RESEARCH ARTICLE

CANCER GENOMICS

# Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition

Karol Nowicki-Osuch[1]†‡, Lizhe Zhuang[1]†, Sriganesh Jammula[2], Christopher W. Bleaney[3], Krishnaa T. Mahbubani[4,5], Ginny Devonshire[2], Annalise Katz-Summercorn[1], Nils Eling[2,6], Anna Wilbrey-Clark[7], Elo Madissoon[7], John Gamble[4,8], Massimiliano Di Pietro[1], Maria O'Donovan[1], Kerstin B. Meyer[7], Kourosh Saeb-Parsy[4,8], Andrew D. Sharrocks[3], Sarah A. Teichmann[7,9], John C. Marioni[2,6,7], Rebecca C. Fitzgerald[1]*

The origin of human metaplastic states and their propensity for cancer is poorly understood. Barrett's esophagus is a common metaplastic condition that increases the risk for esophageal adenocarcinoma, and its cellular origin is enigmatic. To address this, we harvested tissues spanning the gastroesophageal junction from healthy and diseased donors, including isolation of esophageal submucosal glands. A combination of single-cell transcriptomic profiling, in silico lineage tracing from methylation, open chromatin and somatic mutation analyses, and functional studies in organoid models showed that Barrett's esophagus originates from gastric cardia through c-MYC and HNF4A-driven transcriptional programs. Furthermore, our data indicate that esophageal adenocarcinoma likely arises from undifferentiated Barrett's esophagus cell types even in the absence of a pathologically identifiable metaplastic precursor, illuminating early detection strategies.

Metaplasia is usually associated with an increased risk of malignancy and is thought to result from a transcriptional switch within existing cells or an outgrowth of minor cell types, yet its specific origin is often unclear (*1*). It commonly occurs at sites where different epithelial types meet, such as squamo-columnar junctions (SCJ) (*2*). Barrett's esophagus (BE) is an archetypal metaplastic condition comprising a mosaic of gastric and intestinal cell types (*3*, *4*). BE occurs in up to 10% of individuals with gastric reflux and starts at the gastroesophageal junction (GEJ), causing the SCJ to be displaced proximally (*5*). It increases the propensity for esophageal adenocarcinoma (EAC), which has an overall 5-year survival rate of 15% (*6*, *7*). Given that the native esophagus is squamous, the glandular phenotype of EAC is thought to be inextricably linked to BE (*6*). However, about 50% of EAC patients do not have evidence of BE at the time of diagnosis (*8*), calling this current dogma into question.

The controversy could be resolved by determining the origin of BE, which has been hypothesized to originate from many sources, including esophageal submucosal glands (SMG) or various specific cell populations at the GEJ (*9*–*16*) (fig. S1). One major impediment to research is that mouse models used for lineage tracing do not fully resemble human gastroesophageal physiology owing to a keratinized squamous forestomach and a lack of SMG. Furthermore, isolation of SMG is particularly challenging in fresh human tissue. The overall aims of this study were therefore to molecularly characterize all putative cell origins for BE and determine whether all EAC subtypes are derived from BE.

## Results

### *Determining the cellular components of human SMG*

SMG have been studied as the origin of BE, yet never specifically isolated. Using stereomicroscopy (fig. S2 and movie S1), we isolated fresh SMGs from human esophagus. We performed immunostaining with cell markers of CDH1 (pan-epithelial), KRT5 (squamous), KRT8 (columnar), and KRT7 (SMG) (*17*) and used three-dimensional (3D) confocal microscopy to identify all the cellular components: duct cells, oncocytes, mucous cells, and myoepithelial cells (Fig. 1A, fig. S3, and movies S2 and S3) (*18*).

In both the intercalated and main duct of SMG, we observed a P63$^+$KRT5$^+$KRT7$^+$ cell population (Fig. 1B) that resembles the transitional basal progenitors, previously reported to generate BE and reside on the surface of SCJ (*15*, *18*). We also found that oncocytes, characterized by eosinophilic cytoplasm and centrally located nuclei, are common in disease-free donors and sometimes form their own acini (fig. S4, A and B). This is contrary to the literature, which suggests that they are associated with gastroesophageal reflux disease and BE (*19*).

Next, to characterize all cell types within the SMG, we dissociated fresh SMG into single cells and performed single-cell RNA sequencing (scRNA-seq) (*20*). We identified four major epithelial cell types (Fig. 1, C and D, and tables S1 and S2) expressing transcripts previously not associated with SMG, including *MUC5B*, *KRT23*, and *AGR2*, which were confirmed by immunostaining (fig. S4C). Few (<0.1%) SMG cells express the cell division marker *MKI67*, indicating their quiescent status, and contrary to a previous scRNA-seq study (*11*), no significant *OLFM4* was detected (Fig. 1D and fig. S5A). A comparison of our SMG profile with esophageal cells from the Human Cell Atlas (HCA) project (*21*) and a recent SMG study (*11*) demonstrated the sparse nature of SMG in esophageal biopsies. Only a very small fraction of cells in the HCA project, and none of the cells from the previous SMG study (*11*), showed SMG phenotypes (fig. S5, B to E) (*20*).

### *A KRT7$^{high}$ population of SCJ and its lineage relationship with SMG*

The transitional basal cells and residual embryonic cells of the GEJ have also been proposed as the origin for BE (*14*, *15*). We collected normal squamo-columnar junction (N-SCJ) tissue from eight disease-free donors (fig. S6) and performed scRNA-seq to study its cellular composition. In addition to the expected esophageal squamous and gastric columnar cells, we discovered a small distinct *KRT7$^{high}$* population (Fig. 2, A and B), which consisted of P63$^+$KRT5$^+$KRT7$^+$ transitional basal cells (C1 and C2), KRT7$^+$MUC4$^+$ residual embryonic cells (C3), and a *MUC5B$^{high}$* cell type (C4) (Fig. 2, C and D, and table S3). Immunostaining showed that MUC5B is expressed in the suprabasal layer and is in some cases independent from the transitional basal P63$^+$KRT5$^+$KRT7$^+$ cells (Fig. 2, E and F).

When the single-cell transcriptome profiles of N-SCJ cells were compared to those of cells from normal esophagus (NE), normal gastric cardia (NG) (*22*), and SMG from disease-free donors, the *KRT7$^{high}$* population from N-SCJ exhibited the highest similarity to SMG cells with respect to cell types present in our reference set (Figs. 2, G and H, and fig. S7A). Overall, this suggests that the transitional basal progenitors, residual embryonic cells, and the *MUC5B$^{high}$* cells might belong to a single

[1]Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge CB2 0X2, UK. [2]Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK. [3]Faculty of Biology, Medicine and Health, Michael Smith Building, Oxford Road, University of Manchester, Manchester, UK. [4]Cambridge Biorepository for Translational Medicine (CBTM), NIHR Cambridge Biomedical Research Centre, Cambridge, UK. [5]Department of Haematology, University of Cambridge, Cambridge, UK. [6]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [7]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [8]Department of Surgery, University of Cambridge, Cambridge, UK. [9]Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0HE, UK.
*Corresponding author. Email: rcf29@cam.ac.uk
†These authors contributed equally to this work.
‡Present address: Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA.

