that also contains the subset of oesophageal cells that have a gene expression profile consistent with their being OSGs (**Figure 3c,** also see **Supplementary Data 3**). These cells are enriched for *LEFTY1* expression.

To test whether this relationship between BO and native oesophageal cells with columnar characterisation was also seen in patients without BO, we clustered all normal oesophageal cells from patients with and without BO (A, B, D and E, F, respectively). This showed that cells grouped into five clusters (**Supplementary Fig. 5a**), three clusters (1, 2 and 4) were mainly squamous and the remaining two (3 and 5) had more columnar marker-expressing cells. Of the ‘columnar’ clusters, cluster 5 consisted of cells from patients A and B and cluster 3 consisted of cells from patients B, D and E (patients A, B, D had BO, patients E, F had no BO) (**Supplementary Fig. 5b**). Although rare in these data, it is interesting that one of the clusters (cluster 3) containing *TFF3*$^+$ cells also had four cells which were positive for the squamous genes *KRT14* (a gene pair with *KRT5*), *TP63* and *KRT7* (**Supplementary Fig. 5c**). As p63$^+$ KRT7$^+$ cells have been shown to generate intestinal-like epithelial cells in organoid culture upon CDX2 overexpression, it may be possible that these oesophageal cells could be related to the transitional zone progenitor cells previously observed in humans[21].

To confirm whether the relationship between BO cells and OSGs was stronger than the associations with other gland-type cells, we looked across the RNA compositions of cells from other tissues, i.e. gastric gland cells and BO cells that did not express *CHGA* or *MUC2* (to exclude enteroendocrine and goblet cells, respectively; see **Methods** for thresholding), and oesophageal cells that expressed *TFF3* (to exclude squamous cells, **Supplementary Fig. 5d-e**). We also developed BEARscc, an algorithm which uses external controls to simulate technical replicates to check whether a single cell clustering method is robust to technical variability[28]. The ‘score’ metric of BEARscc reflects how frequently cells within a group cluster together, as opposed to with cells from other clusters. We compared manually selected groups of 1) gastric and BO cells, 2) gastric and OSG cells, and 3) BO and OSG cells, from patients with BO (A-D). The BO and OSG cell combination had a higher score than any combination which included gastric cells, or all cells grouped together, suggesting BO and OSG cells have the most stable cell type relationship (**Figure 3d**). Using only these manually selected gastric, BO and OSG cells with additional OSG cells from patients without BO (E-F), unbiased clustering with SC3 also confirmed the strong relationship between BO and OSG cells, with only very few gastric cells clustering with BO or OSG cells (**Supplementary Fig. 6a**). t-SNE, with the inclusion of duodenal cells which expressed the highest levels of

MUC6 to enrich for duodenal Brunner's gland-type cells (**Supplementary Fig. 6b**), also confirmed the strong relationship between BO and OSG cells (**Supplementary Fig. 6c**). This relationship was characterised by high *LEFTY1* expression (**Supplementary Fig. 6d**). Only a small number of genes show differential expression between BO cells and OSG cells that did not express *CHGA* and *MUC2* (to exclude enteroendocrine and goblet cells). Pathway analysis on these genes did not suggest any biological processes that mechanistically distinguish BO and OSG cells (**Supplementary Fig. 6e-f**).

In view of the phenotypic overlap with BO and gastric pylorus, we analysed the transcriptomes of 194 cells from an additional two patients (G-H) with BO (24 oesophageal cells, 32 BO cells, 109 gastric cardia cells and 29 gastric pyloric cells). Clustering of these cells on global and specific gene expression show that gastric cardia and pylorus exhibited similar RNA composition properties (**Supplementary Fig. 7**). The BO cells also expressed several of the gastric genes, but showed differences such as increased *KRT7* and *BPIFB1* expression (**Supplementary Fig. 7b**). Collectively, these data show that oesophageal cells expressing genes seen in OSGs, and not intestinal, gastric or squamous cells, have the greatest RNA composition similarity to BO cells.

ITLN1 and SPINK4 mark early goblet cells

In this study, 19% of BO cells were classified as 'goblet' cells, which is consistent with the requirement in some countries, such as the US[29], for goblet cells to be present for the diagnosis of BO. Goblet cells are classically defined by morphological appearance and MUC2 expression. Applying a threshold set at the tenth centile to include 90% of cells in which at least one transcript was detected from each gene of interest (to reduce biological noise), we found that *MUC2* RNA co-expressed with intelectin 1 (*ITLN1*) and Kazal type 4 serine peptidase inhibitor (*SPINK4*) in 61% of goblet cells from duodenum, gastric and BO samples (**Figure 4a-b**). ITLN1 and SPINK4 have been previously shown to mark goblet cells in the normal gut and some non-gastrointestinal tissues[30,31], but we observed some cells in each tissue type that uniquely expressed *MUC2*, *ITLN1* or *SPINK4*. Therefore we hypothesized that their expression pattern might mark stages of goblet cell development *in vivo*. To test this, we analysed expression of these proteins by immunofluorescence staining of five human colon samples (approximately 500 crypts examined in each sample). ITLN1 and SPINK4 co-staining was consistently present near the crypt base, where undifferentiated cells occur, whereas MUC2 staining was in cells toward the centre and top of the crypts, where terminally differentiated cells are found (**Figure 4c**). This suggests that ITLN1 and SPINK4 might mark an earlier stage of goblet cell differentiation than MUC2 in the intestine.

In the three patients with OSGs found in the 140 squamous endoscopic biopsies from 80 patients with BO, we observed that OSG cells consistently co-expressed ITLN1, and MUC2, but not SPINK4. This may be because SPINK4 positive cells are more 'naïve' in goblet cell differentiation and thus they are present lower in the duct or gland and were not captured within these biopsies (**Figure 4d**). In these same three patients we found a squamous marker (KRT14, which pairs with KRT5 in p63+ cells), a columnar marker (KRT7) and a specialised goblet cells marker (MUC2) expressed in adjacent cells in the same OSG (**Figure 4e**). This intestinal metaplasia in an OSG from a squamous oesophageal biopsy 20mm proximal to the BO margin suggests the ability of OSGs to undergo intestinalisation and may be the source of BO islands[32]. In 30 BO endoscopic mucosal resection (EMR) specimens (from 16 patients) with intestinal metaplasia but no dysplasia present, we also consistently observed cells expressing ITLN1 or SPINK4 without MUC2 (**Figure 4f,** also see **Supplementary Table 2**). Specifically, quantification of triple immunofluorescence staining of eight BO EMR specimens with intestinal metaplasia but no dysplasia taken from five patients showed 41% of MUC2 low cells expressed SPINK4 and/or ITLN1, whereas 28% of cells expressed MUC2 alone (**Supplementary Table 3**). These data suggest that OSGs and BO may contain early goblet cells, as seen in the colon, and that ITLN1 or SPINK4 might mark cells with some goblet cell characteristics that are not yet morphologically identifiable as goblet cells.

