Cancer Discovery: “Research Article”

Single-cell atlas of lineage states, tumor microenvironment and subtype-specific expression programs in gastric cancer

Vikrant Kumar¹#, Kalpana Ramnarayanan¹#, Raghav Sundar¹,²,³,⁴,⁵#, Nisha Padmanbahan¹#, Supriya Srivastava⁶, Mayu Koiwa⁷, Tadahito Yasuda⁷, Vivien Koh²,⁸, Kie Kyo Huang¹, Su Ting Tay¹, Shamaine Wei Ting Ho¹,⁸ Angie Lay Keng Tan¹, Takatsugu Ishimoto⁷, Guowei Kim³,⁹, Asim Shabbir³,⁹, Qingfeng Chen¹⁰,¹¹, Zhang Biyan¹², Shengli Xu¹²,¹³, Kong-Peng Lam¹¹,¹²,¹⁴, Huey Yew Jeffrey Lum¹⁵, Ming Teh¹⁵, Wei Peng Yong²,⁵,⁸, Jimmy Bok-Yan So³,⁵,⁹,¹⁶, Patrick Tan¹,⁵,⁸,¹³,¹⁴,¹⁷,¹⁸

¹Cancer and Stem Cell Biology Program, Duke-NUS Medical School, Singapore; ²Department of Haematology-Oncology, National University Cancer Institute, National University Health System, Singapore; ³Yong Loo Lin School of Medicine, National University of Singapore, Singapore; ⁴The N.1 Institute for Health, National University of Singapore, Singapore; ⁵Singapore Gastric Cancer Consortium, Singapore; ⁶Department of Medicine, National University of Singapore, Singapore; ⁷Gastrointestinal Cancer Biology, International Research Center for Medical Sciences (IRCMS), Kumamoto University, Kumamoto, Japan; ⁸Cancer Science Institute of Singapore, National University of Singapore; ⁹Department of Surgery, University Surgical Cluster, National University Health System, Singapore; ¹⁰Institute of Molecular and Cell Biology, Agency for Science, Technology and Research, Singapore; ¹¹Department of Microbiology & Immunology, Yong Loo Lin School of Medicine; ¹²Singapore Immunology Network (SIgN), A*STAR; ¹³Department of Physiology, National University of Singapore, Singapore; ¹⁴School of Biological Sciences, Nanyang Technological University, Singapore; ¹⁵Department of Pathology, National University Health System, Singapore; ¹⁶Division of Surgical Oncology, National University Cancer Institute, Singapore; ¹⁷Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore; ¹⁸SingHealth/Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore.

#These authors contributed equally to the study

*Corresponding author
Prof Patrick Tan
Duke-NUS Medical School
8 College Road, Singapore 169857, Singapore.
Phone: +65-6516-1783
Fax: +65-6221-2402
E-mail: gmstanp@duke-nus.edu.sg

Running title. Single-cell atlas of gastric cancer subtypes

Keywords. Single-cell sequencing, gastric cancer, diffuse, CAF, spatial transcriptomics
**Funding acknowledgement.** RS is supported by the National Medical Research Council (NMRC), Singapore (NMRC/Fellowship/0059/2018 and MOH-000627). PT is supported by Duke-NUS Medical School and the Biomedical Research Council, Agency for Science, Technology and Research. This work was also supported by NMRC grants NR13NMR111OM, NMRC/STaR/0026/2015 and MOH-OFLCG18May-0003. Funding was also provided by Cancer Science Institute of Singapore, NUS, under the National Research Foundation Singapore and the Singapore Ministry of Education Research Centres of Excellence initiative. The funders of the study had no role in study design, data collection, analysis, interpretation, or writing of the manuscript. PT had full access to all data in the study and as corresponding author had final responsibility for the decision to submit for publication.