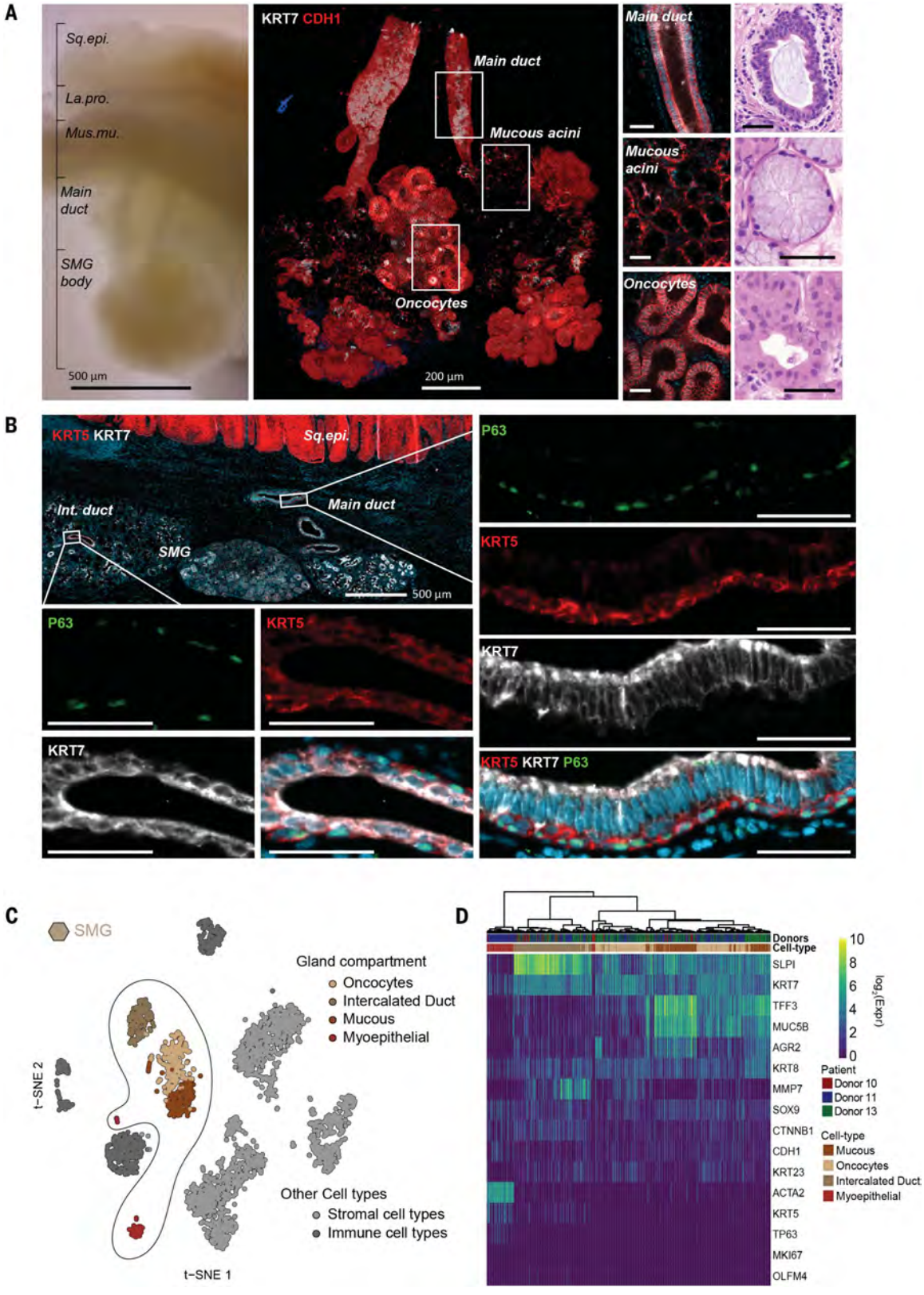


**Fig. 1. Characterization and single-cell profile of human SMG.** (**A**) Whole-mount SMG from disease-free donors, with esophagus squamous epithelium (Sq. epi.), lamina propria (La. pro.), muscularis mucosa (Mus. mu.), duct, and the SMG body. SMG were imaged by confocal microscopy; duct, mucous cells, and oncocytes were confirmed by hematoxylin and eosin (H&E) staining. (**B**) Immunofluorescence of SMG sections. Intercalated (Int.) and main duct regions of SMG were magnified to show the $P63^{+}KRT5^{+}KRT7^{+}$ cells. (**C**) t-SNE (t-distributed stochastic neighbor embedding) projection of scRNA-seq of SMG cells from three disease-free donors. Epithelial cells are circled. (**D**) Heatmap of expression of known SMG marker genes. Only epithelial cells of SMG are shown. For (A) and (B), scale bars are 50 µm unless otherwise indicated.