OLFM4 marks a stem-like transcript profile in BO and OSG epithelium

StemID is a published workflow designed to find cells with stem-like properties in single cell RNA-seq data by calculating a 'stem-ness' score based on the entropy of cell clusters and the number of links between clusters[33,34]. As a control we analysed duodenum cells from BO patients (A-D) and found the highest scoring cluster was enriched for *LGR5* expression, consistent with *LGR5* being a known marker of intestinal stem cells[35,36]. Applying StemID to the remaining individual tissues from the same patients did not identify any well-known stem cell markers (**Supplementary Fig. 8a-b**), even though a small number of LGR5 positive cells are present in all tissues sequenced (**Supplementary Fig. 1**). Since a recent study showed that BO contains pluripotent cells[37] and in view of the striking transcript profile overlap between OSG and BO cells, we therefore analysed all BO and OSG cells using StemID (patients A-F). Interestingly, the highest scoring cluster was enriched for the stem-cell associated gene *OLFM4* (**Figure 5a**, blue asterisk). BO cells from all four patients with BO (A-D) contributed to this cluster, and oesophageal cells from two patients with BO (A and B) (**Supplementary Fig. 8c**). The second highest scoring cell cluster (**Figure 5a**, red asterisk) was enriched for *LYZ*, a marker of Paneth cells, which are long-lived secretory cells found adjacent to the stem cell niche in the intestinal crypt base. *OLFM4* has been shown to associate with *LGR5* expression and marks stem cells in intestinal tissue in normal and metaplastic contexts[38,39]. Consistent with this, immunohistochemical staining detected OLFM4 expression in human colon crypt bases, where stem cells are known to be located (**Figure 5b**). In 8 BO sections from 7 patients, we observed that OLFM4 protein expression was less restricted to the crypt base (**Figure 5c**), similar to previous observations of LGR5 expression patterns in BO[12] and in contrast to the expression of OLFM4 in control tissues (**Supplementary Fig. 8d**). In OSGs beneath normal squamous epithelium, OLFM4 positive cells were seen within the gland structures (**Figure 5d**). Interestingly, OLFM4 staining in OSGs from patients without BO was much more restricted than seen in OSGs taken from patients with BO (**Figure 5d, e**), although the number of cases examined is limited.

Notably, *OLFM4* has a higher mean expression in the *LEFTY1*-positive clusters (B2/B3) compared to the clusters expressing known markers of the differentiated goblet (*MUC2*) and enteroendocrine (*CHGA*) lineages (**Figure 2a**, B1 and B4, respectively). To examine co-expression of *OLFM4*, *LEFTY1*, *MUC2* and *CHGA* in individual cells, we applied a threshold at the tenth centile to include 90% of cells in which at least one transcript was detected from each gene of interest. Using this threshold, half of the BO cells express *LEFTY1* and *OLFM4*, alone or in combination (29% *OLFM4* and *LEFTY1*; 13% *OLFM4* only; 11% *LEFTY1* only). *LEFTY1* and *OLFM4* positive BO cells rarely co-expressed *MUC2* or *CHGA* (**Supplementary Fig. 8e)**. Together, these data suggest that B2/B3 represent a cell population that harbours BO progenitor cells.

Discussion

Our single cell RNA-seq data has resolved cell sub-populations in gastrointestinal epithelia and shown a profound similarity in the transcript profile between OSG cells and BO cells. This is supported by our observation that this sub-population of BO cells and OSGs express the stem cell-associated gene *OLFM4*, in line with the notion that these populations might contain similar progenitor cells. Glandular epithelial cells are replaced by squamous epithelium during development of the oesophagus and OSGs are functionally important structures formed from remaining glandular epithelium[40]. It is thus not surprising that the developmental gene *LEFTY1* is expressed in OSGs, and that as these structures expand during the development of BO, increased levels of LEFTY1 and OLFM4 are observed in these tissues. Notably, *LEFTY1* is regulated by TGF-β signalling and bone morphogenic proteins (BMPs)[41,42]. Since TGF-β is often perturbed in BO, and BMPs have been shown to play a major role in the development of a BO like phenotype, it will be interesting to explore these relationships further[43,44].

Additionally, our findings support a previously proposed hypothesis that BO may originate from OSGs. This model suggests that acid and bile reflux-induced damage to the oesophagus is ‘repaired’ by the expansion or selection of OSGs, which contain progenitors that may express OLFM4 and have alkaline secretions, and are thus able to play a role in protecting the oesophagus from gastro-oesophageal reflux damage. Further consideration of the functional overlap of other secretory structures with BO and OSGs, such as salivary and mammary glands may help our understanding of an adaptive response to injury that drives metaplasia. Studies are also needed to experimentally demonstrate the potential of OSG cells, p63$^+$ or p63$^-$ OSGs in particular, to develop into BO cells and OAC.

Given that rodents lack OSGs, and the lack of an *in vitro* model of human oesophageal glands, analysis of human biopsies currently provides the most reliable approach to dissect the cell relationships of BO. Future improvements in single cell sequencing techniques may enable more systematic genetic confirmation of the cellular origin of BO through DNA analysis and also allow higher throughput, to reduce any potential selection bias inherent in the methodology we have used, especially with respect to gastric cells, which were likely to have been detrimentally affected by acid exposure. Also, it is important to note that our study cannot definitively identify the origins of OAC. Future studies are needed to address the relationship between BO and OAC on a cellular level, and how this relates to recent work suggesting that OAC is highly similar to a sub-set of gastric cancers[45].

Finally we showed that SPINK4 and ITLN1 seem to identify an earlier stage of intestinal metaplasia than marked by MUC2, given that they are expressed lower in intestinal crypts than MUC2 and can be seen without MUC2 in BO. Of clinical importance, our results suggest that intestinal goblet cell characteristics exist even in the absence of morphologically identifiable goblet cells, supporting the view that diagnosis of BO should not require the detection of goblet cells. Together, our findings help characterize BO in humans. In addition, this study demonstrates the power of single cell analysis of clinical samples to uncover biological relationships among cell types and cellular heterogeneity in healthy and diseased tissues.

Methods

Sampling

Patients attending routine endoscopic surveillance of BO and patients with mild reflux symptoms undergoing gastroscopy for diagnostic purposes gave written informed consent and provided samples (patients A-F and I-Q, study authorised by South Central - Oxford C Research Ethics Committee: 09/H0606/5+5; patients G-H, study authorised by Yorkshire & The Humber - Sheffield Research Ethics Committee: 16/YH/0247). Patient numbers were chosen to provide suitable biological replicates, and cells sequenced to provide balanced sample sizes at sequencing input. Double bite quadrantic 2mm biopsies were obtained endoscopically using standard biopsy forceps (Radial Jaw 4 Standard Capacity, Boston Scientific, Natick, USA) from a central region of the BO segment avoiding the proximal BO margin as well as the oesophagogastric junction. Control samples were taken from the second part of the duodenum, the stomach 20mm distal to the gastro-oesophageal junction and the normal oesophageal squamous epithelium at least 20mm clear of the most proximal extent of BO. Each sample was fragmented and then pooled to ensure all sampling sites were represented in each investigative modality. Fragments pools were divided into three groups for histological verification, whole-tissue RNA-seq and single cell RNA-seq (**Figure 1a**). Patients were selected based on their previously known pathological features (**Supplementary Table 1** and **Supplementary Fig. 9**). Patients without BO described 0-2 reflux episodes per week with normal endoscopic appearances of the upper gastrointestinal tract on endoscopic examination, and no histological evidence of oesophagitis in the processed samples.

