Supplementary Figure, Table Legend, and Methods

Supplementary Figure 1: Hematoxylin-eosin staining (HE) and EBV infection detection using Paraffin embedded tissue section in all 9 enrolled cancer patients.

Supplementary Figure 2: tSNE plot of the 27,677 cells profiled here, with each cell color-coded for (left to right): the number of genes (nGene) and the number of transcripts (nUMI) detected in each cell.

Supplementary Figure 3: (A) Volcano plot showing differentially expressed genes (DEGs) across paired normal and tumor samples of gastric cancer in TCGA dataset (B) Heat map showing differentially expressed genes (DEGs) across paired normal and tumor samples of gastric cancer in TCGA dataset. (C) Expression of representative tumor associated and normal associated marker genes in TCGA samples showed using bar chart. (D) The percentage of non-malignant cells and malignant cells in all epithelial cells across surgical resected samples and targeted biopsies in scRNA-seq data.

Supplementary Figure 4: (A) Bar chart showing the CNV signaling among each sample. (B) tSNE plot of colored by CNV signaling of all 10,411 epithelial cells. (C) Heatmap showing TCGA STAD CNV estimated using the GISTIC2 method (Downloaded from https://xenabrowser.net/). (D) Heatmap showing TCGA STAD thresholded value of estimated CNV from GISTIC2 (Downloaded from https://xenabrowser.net/).

Supplementary Figure 5: (A) Violin plots showing the expression of MHC-II genes in chief cells, SPEM, neck cells, and surface cell (B) Pseudo-temporal expression dynamics of specific representative genes between chief cells, SPEM and neck cells.

Supplementary Figure 6: Heterogeneous expression of KRT20 in two intestinal-histology samples (IGC1 and IGC4).

Supplementary Figure 7: Expression of gastric-phenotype marker Claudin18 and intestinal or differentiation marker KRT20 in diffuse- and intestinal-histology gastric tumor tissues.

Supplementary Figure 8: Validation of presence of GA-FG-CCP in two independent bulk sequencing datasets and large-scale histological images (A) Representative HE staining image of IGC3 patient composed of differentiated spherical cells (B) Heatmap showing the relative expression of top 50 differentially expressed genes in C2 and C4. Right, average expression of all normal epithelial cells for heatmap genes, colored by red if average expression < 0.5 (C) Validation of presence of GA-FG-CCP in Singapore dataset (D) Representative HE staining image of GA-FG-CCP type patient from GDC Data Portal website.

Supplementary Figure 9: HE staining image of 9 inferred GA-FG-CCP type patients from GDC Data Portal website.

Supplementary Table Legend

Online supplementary table S1, Top 100 signature genes of tumor samples versus normal samples from TCGA dataset.

Online supplementary table S2, Top 100 signature genes of malignant cells versus non-malignant cells at single cell level.

Online supplementary table S3, Marker genes in each malignant epithelial cell clusters.

Methods:

Preparation of fresh tissue material and its dissociation into single cells

There were 12 tissue samples in our study, and all study participants provided written informed consent and study protocols were approved by the ethical review community of the Fifth Medical Center of Chinese PLA General Hospital (No. ky-2017-8-11). All of freshly resected biopsy specimens were obtained from one or two sites by conventional upper gastrointestinal endoscopy using Radial Jaw 4 jumbo forceps (Boston Scientific, Natick, MA). Fresh samples were then washed with phosphate-buffered saline (PBS) and divided into two equal parts. One part was processed to generate single-cell suspensions for scRNA-seq, while the second part was used for other experiments, including EBER detection and immunofluorescence examination. Tissue digestion was performed in digestive enzyme mixture which contains 10 ml pre-warmed RPMI1640 (ThermoFisher Scientific), 2 mg/ml dispase (Roche), 1 mg/ml type IV collagenase (Sigma) and 10 U/μl DNase I (Roche) for 30 minutes at 37°C. The reaction was deactivated by the addition of 10% FBS (ThermoFisher Scientific). Cell suspensions were filtered using a 70 μm filter and then centrifuged at 500 rpm for 6 min at 4°C to pellet dead cells and red blood cells. The cells were washed twice and re-suspended in PBS with 0.5% bovine serum albumin (BSA, Sigma).

Library preparation and sequencing

FACS-sorted viable cells were loaded into a well in a microfluidic chip to generate a cDNA library using a droplet based sequencing platform (10x Genomics). Single-cell transcriptomic amplification and library preparation were performed using single-cell 3′ v2 chemistry (10x Genomics) according to manufacturer's instructions. The libraries were then pooled and sequenced across six lanes on Illumina HiSeq X-10 or NovaSeq 6000 system.