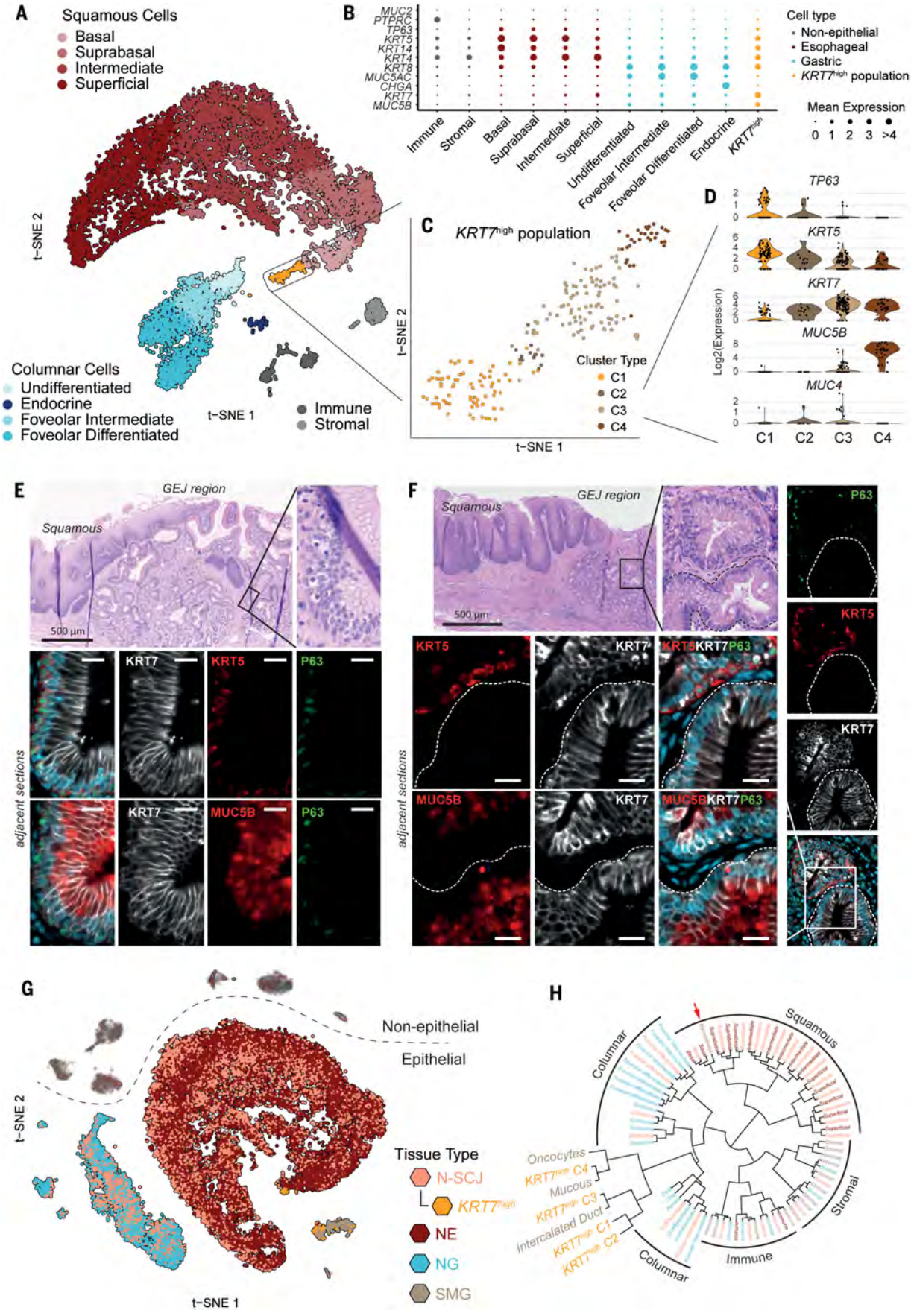

**Fig. 2. A *KRT7*$^{high}$ epithelial population at the GEJ region.** (**A**) t-SNE projection of scRNA-seq of cells from N-SCJ. The *KRT7*$^{high}$ population is highlighted in yellow. (**B**) Bubble plot of selected marker genes relevant to squamous, gastric, and SMG phenotypes. (**C**) t-SNE projection and clustering results of scRNA-seq of the *KRT7*$^{high}$ population. (**D**) Violin plots of gene expression in C1 to C4 cell populations. Marker genes for the transitional basal progenitor cells (P63$^+$KRT5$^+$KRT7$^+$) and residual embryonic cells (KRT7$^+$MUC4$^+$) are shown. (**E** and **F**) H&E staining and immunofluorescence of disease-free donors. H&E confirmed that these cells are not located at SMG but at the GEJ. Adjacent sections were co-stained for KRT5/KRT7/P63 and MUC5B/KRT7/P63, respectively. Note the single-layered MUC5B$^+$ epithelium (marked by a dashed line and corresponding to C4 of the *KRT7*$^{high}$ population). Scale bars are 20 μm unless otherwise indicated. (**G**) t-SNE projections of scRNA-seq of cells from N-SCJ, NE, NG, and SMG. Color denotes tissue types; epithelial cell types are shown in full color, and nonepithelial cell types are shaded out. (**H**) Similarity (Euclidean distance) between cell clusters from all of the tissue types. Clusters C1 to C4 and SMG cell types are highlighted with bigger font. The red arrow indicates myoepithelial cells. Font color denotes tissue types, same as in (G).

*KRT7*$^{high}$ lineage that are at different stages of differentiation, and it is possible that they share the same origin with SMG (fig. S7B).

### *scRNA-seq profiling of BE and normal gastroesophageal tissues*

Having established the phenotypes of SMG and N-SCJ regions, we expanded scRNA-seq analysis to 43,000 cells from 43 samples collected across seven different tissue sites from 14 donors and BE patients (Fig. 3A, fig. S8, and table S1). All major cell types in NE, NG, and ND (normal duodenum, which served as an intestinal cell reference for BE)—including squamous basal and superficial cells, gastric foveolar, endocrine, parietal, and chief cells—were successfully identified (fig. S9, A to F, and tables S4 to S6). Next, we analyzed BE samples, including the BE squamo-columnar junction (B-SCJ; fig. S6B) (*23*). Within BE, we identified columnar cells resembling gastric foveolar cells (*MUC5AC*, *KRT20*), goblet cells (*MUC2*, *TFF3*), and previously uncharacterized poorly differentiated enteroendocrine-like cells (*NEUROG3*, *CHGA*) (fig. S10 and table S7). Furthermore, we identified an undifferentiated BE cell type characterized by lack of differentiation features but showing *OLFM4* expression, a marker of intestinal and BE stem cells (*11*, *24*). This undifferentiated cell type was also confirmed by intercellular transcriptional variability (also referred to as entropy) (*25*) (fig. S11).

To reconstruct the lineage relations of all cell types, we used hierarchical clustering of cell type–specific consensus transcriptomes. Except for goblet cells, BE cells (from both B-SCJ and BE) showed the strongest similarities to gastric cardia cells (from both NG and N-SCJ). This was consistent through all stages of differentiation for gastric and BE cells (Fig. 3B and fig. S12). In addition, BE cells are distinct from any of the SMG cell types and the N-SCJ *KRT7*$^{high}$ clusters. We did not observe *KRT7*$^{high}$ or any intermediate cell populations in the B-SCJ samples, suggesting that cell transdifferentiation from NE to BE is unlikely (fig. S13) (*23*).

### *Genetic and epigenetic similarities between BE and gastric cardia*

In view of the observation that BE cell types resemble the transcriptional profiles of their NG counterparts, we used three lineage-tracing methods (*26*–*28*) to further investigate this. First, methylation profiles were generated from fresh samples of NE, NG, SMG, and BE. Principal components analysis from the top 10,000 most variable probes confirmed that NE and SMG share similar methylation profiles, suggesting a close embryological origin, and have distinctive features not detected in BE samples. BE methylation profiles resembled NG and did not overlap with NE or SMG (Fig. 3C).