Cell isolation

Sample fragments were placed directly into a digestion solution (made with 1x phosphate buffered solution (Gibco™), 2mM EDTA, 100U ml^{-1} type I collagenase (Worthington Biochemical Company®), sodium phosphate (5.6mM), monopotassium phosphate (8mM), sodium chloride (96mM), potassium chloride (1.6mM), sucrose (44mM), D-Sorbitol (55mM), Dl-Dithiotreitol (0.5mM)) and gently oscillated at 4°C for 60 minutes. Samples were then further fragmented with scissors and briefly manually triturated with a p1000 pipette. Fragments were allowed to settle and the cell-containing supernatant filtered (Sysmex Celltrics® 100 micron) into a 15ml Falcon tube. This process was repeated 3 times and the product centrifuged at 300g for 20 minutes at 4°C to create a cell pellet which was resuspended in sorting buffer (1x phosphate buffered solution (Gibco™), 2mM EDTA and 5% heat inactivated fetal bovine serum (Sigma-Aldrich®)). A small amount of each sample was pooled for labelling controls. Pre-conjugated CD45-FITC (1:10, mouse monoclonal, cat. 130-080-202, Miltenyi Biotec)[46] and EpCAM-PE (1:10, mouse monoclonal, cat. 130-110-999, Miltenyi Biotec)[47] antibodies were added to cell suspensions to help identify epithelial and immune cells, respectively, and they were incubated/washed according to manufacturer's advice. DAPI (1:2000, Sigma-Aldrich®) was added to cell suspensions immediately prior to sort. FACS was carried out using a BD Biosciences FACS Aria IIIu platform with 70μm nozzle in the case of the first four patients and the additional squamous samples, and a Sony SH800S Cell Sorter with 100μm chip in the second batch of two patients including the pyloric samples. Cells were selected based on size and singlet gating to saturate cell output while minimising debris passed to subsequent gates. Size and singlet gating were then adjusted to capture of EpCAM+ cells, on the basis that these would represent a range of epithelial cells and minimise debris selection (**Supplementary Fig. 10a**). Resultant cells were sorted directly into 96 well plates (Life Technologies™ MicroAmp® Optical 96-well Reaction Plate) pre-prepared with 2μl 0.2% Triton™ X-100 (Sigma-Aldrich®) and RNAse inhibitor (Takara Recombinant RNase Inhibitor) at 19:1 and then immediately frozen on dry ice. To confirm spectral accuracy, compensation bead controls and pooled cell suspensions were used for fluorescence-minus-one controls where possible. Each plate was re-permuted to avoid batch effects at the next stages of preparation, with no single plate containing cells from only a single patient or tissue type. Variable patterns of 6 blank wells were also prepared in each plate, 3 of which had a 10pg of brain total RNA (Agilent Technologies) added as a positive control. A single 100 cell pool was also sorted in experiments involving pyloric cells (patients G-H) to provide a bulk control as whole tissue RNA-seq was not performed in these patients. To check for bias in cell selection, index sorting was carried out in most experiments to analyse expression of antibodies in relation to tissue type and subsequent data quality (**Supplementary Fig. 10b-d**). Using the input metrics available up to the point of sequencing, logistic regression was also undertaken to see if higher quality cell data could be predicted before sequencing. While the length of the experiment tended towards having an effect on data quality, recorded metrics at FACS could not accurately predict whether a cell would meet a read count threshold (**Supplementary Fig. 10d**).

Single cell RNA-seq

Transcriptome libraries were prepared using a Biomek FX liquid handling instrument (Beckman Coulter) with a custom adaptation of the published smart-seq2 method[48,49], with minor modifications, and Nextera XT (Illumina®) methodology with custom, unique index primers after tagmentation and ERCC spike-in at a dilution of 1:100,000. Libraries were sequenced using the Illumina® HiSeq 4000 platform, aiming for 3.5×10^5 reads per cell at 75bp paired end.

Bulk RNA-seq

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

Data analysis

All data were mapped using STAR[50] (release 2.5.2a) to the hg19 version of the human genome with transcriptome annotations from Gencode (release 25). Counts tables were made with HTSeq[51]. Cells were excluded that didn't meet a threshold set to exclude all negative controls and outliers, and includes all remaining positive controls, see **Supplementary Fig. 11a-c**). For example, this was fewer than 25,119 fragments mapping to the transcriptome in the first experiment (patients A-D). No oesophageal cells from patient C passed this quality control threshold. To check biological relevance, counts from the most abundant cell population from a single patient and tissue were summed and correlated against bulk RNA-seq expression (**Supplementary Fig. 11d**). Counts were trimmed mean of M-values (TMM)-normalised and fragments per kilobase million (FPKM) values were calculated. Genes with less than 4 FPKM in at least 3 cells were filtered out. After re-normalisation, expression values were converted to transcripts per kilobase million (TPM). A further gene filtering step was included to remove highly expressed genes with low variability across all samples (cells in the top decile for mean expression and below the fifth centile for coefficient of variation). SC3[52] was used to provide cell cluster information. Cluster robustness to experimental technical variation was tested using BEARscc[28] which models technical noise from ERCC spike-in measurements. Cluster number, k, was chosen manually using the distribution of cluster-wise mean silhouette widths across clusters in all 250 simulated technical replicates for each cluster number k (2 to 8 for individual tissue and 1 to 15 for all tissues). Where box plots are used, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles), the whiskers extend from the hinge to the largest or smallest values at most 1.5 x inter-quartile range from the hinge. Data beyond the whiskers are outliers and are plotted individually. t-SNE data were generated using the Barnes-Hut implementation of t-SNE[53] in R. Differential expression analysis was carried out between cell groups using edgeR[54] from normalized counts according to the package manual. P values used were determined by permutation test at 5% (250-1000 permutations) to allow for multiple comparisons or, in cases of unbalanced sample numbers, converted to false discovery rates (FDR) by the Benjamini-Hochberg procedure. Pathway analysis was performed using goseq[55] to identify over or under represented ontological terms. Identification of stem-like cells was performed using RaceID2 and StemID, please see https://github.com/dgrun/StemID for more details[33,34]. Further results from this analysis showing differentially expressed genes in high stem-scoring clusters are available in **Supplementary Data 4**. Where gene expression is described in binary terms, the threshold was set to include or exclude 90% of cells with the highest expression of a given gene, to allow for biological noise.

Immunohistochemistry and immunofluorescence on human tissue

Oesophageal samples from oesophagectomy specimens (5 patients) containing normal mucosa and gland structures and endoscopic mucosal resection specimens (30 patients) with Barrett's oesophagus were obtained from the Oxford Radcliffe and Translational Gastroenterology Unit biobanks. Sections were de-waxed, rehydrated and incubated with 3% hydrogen peroxide in methanol to block endogenous peroxidase activity (10 minutes, room temperature). Antigen retrieval was carried out using 10mM sodium citrate, pH6 at 100°C for 10 minutes. Sections were then blocked with normal goat serum (at room temperature) and incubated overnight at 4 °C with a primary antibody against anti-KRT14 (IHC, 1:1000, rabbit polyclonal, cat. PRB-155P, BioLegend), anti-TFF3 (IHC, 1:1000, mouse monoclonal, cat. WH0007033M1, Sigma-Aldrich®)[56], anti-MUC2 (IHC, 1:300, rabbit polyclonal, cat. SC-15334, Santa Cruz Biotechnology)[57], anti-CHGA (IHC, 1:500, rabbit polyclonal, cat. ab15160, Abcam)[58], anti-KRT7 (IHC, 1:4000, rabbit monoclonal, cat. ab181598, Abcam)[59], anti-LEFTY1 (IHC, 1:1000, D7E3G rabbit polyclonal, cat. 12647, Cell Signalling), anti-OLFM4 (IHC, 1:200, D1E4M rabbit monoclonal, cat. 14369, Cell Signalling Technology®), anti-ITLN1 (IHC/IF, 1:500, sheep polyclonal, cat. AF4254, R&D systems)[60], anti-MUC2 (IF, 1:300, mouse monoclonal, cat. ab11197, Abcam)[61] or anti-SPINK4 (IF, 1:500, rabbit polyclonal, cat. HPA007286, Sigma-Aldrich®)[62]. For immunohistochemical staining, samples were then treated with biotinylated secondary antibody (Vector Labs; 1:250) for 40 minutes at room temperature. The staining reaction was worked up using the Vector Elite ABC kit and counterstained with haematoxylin. Samples were examined by a pathologist using a histology microscope. For immunofluorescent staining, expression was detected using Alexa Fluor (1:250, Molecular Probes) for one hour. DAPI (1:2000, Sigma-Aldrich®) was used to stain nucleic acids. Samples were observed using a confocal microscope system (LSM 710; Carl Zeiss). The limited amount of material obtained from patients precluded the use of each described staining technique on every sample collected.

Data availability

Single cell and bulk RNA-seq counts data and the cell cluster assignments for each analysis are supplied in the **Supplementary Data Files 5-7**. Raw data are available in the European Genome-phenome Archive, following the necessary consents to protect donor anonymity (accession # EGAS00001003144). All other data available upon request.

References

1 Zagari, R. M. *et al.* Prevalence of upper gastrointestinal endoscopic findings in the community: A systematic review of studies in unselected samples of subjects. *J Gastroenterol Hepatol* **31**, 1527-1538 (2016).

2 CRUK. <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer> (2016).

3 Solaymani-Dodaran, M., Logan, R. F. A., West, J., Card, T. & Coupland, C. Risk of oesophageal cancer in Barrett's oesophagus and gastro-oesophageal reflux. *Gut* **53**, 1070-1074 (2004).