Single-cell gene expression quantification, quality control, and cell type determination

The sequencing data from 10x Genomics were aligned and quantified using the CellRanger software package (version 3.1) against the human reference genome (hg19). Raw gene expression matrices were imported and processed using the Seurat R package (version 3.1.3). Standard scRNA-seq filtering excludes low-quality cells with >15-25% mitochondrial UMIs, whereas the filtering criterion of human liver hepatocytes[1] and mouse kidney epithelium[2] was set at 50% because these cells usually have high mitochondrial content. As we know, gastric epithelial cells are located in the environment that contains digestive enzymes and stomach acid. The rate of cell renewal is fast (Gastric pit cells are replaced every 2-4 days). Moreover, gastric epithelial cells such as chief cells and parietal cells serve as important roles in energy metabolism. Gastric tissues are similar with kidney and liver tissues which are all active in metabolism and thus the gastric epithelium may have high mitochondrial content. Therefore, we filtered low-quality cells by referring to the quality control process of MacParland SA et al[1] and Park J et al[2],

following two measurements: 1) cells had either fewer than 1001 UMIs, over 6,000 or less than 501 expressed genes, or over 50% UMIs derived from the mitochondrial genome and 2) cells had an average expression level of less than 2 for a curated list of housekeeping genes. Gene expression matrices of the remaining 27,677 cells were normalized to the total cellular UMI count. The normalized expression was scaled by regressing out the total cellular UMI counts. Highly variable genes were calculated using Seurat “FindVariableGenes” function according to an average expression between 0.125 and 3 as well as a quantile-normalized variance exceeding 0.5. Then we performed principal component analysis (PCA) using HVGs, and significant top 20 principal components (PCs) were selected to perform tSNE dimensionality reduction. All of cells were clustered using DBClustDimension function and clusters with cell numbers less than 50 were deserted. Unbiased clustering generated 14 main clusters and were annotated to 9 known cell types according to canonical marker genes.

Recognition of malignant and non-malignant epithelial cells

Bulk RNA-seq data of stomach adenocarcinoma (dataset ID: tcga_RSEM_gene_tpm) were downloaded from the UCSC Xena website (https://xenabrowser.net/), along with the curated clinical data (dataset ID: Survival_SupplementalTable_S1_20171025_xena_sp). Differentially expressed genes were first calculated between matched tumour and normal TCGA samples using the limma package. Then, each epithelial cell in our study was assigned an initial malignant score and a non-malignant score simultaneously (method described below). Putative malignant and non-malignant epithelial cells were defined

based on the two scores using the k-means clustering algorithm. The initial recognition derived from the TCGA bulk tissues is biased due to the inclusion of non-epithelial cells. Thus, we next generated differentially expressed genes between putative malignant and non-malignant epithelial cells, re-calculated the malignant/non-malignant scores and classified epithelial cells, as described above. We repeated the process iteratively until the classification result was consistent.

Definition of malignant and non-malignant scores

Signature genes for malignant scores were selected as the top 50 highly expressed genes (adjusted p value < 0.01) in tumour tissues (the initial step) or malignant cells (the subsequent steps). Signature genes for non-malignant scores were selected as the top 50 highly expressed genes (adjusted p value < 0.01) in normal tissues (the initial step) or non-malignant cells (the subsequent steps). Malignant scores and non-malignant scores were calculated using the “AddModuleScore” function in “Seruat” R package.

Scoring cells and bulk samples for gene signatures

The scoring for cells and bulk samples were calculated using the “AddModuleScore” function in “Seurat” R package. This method was previously described in Puram SV et al[3] and has been often applied in cell scoring and bulk tissue scoring. Given a certain pre-defined set of genes (G_j), we generated for each cell i, a score, $SC_j(i)$, quantifying the degree to which sample i expressed G_j. $SC_j(i)$ was calculated by subtracting the average relative expression (Er) of a control gene set G_j^{cont} from the average relative expression of

G_j. The control gene set G_j^{cont} is 100-fold of G_j and has a comparable distribution of expression levels to that of G_j. First, all analyzed genes are binned into 25 bins of equal size based on their aggregate expression levels. Next, we randomly select 100 genes from the same bin for each gene in G_j, such that its average expression is analogous to averaging over 100 randomly-selected gene-sets of the same size as G_j. A similar approach was used to define bulk sample scores in our study.