Second, we studied lineage relatedness using ATAC-seq (assay for transposase-accessible chromatin with sequencing) that enabled profiling of the open chromatin landscape of tissues from two independent cohorts. We identified the top 10,000 most variable peaks of open chromatin across all the tissue types and performed hierarchical clustering of normalized reads in peaks. The results were concordant with the methylation profiling—NG and BE share the most similar features of chromatin accessibility, whereas SMG clustered with NE (Fig. 3D).

Third, we interrogated the 60× whole-genome sequencing (WGS) data of matched endoscopic biopsies from five patients (SMG not included) (*18*) to identify any spontaneously accumulated somatic mutations that were shared between BE and its putative tissue of origin (*29*). To do this, we identified clonal mutations in BE and then determined their presence in NE or NG (Fig. 3E). We found such mutations in four out of five patients, which were all shared between BE and NG (Fig. 3F and fig. S14). The results were further confirmed by targeted Sanger sequencing in a separate set of tissues from the same patients (fig. S14).

### *c-MYC and HNF4A drive the transcriptional programs of BE cells*

To investigate the transcriptional mechanisms underlying the development of BE, we performed gene set enrichment analysis (GSEA) (*30*) and causal analysis (*31*) of genes differentially expressed between different stages of BE and NG differentiation (Fig. 4, A and B). We reasoned that the differences between the undifferentiated cells of each tissue would be key for BE development. Differential analysis of NG and BE undifferentiated phenotypes identified activation of *CDX1*, *CDX2*, and *MYC* modules in BE (Fig. 4C). Further comparison of BE columnar differentiated cells with undifferentiated cells showed activation of the transcription factor *HNF4A* in the former and *MYC* in the latter (Fig. 4D and fig. S15). These results were corroborated by bulk ATAC-seq analysis. A comparison of regions that are more accessible in BE versus NG tissues identified binding motifs for a variety of transcription factors associated with gastrointestinal development enriched in BE. This included *CDX2* and *HNF4A*, further confirming the scRNA-seq analysis (fig. S16, A and B) (*32*).

At the protein level, c-MYC was detected in 61 out of 66 BE biopsies (25 patients) containing intestinal metaplasia, with expression confined to the bottom of the crypts (Fig. 4E). There was no c-MYC expression in 64 normal gastric biopsies (28 patients). Similarly, all 66 BE biopsies and none of the 64 gastric biopsies were positive for HNF4A. In contrast to c-MYC expression, the expression of HNF4A was present in the top two-thirds of the BE crypts (Fig. 4E).

To elucidate the functional relevance of these findings, we established organoid cultures from disease-free NG tissues and transduced them with exogenous c-MYC and HNF4A marked by green fluorescent protein (GFP) labeling (Fig. 4, F and G). Because NG and BE organoids are morphologically indistinguishable with no identifiable goblet cells, a gene panel was used (Fig. 4H) to evaluate the phenotype of transduced NG organoids. Ectopic expression of c-MYC led to an up-regulation of genes characteristic of undifferentiated BE cells, including *NCL* and *LEFTY1*, whereas HNF4A led to up-regulation of a wider BE gene set comprising *FBP1*, *CLDN3*, *LEFTY1*, *TFF3*, and particularly *CDX2* (Fig. 4I and fig. S17A). ATAC-seq also showed increased accessibility at the promotor regions of *CLDN3* and *CDX2* in BE (fig. S16C). The regulatory roles of c-MYC and HNF4A were also confirmed in the normal gastric cell line HFE145 (fig. S17, B and C). These results demonstrate that ectopic expression of c-MYC and HNF4A in gastric cells drives expression of genes relevant for the BE phenotype.

### *EAC likely originates from undifferentiated BE cells regardless of its clinical features*

Subtypes of EAC based on the presence or absence of BE have different prognoses, raising