4 Cunningham, D., Okines, A. F. & Ashley, S. Capecitabine and oxaliplatin for advanced esophagogastric cancer. *N Engl J Med* **362**, 858-859 (2010).

5 Kavanagh, M. E. *et al.* The esophagitis to adenocarcinoma sequence; the role of inflammation. *Cancer Lett* **345**, 182-189 (2014).

6 Chaves, P. *et al.* Gastric and intestinal differentiation in Barrett's metaplasia and associated adenocarcinoma. *Dis Esophagus* **18**, 383-387 (2005).

7 Griffin, M. & Sweeney, E. C. The relationship of endocrine cells, dysplasia and carcinoembryonic antigen in Barrett's mucosa to adenocarcinoma of the oesophagus. *Histopathology* **11**, 53-62 (1987).

8 Krishnamurthy, S. & Dayal, Y. Pancreatic metaplasia in Barrett's esophagus. An immunohistochemical study. *Am J Surg Pathol* **19**, 1172-1180 (1995).

9 Quante, M. *et al.* Bile acid and inflammation activate gastric cardia stem cells in a mouse model of Barrett-like metaplasia. *Cancer Cell* **21**, 36-51 (2012).

10 White, N. M. *et al.* Barrett's esophagus and cardiac intestinal metaplasia: two conditions within the same spectrum. *Can J Gastroenterol* **22**, 369-375 (2008).

11 Paull, A. *et al.* The histologic spectrum of Barrett's esophagus. *N Engl J Med* **295**, 476-480 (1976).

12 Lavery, D. L. *et al.* The stem cell organisation, and the proliferative and gene expression profile of Barrett's epithelium, replicates pyloric-type gastric glands. *Gut* **63**, 1854-1863 (2014).

13 Hu, Y. *et al.* The pathogenesis of Barrett's esophagus: secondary bile acids upregulate intestinal differentiation factor CDX2 expression in esophageal cells. *J Gastrointest Surg* **11**, 827-834 (2007).

14 Leedham, S. J. *et al.* Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human Barrett's oesophagus. *Gut* **57**, 1041-1048 (2008).

15 Lorinc, E., Mellblom, L. & Oberg, S. The immunophenotypic relationship between the submucosal gland unit, columnar metaplasia and squamous islands in the columnar-lined oesophagus. *Histopathology* **67**, 792-798 (2015).

16 Coad, R. A. *et al.* 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* **206**, 388-394 (2005).

17 von Furstenberg, R. J. *et al.* Porcine Esophageal Submucosal Gland Culture Model Shows Capacity for Proliferation and Differentiation. *Cell Mol Gastroenterol Hepatol* **4**, 385-404 (2017).

18 Hutchinson, L. *et al.* 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* **20**, 11-17 (2011).

19 Wang, X. *et al.* Residual embryonic cells as precursors of a Barrett's-like metaplasia. *Cell* **145**, 1023-1035 (2011).

20 Jiang, M. *et al.* Transitional basal cells at the squamous-columnar junction generate Barrett's oesophagus. *Nature* **550**, 529-533 (2017).

21 Macke, R. A. *et al.* Barrett's esophagus and animal models. *Ann N Y Acad Sci* **1232**, 392-400 (2011).

22 McDonald, S. A., Lavery, D., Wright, N. A. & Jansen, M. Barrett oesophagus: lessons on its origins from the lesion itself. *Nat Rev Gastroenterol Hepatol* **12**, 50-60 (2015).

23 Hicks, S. C., Teng, M. & Irizarry, R. A. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv*, doi:10.1101/025528 (2015).

24 Long, J. D. & Orlando, R. C. Esophageal submucosal glands: structure and function. *Am J Gastroenterol* **94**, 2818-2824 (1999).

25 Ormsby, A. H. *et al.* Cytokeratin immunoreactivity patterns in the diagnosis of short-segment Barrett's esophagus. *Gastroenterology* **119**, 683-690 (2000).

26 Shearer, C., Going, J., Neilson, L., Mackay, C. & Stuart, R. C. Cytokeratin 7 and 20 expression in intestinal metaplasia of the distal oesophagus: relationship to gastro-oesophageal reflux disease. *Histopathology* **47**, 268-275 (2005).

27 Hamada, H., Meno, C., Watanabe, D. & Saijoh, Y. Establishment of vertebrate left-right asymmetry. *Nature Reviews Genetics* **3**, 103-113 (2002).

28 Severson, D. T., Owen, R. P., White, M. J., Lu, X. & Schuster-Bockler, B. BEARscc determines robustness of single-cell clusters using simulated technical replicates. *Nat Commun* **9**, 1187 (2018).

29 Spechler, S. J. *et al.* American Gastroenterological Association Medical Position Statement on the Management of Barrett's Esophagus. *Gastroenterology* **140**, 1084-1091 (2011).

30 Washimi, K. *et al.* Specific Expression of Human Intelectin-1 in Malignant Pleural Mesothelioma and Gastrointestinal Goblet Cells. *Plos One* **7** (2012).

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

32 Sharma, P., Morales, T. G., Bhattacharyya, A., Garewal, H. S. & Sampliner, R. E. Squamous islands in Barrett's esophagus: what lies underneath? *Am J Gastroenterol* **93**, 332-335 (1998).

33 Grun, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251-255 (2015).

34 Grun, D. *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* **19**, 266-277 (2016).

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

36 Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* **459**, 262-265 (2009).

37 Yamamoto, Y. *et al.* Mutational spectrum of Barrett's stem cells suggests paths to initiation of a precancerous lesion. *Nat Commun* **7**, 10380 (2016).

38 van der Flier, L. G., Haegebarth, A., Stange, D. E., van de Wetering, M. & Clevers, H. OLFM4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. *Gastroenterology* **137**, 15-17 (2009).

39 Jang, B. G., Lee, B. L. & Kim, W. H. Intestinal Stem Cell Markers in the Intestinal Metaplasia of Stomach and Barrett's Esophagus. *PLoS One* **10** (2015).

40 Rishniw, M. *et al.* Molecular aspects of esophageal development. *Barrett's Esophagus: The 10th Oeso World Congress Proceedings* **1232**, 309-315 (2011).

41 Miyata, N. *et al.* Transforming Growth Factor beta and Ras/MEK/ERK Signaling Regulate the Expression Level of a Novel Tumor Suppressor Lefty. *Pancreas* **41**, 745-752 (2012).

42 Smith, K. A. *et al.* Bmp and nodal independently regulate lefty1 expression to maintain unilateral nodal activity during left-right axis specification in zebrafish. *PLoS Genet* **7** (2011).

43 Hyland, P. L. *et al.* Global Changes in Gene Expression of Barrett's Esophagus Compared to Normal Squamous Esophagus and Gastric Cardia Tissues. *Plos One* **9** (2014).

44 Mari, L. *et al.* A pSMAD/CDX2 complex is essential for the intestinalization of epithelial metaplasia. *Cell reports* **7**, 1197-1210 (2014).

45 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169-175 (2017).

46 Kurian, L. *et al.* Conversion of human fibroblasts to angioblast-like progenitor cells. *Nature Methods* **10**, 77-U116 (2013).

47 Metsuyanim, S. *et al.* Expression of Stem Cell Markers in the Human Fetal Kidney. *Plos One* **4** (2009).

48 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1098 (2013).

49 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181 (2014).

50 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

51 Anders, S., Pyl, P. T. & Huber, W. HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

52 Kiselev, V. Y. *et al.* SC3 - consensus clustering of single-cell RNA-Seq data. *Nature Methods* **14**, 483-486 (2017).

53 van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* **15**, 3221-3245 (2014).

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

55 Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**, R14 (2010).

56 Perera, O. *et al.* Trefoil factor 3 (TFF3) enhances the oncogenic characteristics of prostate carcinoma cells and reduces sensitivity to ionising radiation. *Cancer Letters* **361**, 104-111 (2015).

57 Kosinsky, R. L. *et al.* Usp22 deficiency impairs intestinal epithelial lineage specification in vivo. *Oncotarget* **6**, 37906-37918 (2015).