Searching GA-FG-CCP in bulk transcriptomic dataset

The existence of GA-FG-CCP in bulk samples was validated in TCGA STAD dataset. Top 50 differentially expressed genes in C2 and C4 were selected as two signatures to recognize GA-FG-CCP from intestinal-type samples. To minimize the influence of normal cells in bulk tissues, signature genes with average expression over 0.5 in normal epithelium were discarded, such as LIPG, PGC and PGA3. Non-negative matrix factorization (using NMF R package) was performed on the centred expression data of intestinal-type samples, by converting negative values to zero. We integrated the hierarchical clustering results generated by NMF reconstruction with different factors (from 2 to 15) and defined the sample-sample distances as the co-existence times of sample-pairs in the same cluster. The sample-sample distance matrix was used to plot the heatmap with hierarchical clustering result. We also changed the maximum number of factors in NMF from 15 to 20, 25, 30 and observed similar clusters corresponding to GA-FG-CCP.

Tumor differentiation score and survival analysis

We first calculated the *KRT20*-correlated genes using the threshold 0.2. Among the 128 genes with correlation coefficient > 0.2, eight genes have been reported to be closely related to epithelium differentiation (online supplementary table S3), *i.e. KRT20*, *PHGR1*, *MDK*, *CHDR2*, *RARRES3*, *GPA33*, *SLC5A1* and *MUC13*. We used these genes to define differentiation scores for each malignant cell and tumor sample. The differentiation scores were calculated using the "AddModuleScore" function in "Seurat" R package. Then an online tool, Cutoff Finder[4] (http://molpath.charite.de/cutoff/) was used to find an optimal cutoff, which can divide tumors into low differentiation group and high differentiation group based on differentiation scores. This method fits Cox proportional hazard models to the dichotomized variable and the survival variable. Survival analysis is performed using the "coxph" function in "survival" R package. The optimal cutoff is defined as the point with the most significant split. Finally, the survival curve was plotted by the "survminer" R package.

EBV detection using in situ hybridization (ISH)

EBV detection using the ISH assay was performed using formalin-fixed paraffin-embedded tissue sections with a commercially available EBV oligonucleotide probe complementary to EBER-1 (PanPath, Amsterdam, Netherlands) according to the manufacturer's instructions. Positive signals were recognized as dark brown nuclear staining by light microscopy.

Multiplex immunohistochemistry (IHC) staining on human stomach tumour tissue

Multiplex IHC staining was performed on 4-μm-thick, formalin-fixed, paraffin-embedded sections using an Opal multiplex IHC system (NEL811001KT, PerkinElmer) according to the manufacturer's instructions. Briefly, sections were incubated with 3% hydrogen peroxide for 10 min at room temperature, then incubated with a single primary antibody, and then washed and incubated with a probe antibody specific to the species of the primary antibody for 10 min, washed and then incubated for a further 10 min with a horseradish-peroxidase (HRP)-conjugated antibody specific to the probe. Following this, sections were washed and then incubated with opal fluorophores at a 1/50 dilution made up in tyramide signal amplification (TSA) reagent (Opal 7-Color IHC, Perkin Elmer, USA). Images were taken using Vectra Polaris automated quantitative pathology system (PerkinElmer) and image analyzed by inForm 2.3.0 software (PerkinElmer). For IHC, LY6K mouse mAb (IHC, 1:100, cat. sc-393560, santa cruz biotechnology) was used to stain EBV (+) and EBV (-) tumour tissues sections. For ISH, the primary antibodies used for staining were divided into three panels to characterize the expression module in SPEM, GA-FG-CCP (C4 subgroup) and EBV+ (C5 subgroup) GA.

Panel 1 included antibodies against MUC6 mouse mAb (IF, 1:200, cat. ab212648, Abcam), PGA3 mouse mAb (IF, 1:200, cat. ab50123, Abcam), TFF2 (IF,1:1000, cat.13681-1-AP, proteintech) and DAPI.

Panel 2 included antibodies against MUC6 mouse mAb (IF, 1:200, cat. ab212648, Abcam), PGA3 mouse mAb (IF, 1:200, cat. ab50123, Abcam) and DAPI.

Panel 3 included antibodies against HLA-DR rabbit mAb (IF, 1:100, cat. ab92511, Abcam),

KRT18 mouse mAb (IF, 1:100, cat. ab668, Abcam) and DAPI.

Reference:

[1] MacParland SA, Liu JC, Ma X-Z, et al., Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations, Nature communications 9 (2018) 4383-4383.

[2] Park J, Shrestha R, Qiu C, et al., Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease, Science 360 (2018) 758-763.

[3] Puram SV, Tirosh I, Parikh AS, et al., Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer, Cell 171 (2017) 1611.

[4] Budczies J, Klauschen F, Sinn BV, et al., Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization, PLoS One 7 (2012) e51862.