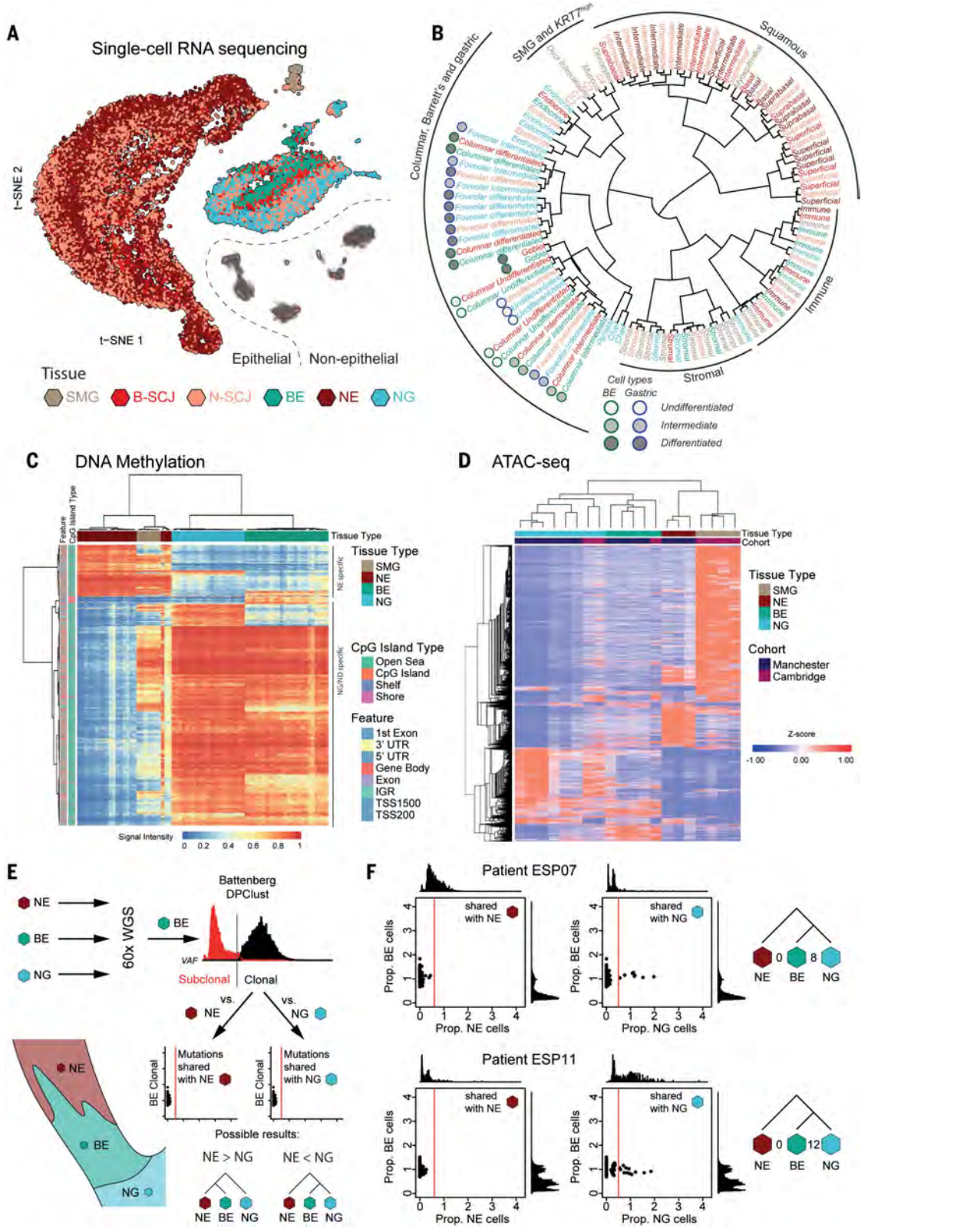


**Fig. 3. BE shows epigenetic, genetic, and transcriptional similarity with gastric cardia.** (**A**) t-SNE projection of scRNA-seq of cells from all tissue sites. (**B**) Similarity (Euclidean distance) between all the cell clusters. Font color denotes tissue sites, same as in (A); undifferentiated, intermediate, and differentiated cells of BE and NG are marked by green and blue circles, respectively. (**C**) Clustering of 10,000 highly variable probes of bulk methylation data from BE, NG, NE, and SMG. IGR, intergenic region; TSS, transcription start site; UTR, untranslated region. (**D**) Hierarchical clustering of the 10,000 most variable peaks of open chromatin regions in bulk ATAC-seq data. (**E**) Schematics of somatic mutation analysis and pairwise comparison from 60× WGS data. (**F**) Pairwise comparison of BE versus NE and BE versus NG in two BE patients; their relationship is indicated by the similarity plot based on shared somatic mutations.

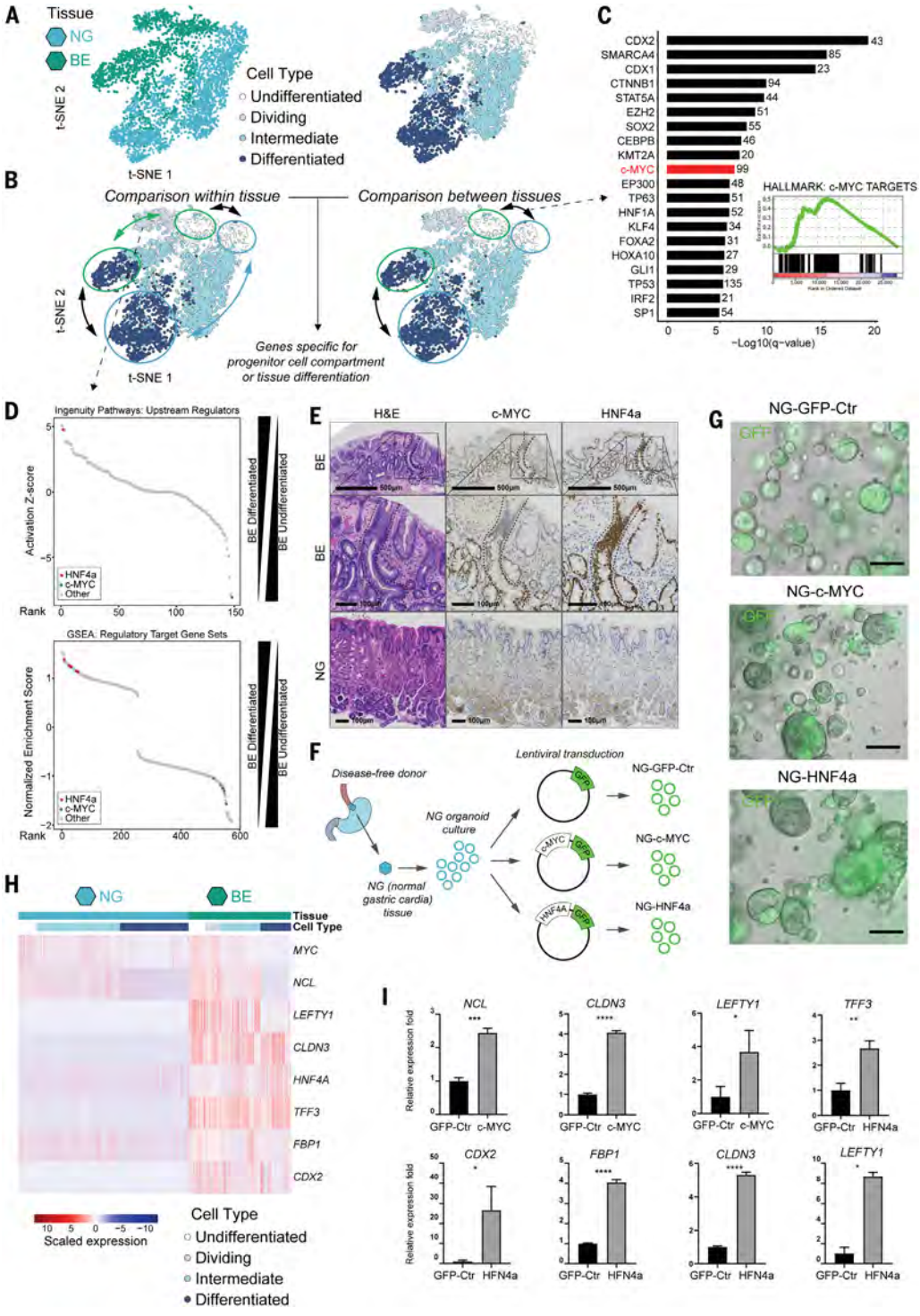


**Fig. 4. c-MYC and HNF4A drive the transcriptional programs of BE cells.** (**A**) t-SNE projection of scRNA-seq of selected epithelial cells from NG and BE with tissue sites (left) and cell subtype (right) overlaid. (**B**) Schematics of differential analysis between cell types used for identification of marker genes. (**C**) Causal analysis of pathways that are up-regulated when comparing BE undifferentiated and NG undifferentiated cells. The top 20 pathways with a positive *z*-score are shown; the number indicates the number of genes common between the dataset and the annotated pathway. The inset shows GSEA using Hallmark gene sets of BE undifferentiated and NG undifferentiated cells. (**D**) (Top) causal analysis of pathways up- and down-regulated when comparing BE undifferentiated and foveolar-like cells. Only pathways with a *q* value < 0.05 and a *z*-score of activation above 1 or below −1 are shown. (Bottom) GSEA using C3 gene sets database of BE undifferentiated and foveolar-like cells. (**E**) Representative immunohistochemical staining of MYC and HNF4A in BE and NG biopsies with an individual crypt highlighted. (**F**) Schematics of generation and transduction of NG organoids. (**G**) Fluorescence imaging of GFP indicating successful transduction of organoids. Ctr, control. (**H**) Marker genes highly expressed in BE cells versus NG. $Log_2$-transformed, normalized counts were *z*-score scaled per gene for visualization purposes. (**I**) Relative expression fold change of BE marker genes in transduced NG organoids upon overexpression of *c-MYC* or *HNF4A*. Error bars represent standard deviation. Statistical analysis: Paired *t* test; **p* < 0.05; ***p* < 0.01; ****p* < 0.001; *****p* < 0.0001.