58 Chang, P. C. *et al.* Autophagy Pathway Is Required for IL-6 Induced Neuroendocrine Differentiation and Chemoresistance of Prostate Cancer LNCaP Cells. *Plos One* **9** (2014).

59 Fransen-Pettersson, N. *et al.* A New Mouse Model That Spontaneously Develops Chronic Liver Inflammation and Fibrosis. *Plos One* **11** (2016).

60 Greulich, S. *et al.* Cardioprotective Properties of Omentin-1 in Type 2 Diabetes: Evidence from Clinical and In Vitro Studies. *Plos One* **8** (2013).

61 He, Y. F. *et al.* High MUC2 Expression in Ovarian Cancer Is Inversely Associated with the M1/M2 Ratio of Tumor-Associated Macrophages and Patient Survival Time. *Plos One* **8** (2013).

62 Wapenaar, M. C. *et al.* The SPINK gene family and celiac disease susceptibility. *Immunogenetics* **59**, 349-357 (2007).

Author contributions

R.P.O. and M.J.W. collected biopsy samples and prepared them for sequencing. M.J.W. carried out the immunoreactive staining and imaging. M.J.W. and C.R.P. processed the FFPE samples. R.P.O. and D.S.T. carried out RNA-seq mapping and data analysis. B.B., A.B., M.M. and N.D.M. helped to design and curate the clinical data and sample collection. R.G. and L.M.W. provided pathological interpretation of all samples used. A.G., P.P. and D.B. generated all sequencing data used. C.P.P. provided computational oversight of the data analysis. B.S.-B. provided overall supervision of the computational analysis of the data and X.L. provided overall supervision of the project. The manuscript was written by R.P.O., M.J.W. and X.L., with assistance from B.S.-B. and D.S.T. Figures were prepared by R.P.O., M.J.W. and D.S.T.

Competing interests

The authors declare no competing interests. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Acknowledgments

The work is mainly funded by the Ludwig Institute for Cancer Research Ltd. with additional support from the NIHR Biomedical Research Centre and a CRUK Accelerator Award (C328/A21998). R.P.O. received funding from Oxford Health Services Research Committee and Oxford University Clinical Academic Graduate School. M.J.W. was supported by Cancer Research UK (C5255/A19498, through an Oxford Cancer Research Centre Clinical Research Training Fellowship). D.T.S. was supported by Nuffield Department of Clinical Medicine and the Clarendon Fund. We thank Mary Muers and Françoise Howe for helping with critical reading of the manuscript, Andrew Roth for comments on statistical analysis of the data, Sally-Ann Clark and Paul Sopp for providing FACS expertise, John Findlay for help with ethical approval, and Rory Bowden, Amy Trebes and the High-Throughput Genomics team at the Wellcome Trust Centre for Human Genetics, Oxford for assistance with sequencing.

Figure 1. Single cell RNA sequencing identifies cell groups in normal upper gastrointestinal epithelia

(a) Endoscopic sampling sites (yellow, oesophagus; green, gastric cardia; purple, duodenum; orange, Barrett's oesophagus) with summary of how tissues from patients were used. 2-4 biopsies were taken at each site. Patients without BO were sampled from the lower oesophagus 20mm proximal to the squamous-columnar junction. **(b)** From bulk RNA-seq data derived from samples from 13 patients with BO, heatmap of genes differentially expressed between any tissue type (analysis of variance-like test, false discovery rate (FDR) $< 1x10^{-12}$) with tissue hierarchy determined by nearest neighbour. Tissue indicated by colours as in **a**. One duodenal sample from patient Q failed to produce usable data and was excluded. **(c)** From bulk RNA-seq data, heatmap of expression of mucin and trefoil factor genes with tissue hierarchy determined by nearest neighbour, in samples from 13 patients with BO. **(d)** Upper panels show the cluster consensus matrices for single cells from normal tissue sites in four BO patients. Blue-to-red colours denote the frequency with which cells are grouped together in 250 repeat clusterings of simulated technical replicates (see Methods). Cell clusters are indicated by coloured bars below the matrices. In lower panels, heatmaps show expression of known functionally relevant genes that were differentially expressed between cell clusters (>4 fold change, FDR <1e-5). **(e)** Haematoxylin and eosin staining of normal oesophagus taken from the proximal part of an oesophagectomy specimen resected for Siewert type III junctional tumour in a patient with no BO, showing OSGs (red arrow), OSG ducts (black arrow) and squamous epithelium (marked with dotted black line). Scale bar 500µm. **(f)** Immunohistochemical staining of KRT14, TFF3 and KRT7 (left, middle and right images, respectively) in adjacent sections from the same specimen as **e,** showing OSG ducts (black arrows) and OSGs (red arrows) and squamous epithelium (marked with dotted black line). Scale bar 500µm. OSG, oesophageal submucosal gland.

Figure 2. *LEFTY1* and *OLFM4* are mainly expressed in Barrett's oesophagus cells that do not express differentiated secretory cell markers

(a) Upper panel, cluster consensus matrix of BO cells from 4 BO patients (n=371 cells). Blue-to-red colours denote the frequency with which cells are grouped together in 250 repeat clusterings of simulated technical replicates (see **Methods**). Clusters (B1-B4) are indicated by the coloured bars below. Lower panel, heatmaps showing expression of selected functionally relevant genes that are differentially expressed between cell clusters (>4 fold change, FDR <1e-5). **(b)** Immunohistochemical staining of MUC2, LEFTY1 and CHGA in sections derived from the same BO resection specimen. Black arrows indicate goblet cells on all sections (positively stained for MUC2; negative for LEFTY1 and CHGA) Scale bars are 50µm. **(c)** Immunohistochemical staining of LEFTY1 in an OSG from a normal squamous endoscopic biopsy obtained from a patient with BO. Scale bars are 300µm and 50µm in enlarged image.

Figure 3. The majority of Barrett's oesophagus cells have a similar transcript profile to oesophageal submucosal gland (OSG) cells