the question as to whether all EAC arises from BE (*8*). Here, we took advantage of our scRNA-seq profiles to interrogate bulk RNA-seq data from esophageal cancer (*33*–*35*). MuSiC (multi-subject single cell deconvolution) analysis of an independent dataset of bulk RNA-seq correctly assigned scRNA-seq profiles to individual tissue types and was able to distinguish BE columnar cells and NG foveolar cells (figs. S18 and S19). Next, we compared the two phenotypically distinct types of esophageal cancer as confirmation of the method. As expected, bulk RNA-seq data of 81 ESCC (esophageal squamous cell carcinoma) and 321 EAC samples (*36*, *37*) assigned phenotypes correctly (Fig. 5, A and B). The EAC similarity to BE cells was independent of clinical features including TNM (tumor/node/metastasis) stage, the Siewert classification, and the presence of adjacent BE. The expression of differentiated BE cells (BE columnar and goblet

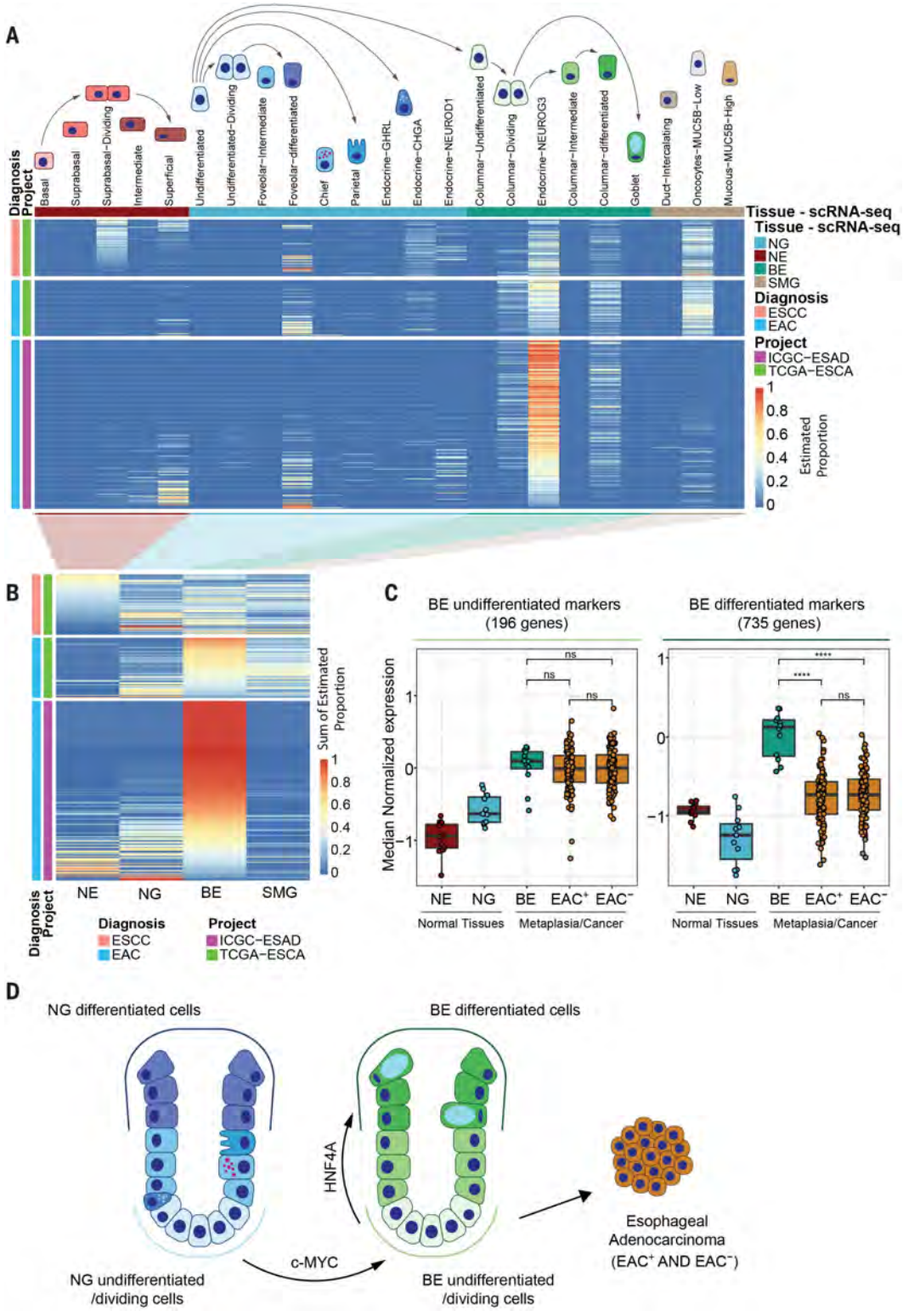


**Fig. 5. Deconvolution analyses of bulk human EAC and ESCC transcriptomes.** (**A**) Contribution of individual cell phenotypes originating from NE, BE, NG, and SMG to the phenotypes of esophageal cancers. The estimation was performed using MuSiC after removal of nonepithelial cell types. Cancer samples (rows) are sorted by the contribution of NE phenotype or BE phenotype for ESCC and EAC, respectively. ICGC-ESAD, International Cancer Genome Consortium–Esophageal Adenocarcinoma; TCGA-ESCA, The Cancer Genome Atlas–Esophageal Carcinoma. (**B**) Contribution of combined scRNA-seq–based tissue phenotypes to esophageal cancer phenotypes. Individual tissue contributions are sums of estimated proportions of individual cells from that tissue type. (**C**) Contribution of undifferentiated and endocrine-like (left) or differentiated (foveolar-like and goblet; right) phenotypes of BE to transcriptomes of NE, NG, BE, and EAC samples with or without adjacent BE ($EAC^+$ or $EAC^-$). Each data point represents a median normalized expression of BE undifferentiated or differentiated marker genes for individual samples. Individual $\log_2$ expression values were normalized to the mean $\log_2$ expression of the given gene in BE samples. Boxes indicate interquartile range (IQR), and whiskers extend to the value no further than 1.5 fold of IQR from the boxes. Statistical analysis: Unpaired Wilcoxon test; ****$p < 0.0001$; ns, not significant. (**D**) Schematic of EAC development. c-MYC and HNF4A drive transcriptional programs of BE from NG, and undifferentiated cells of BE serve as the precursor to EAC.