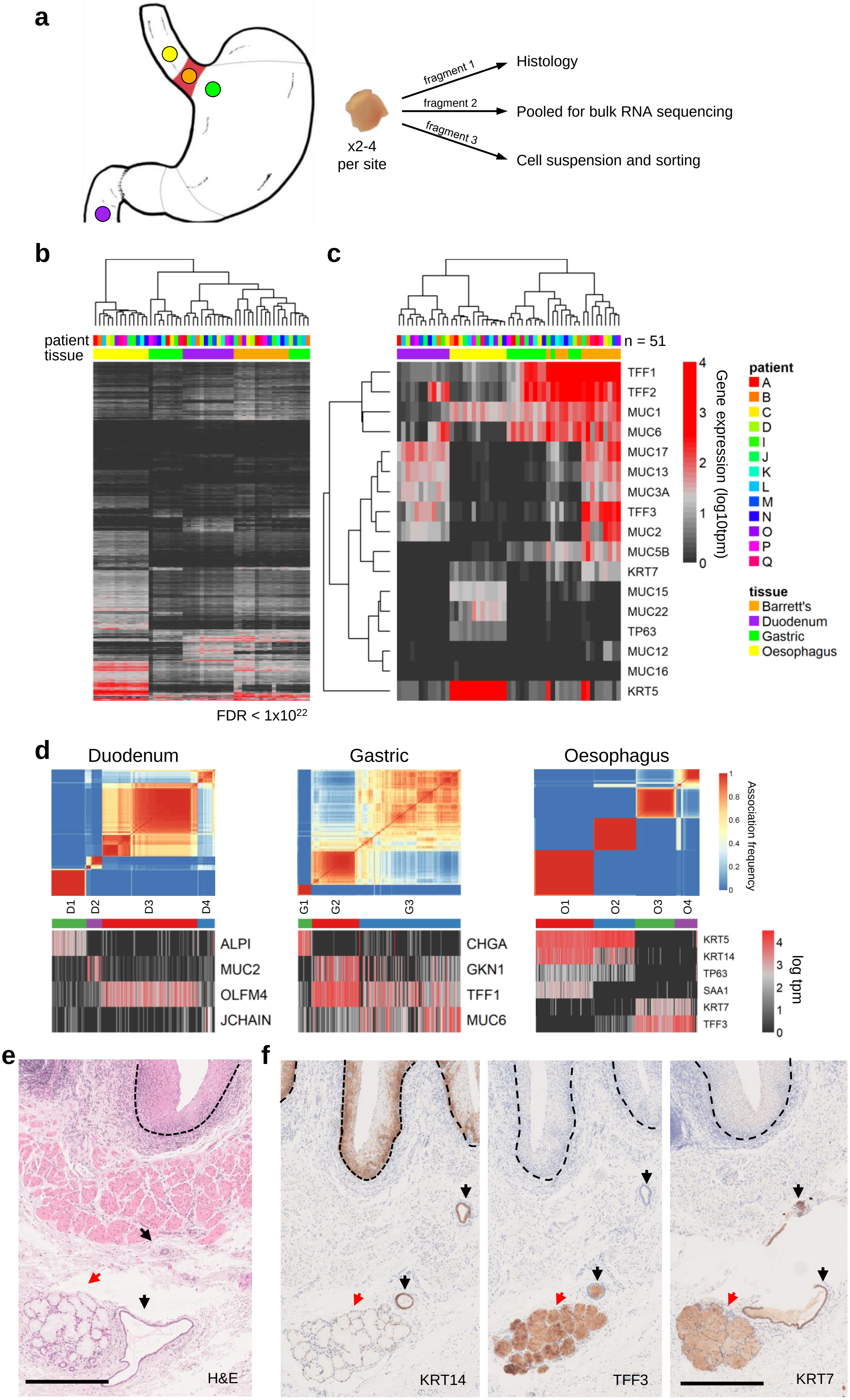
(a) t-Distributed Stochastic Neighbour Embedding (t-SNE) plots of cells from all samples from four BO patients (n=1107 including brain control), showing similarity of cells in two dimensions, coloured by tissue type (yellow, oesophagus; green, gastric cardia; purple, duodenum; orange, Barrett's oesophagus; pink, brain). Brain was used as a control. **(b)** t-SNE plot of cells from four BO patient samples (A-D), as in **a**, coloured by how cells contribute to clusters generated by SC3 analysis with 250 repeat clusterings of simulated technical replicates (see **Methods**). Names given to the clusters are based on expression of known marker genes (see text and **Supplementary Fig. 3**). (**c**) Sankey diagram showing how each tissue type sampled contributes to the clusters shown in **b**. Colours and labels on the left indicate sampled tissue (as in **a**); colours and labels on the right indicate cluster (as in **b**). **(d)** Mean BEARscc score for each grouping of 'gland-like' cells (n=372), which are a sub-set of gastric (G, n=175), BO (n=78) and OSG cells (n=119): excluding gastric and BO cells that expressed *CHGA* or *MUC2* (to exclude enteroendocrine and goblet cells, respectively) and excluding oesophageal cells that did not express *TFF3* (to exclude squamous cells). 'Ensemble' refers to all cells grouped together. Thresholds were set at the tenth centile of cells in which at least one transcript was detected from each gene of interest.

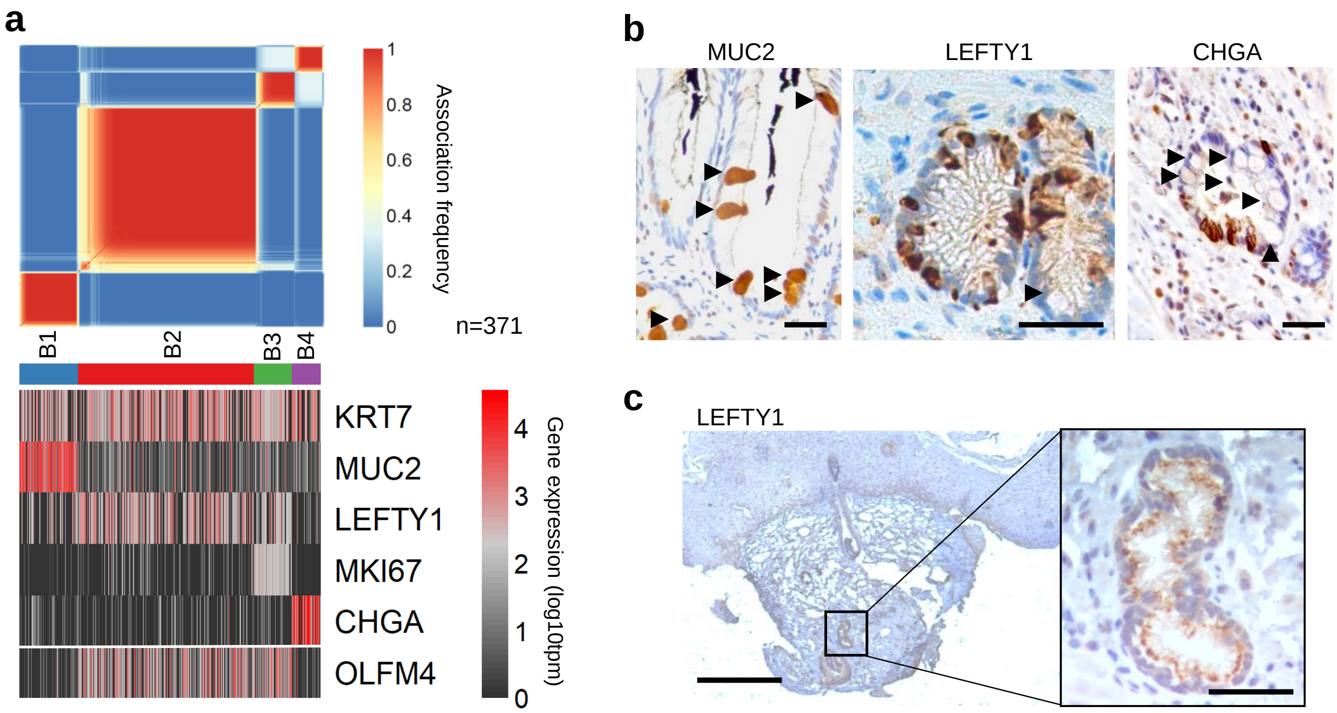
Figure 4. SPINK4 and ITLN1 mark early goblet cells