cell markers; tables S7 to S9) decreased during the transition from BE to EAC (Fig. 5C). Conversely, the undifferentiated BE phenotype (defined as a combination of genes specific to BE undifferentiated and endocrine-like cells; tables S9 and S10) was maintained during the BE-to-EAC transition (Fig. 5C). A corresponding analysis with differentiated and undifferentiated NG cell markers did not show any similarity with EAC (fig. S20). This suggests a unified origin of EAC from a metaplastic precursor, even in cases where BE metaplasia is not apparent at diagnosis or in the pathological specimen.

## Discussion

Our results suggest that undifferentiated gastric cells from the cardia give rise to BE via transcriptional programs driven by c-MYC and HNF4A (Fig. 5D). Furthermore, transcriptional profiling of EAC showed expression of markers found in undifferentiated BE cells, regardless of whether metaplasia precursors were identifiable histologically.

The fierce debate over the origin of BE can be attributed to reliance on mouse models and limited human samples. For example, both the residual embryonic cell (*14*) and transitional basal cell (*15*) hypotheses did not consider human SMG and interpreted expansion of $KRT7^+$ cells at the mouse gastroesophageal region as indicative of BE development, yet our data demonstrate that $KRT7^+$ cells are common in the BE-free SCJ and SMG. Although caution is required before treating the mouse metaplasia-like phenotype as equivalent to clinical BE, mouse models remain a useful tool to investigate environmental cues and EAC pathogenesis (*13*, *38*, *39*). Studies of SMG have been limited by difficulty in its isolation. In formalin-fixed tissue sections, Leedham and colleagues found a single BE gland that shared the same mutational profile with the adjacent SMG duct (*9*). Recently, Owen and colleagues used scRNA-seq to profile cells from human endoscopic pinch biopsies of NE, NG, and BE (*11*) and deduced that BE arose from SMG, but without isolating the SMG specifically.

c-MYC and HNF4A have been studied in the BE context (*40*, *41*), but a NE origin was always assumed, and we demonstrated that NG should be the control tissue. Although BE and EAC were previously studied with gene expression microarrays in bulk tissues (*41*, *42*), the genesis of EAC had been difficult to address before single-cell profiling techniques. Our results suggest that EAC may derive from gastric cells via a BE-like metaplasia, even though the evolutionary trajectory and prognosis can vary between patients (*8*). Moreover, it is possible that intestinal metaplasia phenotypes, such as BE or gastric intestinal metaplasia, are an indispensable step toward neoplasia. This is in keeping with a TCGA (The Cancer Genome Atlas) study that concluded that EAC falls within a spectrum of gastroesophageal adenocarcinomas (*37*).

The strengths of this study include comprehensive multi-omic profiling of freshly isolated human cells from superficial to submucosal compartments across the GEJ in healthy and diseased individuals, which has been challenging to achieve in human. However, there are limitations to this study. Transcriptional similarity does not prove causality, and it is possible that minuscule cell populations were undetected or lost during tissue preparation. Single-cell–based, deep somatic lineage tracing with paired scRNA-seq of human samples will be informative to address these limitations when the technologies are mature.

This study provides several orthogonal lines of evidence for a gastric origin for BE, which is likely to be a requisite step for neoplastic progression at this site. We hope that these data will pave the way for further research and clinical early detection and cancer prevention strategies.

## REFERENCES AND NOTES


1. J. M. Slack, *Lancet* **328**, 268–271 (1986).
2. M. Herfs *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10516–10521 (2012).
3. S. A. C. McDonald, T. A. Graham, D. L. Lavery, N. A. Wright, M. Jansen, *Cell. Mol. Gastroenterol. Hepatol.* **1**, 41–54 (2014).
4. Y. Peters *et al.*, *Nat. Rev. Dis. Primers* **5**, 35 (2019).
5. E. C. Lin, J. Holub, D. Lieberman, C. Hur, *Clin. Gastroenterol. Hepatol.* **17**, 857–863 (2019).
6. E. C. Smyth *et al.*, *Nat. Rev. Dis. Primers* **3**, 17048 (2017).
7. M. Arnold *et al.*, *Lancet Oncol.* **20**, 1493–1505 (2019).
8. T. Sawas *et al.*, *Gastroenterology* **155**, 1720–1728.e4 (2018).
9. S. J. Leedham *et al.*, *Gut* **57**, 1041–1048 (2008).
10. R. J. von Furstenberg *et al.*, *Cell. Mol. Gastroenterol. Hepatol.* **4**, 385–404 (2017).
11. R. P. Owen *et al.*, *Nat. Commun.* **9**, 4261 (2018).
12. Y. Hu *et al.*, *J. Gastrointest. Surg.* **11**, 827–834 (2007).
13. M. Quante *et al.*, *Cancer Cell* **21**, 36–51 (2012).
14. X. Wang *et al.*, *Cell* **145**, 1023–1035 (2011).
15. M. Jiang *et al.*, *Nature* **550**, 529–533 (2017).
16. L. Hutchinson *et al.*, *Stem Cells Dev.* **20**, 11–17 (2011).
17. G. Gonzalez, Q. Huang, H. Mashimo, *Dis. Esophagus* **29**, 670–680 (2016).
18. See supplementary text 1.
19. D. R. Braxton, D. C. Nickleach, Y. Liu, A. B. Farris III, *Virchows Arch.* **465**, 135–143 (2014).
20. See supplementary text 2.
21. E. Madissoon *et al.*, *Genome Biol.* **21**, 1 (2019).
22. See supplementary text 3.
23. See supplementary text 4.
24. L. G. van der Flier, A. Haegebarth, D. E. Stange, M. van de Wetering, H. Clevers, *Gastroenterology* **137**, 15–17 (2009).
25. D. Grün *et al.*, *Cell Stem Cell* **19**, 266–277 (2016).
26. K. A. Hoadley *et al.*, *Cell* **173**, 291–304.e6 (2018).
27. A. B. Stergachis *et al.*, *Cell* **154**, 888–903 (2013).
28. S. Nik-Zainal *et al.*, *Cell* **149**, 979–993 (2012).
29. See supplementary text 5.
30. A. Subramanian *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).
31. A. Krämer, J. Green, J. Pollard Jr., S. Tugendreich, *Bioinformatics* **30**, 523–530 (2014).
32. See supplementary text 6.
33. M. C. Vladoiu *et al.*, *Nature* **572**, 67–73 (2019).
34. M. D. Young *et al.*, *Science* **361**, 594–599 (2018).
35. See supplementary text 7.
36. A. M. Frankell *et al.*, *Nat. Genet.* **51**, 506–516 (2019).
37. The Cancer Genome Atlas Research Network, *Nature* **541**, 169–175 (2017).
38. N. S. Münch *et al.*, *Gastroenterology* **157**, 492–506.e2 (2019).
39. D. J. Fu *et al.*, *Nat. Commun.* **11**, 84 (2020).
40. S. Wang *et al.*, *Oncogene* **25**, 3346–3356 (2006).
41. C. Rogerson *et al.*, *Genome Res.* **29**, 723–736 (2019).
42. D. B. Stairs *et al.*, *PLOS ONE* **3**, e3534 (2008).
43. S. Jammula *et al.*, *Gastroenterology* **158**, 1682–1697.e1 (2020).
44. E. Britton *et al.*, *PLOS Genet.* **13**, e1006879 (2017).
45. K. Nowicki-Osuch, Supporting code for "Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition." Zenodo (2021); https://doi.org/10.5281/zenodo.4740422.


## ACKNOWLEDGMENTS

We thank all patients, organ donors, and their families who contributed to this study. We thank Cambridge Biorepository for Translational Medicine for collecting deceased organ donor samples. We thank the members of the OCCAMS consortium for the recruitment of patients and provision of samples for data generation. We thank P. Coupland, K. Kania, and CRUK Cambridge Institute Core Genomics facility for scRNA-seq and the Genomic Technologies Core Facility, University of Manchester, for ATAC-sequencing. We thank J. Rogan and E. Cheadle for their assistance in accessing the MCRC Biobank tissues. We acknowledge support from the Human Research Tissue Bank, Cambridge University Hospitals NHS Foundation Trust. We thank D. Shorthouse for critical assessment of the manuscript. **Funding:** This work was supported by grants from the Medical Research Council (RG84369) and CRUK (RG66287 and RG81771/84119) to R.C.F., grants from CRUK (C9545/A29580) to J.C.M., grants from Wellcome Trust (206194) to S.A.T., grants from Chan Zuckerberg Initiative (174169) to K.B.M., and grants from NIHR Cambridge BRC (RG92051) to K.S.-P. C.W.B. was supported by a CRUK-funded clinical training PhD studentship. K.T.M. was supported by a Chan Zuckerberg Initiative Award. A.K.-S. was supported by a Cambridge Cancer Centre Clinical Research Fellowship. **Author contributions:** K.N.-O. and L.Z. conceived and performed the experiments, performed the analysis, and prepared the manuscript and figures. S.J. analyzed the methylation data. C.W.B. performed ATAC-seq and analyzed the data. K.T.M., J.G., and K.S.-P. collected and managed the deceased organ donor samples. G.D. performed alignment and preliminary analysis of WGS data. A.K.-S. collected bulk RNA-seq data from BE and normal samples. N.E. designed and performed the scRNA-seq analysis and participated in preparing the earlier versions of the manuscript. A.W.-C. and E.M. collected and provided the HCA scRNA-seq data. M.D.P. collected endoscopic samples and provided guidance on other sample collections. M.O. supervised surgical samples collection and provided guidance on other sample collections. K.B.M. conceived and supervised the scRNA-seq analysis. A.D.S. conceived and supervised the analysis of ATAC-seq data. S.A.T. conceived the analysis and provided the HCA scRNA-seq data. J.C.M. conceived and supervised the analysis. R.C.F. conceived the experiments and analysis, supervised all research, and wrote the manuscript. **Competing interests:** R.C.F. holds patents related to Cytosponge-TFF3 and related assays that have been licensed by the Medical Research Council to Covidien (now Medtronic). R.C.F. is a co-founder and shareholder in an early detection and digital pathology company Cyted Ltd. The authors declare no other conflicts of interest. **Data and materials availability:** scRNA-seq data are available from the European Genome-phenome Archive (EGA) (EGAD00001005438), www.esophaguscancercellatlas.org (username: cellatlas; password: Barrett2021), and the HCA project (PRJEB31843). Bulk RNA-seq data are available from the EGA (EGAD00001006882 for BE, NE, and NG; and EGAD00001005388 for EAC). WGS data for BE, NE, NG, and ND samples are available from the EGA (EGAD00001006874). Methylation data for SMG samples are available from the EGA (EGAD00010001795), and other tissues were previously published (*43*). ATAC-seq data are available at ArrayExpress (E-MTAB-6751 and E-MTAB-5169) and were previously published (*44*). Cancer RNA-seq data were previously published (*36*, *37*). All code is deposited in Zenodo (*45*).

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6556/760/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S24
Tables S1 to S12
References (*46*–*91*)
MDAR Reproducibility Checklist
Movies S1 to S3

View/request a protocol for this paper from *Bio-protocol*.

11 June 2020; resubmitted 26 January 2021
Accepted 7 June 2021
10.1126/science.abd1449

# Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition

Karol Nowicki-Osuch, Lizhe Zhuang, Sriganesh Jammula, Christopher W. Bleaney, Krishnaa T. Mahbubani, Ginny Devonshire, Annalise Katz-Summercorn, Nils Eling, Anna Wilbrey-Clark, Elo Madissoon, John Gamble, Massimiliano Di Pietro, Maria O'Donovan, Kerstin B. Meyer, Kourosh Saeb-Parsy, Andrew D. Sharrocks, Sarah A. Teichmann, John C. Marioni and Rebecca C. Fitzgerald

*Science* **373** (6556), 760-767.
DOI: 10.1126/science.abd1449

**Identifying the origin of cancer**

Many cancers are classified on the basis of the organ or tissue from which they originated. However, identifying the specific cells and conditions that precede tumorigenesis can help us understand and better treat the resulting disease. Nowicki-Osuch *et al.* used a single-cell approach to investigate the cell of origin for Barrett's esophagus (BE) and the mechanisms leading to the development of esophageal adenocarcinoma (EAC) (see the Perspective by Geboes and Hoorens). Analyses of healthy human esophageal tissues, mutational lineage tracing, and organoid models revealed that BE originates from the gastric cardia and that EAC arises from undifferentiated BE cells. This analysis provides a map of the transcriptional landscape of the healthy esophagus that can be compared with mouse models of disease.

*Science*, abd1449, this issue p. 760; see also abj9797, p. 737