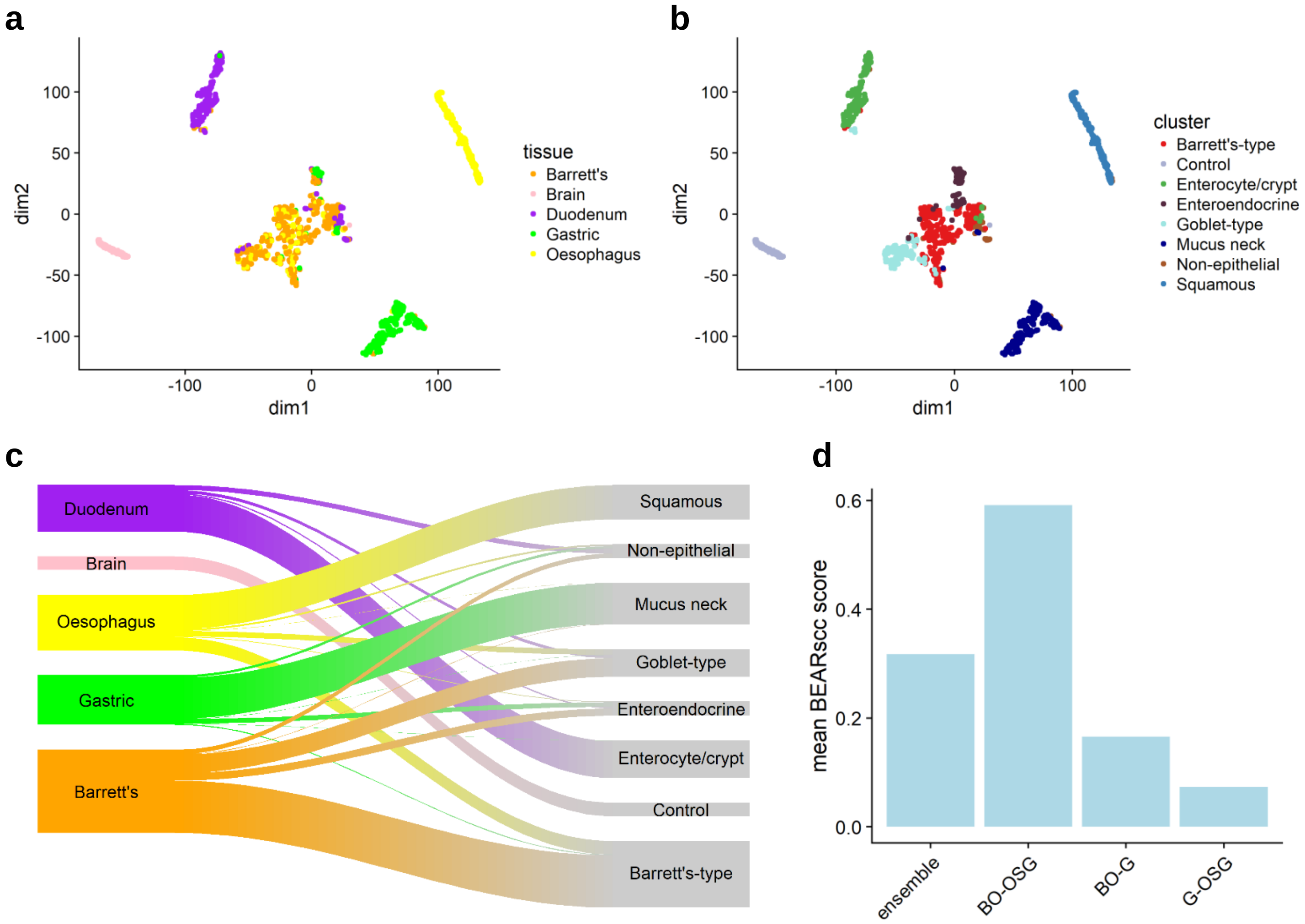
(a) Volcano plot showing fold change and p value of genes differentially expressed in the 'goblet-type' cell cluster as compared to all other cell clusters (see **Figure 3**). Points coloured red indicate genes significant at 5% permutation test. Selected highly significant genes are labelled. **(b)** Bar chart showing the percentage of cells in the 'goblet-type' cell cluster (n=98) expressing *MUC2*, *ITLN1* or *SPINK4* alone or in different combinations (thresholds set at the tenth centile to include 90% of cells in which at least one transcript was detected from each gene). **(c)** Triple immunofluorescence staining images of MUC2 (red), ITLN1 (white) and SPINK4 (green) in normal colon from a resection specimen (blue stain is DAPI). Scale bar 100µm. **(d)** Triple immunofluorescence staining images of MUC2 (red), ITLN1 (white) and SPINK4 (green) in normal oesophageal epithelium obtained by endoscopic biopsy (blue stain is DAPI). OSGs encroaching on the surface epithelium are shown in the enlarged images on the right. Scale bars are 200µm and 50µm in enlarged images. **(e)** Triple immunofluorescence staining images of KRT14 (white), KRT7 (green) and MUC2 (red) in an OSG beneath normal squamous epithelium from an endoscopic biopsy of normal squamous epithelium from a patient with BO biopsy (blue stain is DAPI). Scale bar 50µm. **(f)** Representative immunofluorescence staining of Barrett's EMR specimen containing intestinal metaplasia but no dysplasia for MUC2 (red), ITLN1 (white) and SPINK4 (green); nuclei (DAPI) in blue. Scale bars are 400µm and 100µm in enlarged images.

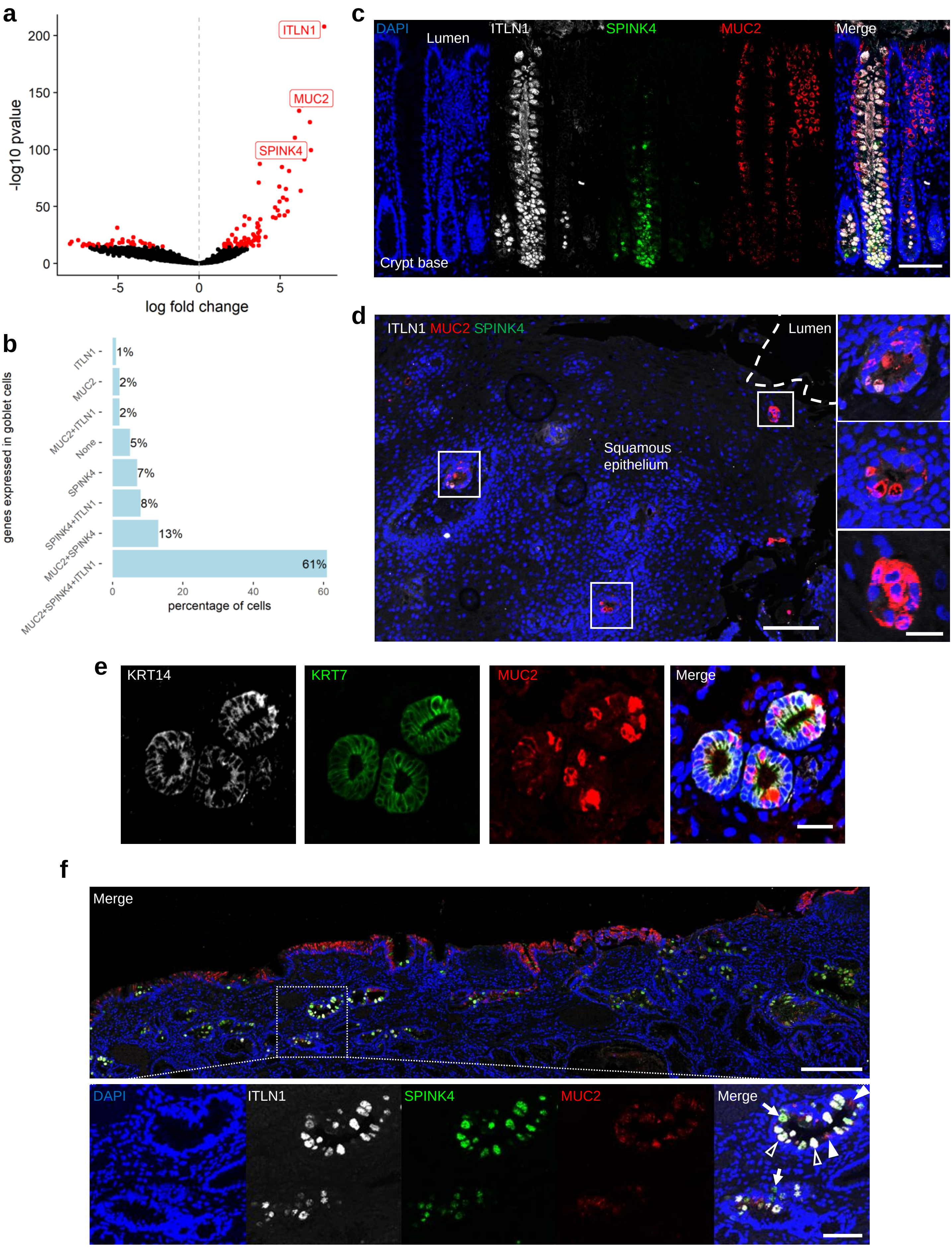
Figure 5. *OLFM4* is upregulated in BO and OSG cells with stem-like transcript profiles

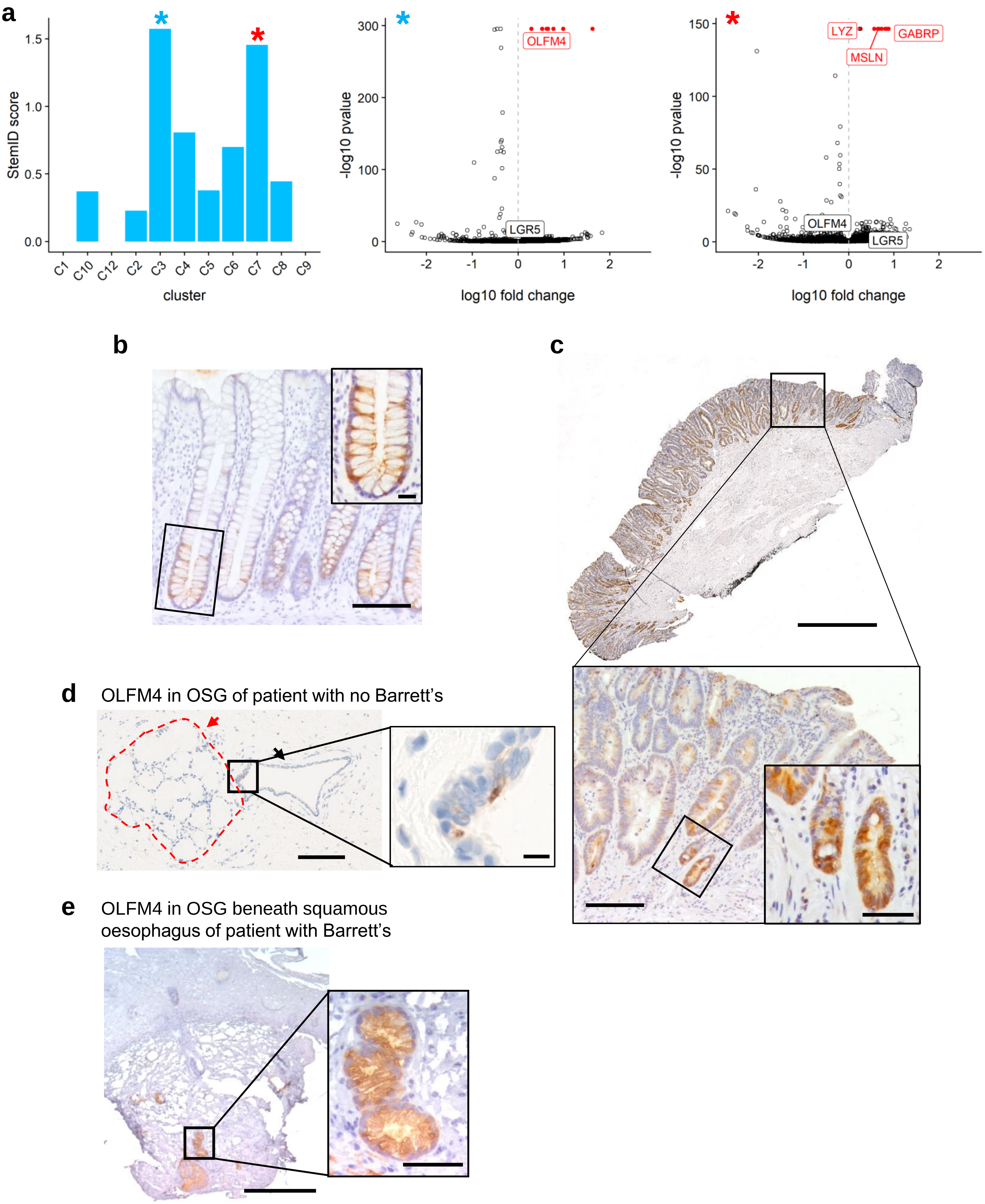
(a) Bar plot on left shows StemID scores across all RaceID2 clusters (see **Methods**) applied to all non-squamous oesophageal cells (BO and oesophageal cells with <5 KRT14 counts to exclude squamous cells, n=533). Scores are calculated from multiplication of the entropy (spread from the cluster mean) and the number of cluster links arising from a given cluster. Differentially expressed genes in the highest scoring cluster (C3, blue asterisk) and second highest scoring cluster (C7, red asterisk) are shown in the volcano plots in the centre and right plots, respectively. Points coloured red indicate the most significant genes with a fold change greater than 2. Selected highly significant genes are labelled. **(b)** Immunohistochemical staining of OLFM4 in human colon (close-up of base of crypt inset). Scale bars are 100µm and 20µm in inset. **(c)** Immunohistochemical staining of OLFM4 in BO mucosal resection containing intestinal metaplasia but no dysplasia, with enlarged image. Scale bars are 1000µm, 200µm in enlarged image and 50µm in inset. **(d)** Immunohistochemical staining of OLFM4 in OSG under normal oesophagus taken from the proximal part of an oesophagectomy specimen resected for Siewert type III junctional tumour in a patient with no BO. Red dashed area and arrow indicates OSG, black arrow indicates OSG duct. Scale bars are 300µm and 20µm in enlarged image. **(e)** Immunohistochemistry in OSGs from endoscopic biopsy of normal squamous oesophagus in patients with BO. Scale bars are 300µm and 50µm in enlarged image.


a
fragment 1
Histology
fragment 2
Pooled for bulk RNA sequencing
fragment 3
Cell suspension and sorting
x2-4
per site
b
patient
tissue
FDR < 1x10^22
c
n = 51
TFF1
TFF2
MUC1
MUC6
MUC17
MUC13
MUC3A
TFF3
MUC2
MUC5B
KRT7
MUC15
MUC22
TP63
MUC12
MUC16
KRT5
4
3
2
1
0
Gene expression (log10tpm)
patient
A
B
C
D
I
J
K
L
M
N
O
P
Q
tissue
Barrett's
Duodenum
Gastric
Oesophagus
d
Duodenum
Gastric
Oesophagus
1
0.8
0.6
0.4
0.2
0
Association frequency
D1
D2
D3
D4
ALPI
MUC2
OLFM4
JCHAIN
G1
G2
G3
CHGA
GKN1
TFF1
MUC6
O1
O2
O3
O4
KRT5
KRT14
TP63
SAA1
KRT7
TFF3
4
3
2
1
0
log tpm
e
H&E
f
KRT14
TFF3
KRT7


a
1
0.8
0.6
0.4
0.2
0
Association frequency
n=371
B1
B2
B3
B4
KRT7
MUC2
LEFTY1
MKI67
CHGA
OLFM4
4
3
2
1
0
Gene expression (log10tpm)
b
MUC2
LEFTY1
CHGA
c
LEFTY1


a
dim2
100
50
0
-50
-100
-100
0
100
dim1
tissue
Barrett's
Brain
Duodenum
Gastric
Oesophagus
b
dim2
100
50
0
-50
-100
-100
0
100
dim1
cluster
Barrett's-type
Control
Enterocyte/crypt
Enteroendocrine
Goblet-type
Mucus neck
Non-epithelial
Squamous
c
Duodenum
Brain
Oesophagus
Gastric
Barrett's
Squamous
Non-epithelial
Mucus neck
Goblet-type
Enteroendocrine
Enterocyte/crypt
Control
Barrett's-type
d
mean BEARscc score
0.6
0.4
0.2
0.0
ensemble
BO-OSG
BO-G
G-OSG


a
200
150
100
50
0
-log10 pvalue
ITLN1
MUC2
SPINK4
-5
0
5
log fold change
b
genes expressed in goblet cells
ITLN1
MUC2
MUC2+ITLN1
None
SPINK4
SPINK4+ITLN1
MUC2+SPINK4
MUC2+SPINK4+ITLN1
1%
2%
2%
5%
7%
8%
13%
61%
0
20
40
60
percentage of cells
c
DAPI
Lumen
ITLN1
SPINK4
MUC2
Merge
Crypt base
d
ITLN1 MUC2 SPINK4
Lumen
Squamous epithelium
e
KRT14
KRT7
MUC2
Merge
f
Merge
DAPI
ITLN1
SPINK4
MUC2
Merge


d OLFM4 in OSG of patient with no Barrett's

e OLFM4 in OSG beneath squamous oesophagus of patient with Barrett's