**Declaration of interests.** PT has stock and other ownership interests in Tempus Healthcare, previous research funding from Kyowa Hakko Kirin and Thermo Fisher Scientific, and patents/other intellectual property through the Agency for Science and Technology Research, Singapore (all outside the submitted work). RS has received honoraria from Bristol-Myers Squibb, Lilly, Roche, Taiho, Astra Zeneca, DKSH and MSD; has advisory activity with Bristol-Myers Squibb, Merck, Eisai, Bayer, Taiho, Novartis, MSD and AstraZeneca; received research funding from MSD and Paxman Coolers; and has received travel grants from AstraZeneca, Eisai, Roche and Taiho Pharmaceutical. All remaining authors have declared no conflicts of interest.
Abstract

Gastric cancer (GC) heterogeneity represents a barrier to disease management. We generated a comprehensive single-cell atlas of GC (>200,000 cells) comprising 48 samples from 31 patients across clinical stages and histological subtypes. We identified 34 distinct cell-lineage states including novel rare cell populations. Many lineage states exhibited distinct cancer-associated expression profiles, individually contributing to a combined tumor-wide molecular collage. We observed increased plasma cell proportions in diffuse-type tumors associated with epithelial-resident KLF2, and stage-wise accrual of cancer-associated fibroblast sub-populations marked by high INHBA and FAP co-expression. Single-cell comparisons between patient-derived organoids (PDOs) and primary tumors highlighted inter- and intra-lineage similarities and differences, demarcating molecular boundaries of PDOs as experimental models. We complemented these findings by spatial transcriptomics, orthogonal validation in independent bulk RNA-seq cohorts, and functional demonstration using in vitro and in vivo models. Our results provide a high-resolution molecular resource of intra- and inter-patient lineage-states across distinct GC subtypes.
Statement of Significance

We profiled gastric malignancies at single-cell resolution and identified increased plasma cell proportions as a novel feature of diffuse-type tumors. We also uncovered distinct CAF subtypes with *INHBA-FAP* high cell populations as predictors of poor clinical prognosis. Our findings highlight potential origins of deregulated cell-states in the gastric tumor ecosystem.
Introduction

Gastric cancer (GC) is a leading cause of global cancer morbidity and mortality (1) with particularly high incidence in Asia, Eastern Europe, and Central America (2). Between individual patients, gastric tumors frequently exhibit high levels of histological, transcriptomic, and (epi)genomic variation, with distinct clinical behaviors and treatment response (“inter-patient heterogeneity”). Factoring this heterogeneity into GC clinical management, and identifying molecular pathways driving hallmarks of GC variation, represent important goals for improving patient outcomes. Although progress has been made in defining specific molecular subtypes of GC through consortia such as The Cancer Genome Atlas (TCGA) and Asian Cancer Research Group (ACRG) (3,4), tangible improvements in patient outcomes based on these findings have been modest, compounded by the growing recognition that GCs also exhibit high levels of within-patient heterogeneity ("intra-patient heterogeneity”; ITH). High-resolution studies probing the molecular extent of GC ITH across a wide range of GC patients are thus required (5), to understand key principles governing GC evolution, selection and adaptation and how clinical care pathways should be adapted to capture GC ITH.

Using “bulk-transcriptome” experimental methods, we and others have previously established that each gastric tumor possesses a personalized expression profile comprising distinct transcriptional programs, contributed by both cancer epithelial cells and other cell types in the tumor microenvironment (TME) (6,7). However, our understanding of mechanisms by which TME-resident cell types such as immune cells, fibroblasts, and blood vessels drive GC phenotypes and clinical trajectories remains nascent (8). While advanced bioinformatic programs have been designed to decompose bulk sequencing data into lineage-specific constituent programs, these deconvolution algorithms are often not able to discern fine-scale tissue lineages, relationships between lineages, rare cell populations, and cell-cell interactions (9). To tackle these challenges, single-cell RNA sequencing (scRNA-seq) is proving to be a powerful tool for characterizing gene expression across thousands of cells simultaneously, enabling comprehensive profiling of different cell types in tumors in distinct biological states and conditions (10). In GC, recent scRNA-seq studies have provided unique insights on different aspects of gastric tumor biology. For example, Wang et al., performed scRNA-seq on malignant cells
from ascitic fluids of GC patients to develop a prognostic signature based on features of malignant peritoneal cells (11). Other scRNA-seq studies of primary GC tumors have been performed, but on limited numbers of samples and cells (8-13 patients, 27,000 to 55,000 cells per study) (12-14). Another limitation of current single-cell sequencing platform is the requirement for tissue dissociation which leads to loss of spatial information. To address this, newer “spatial transcriptomic” platforms such as digital spatial profiling (DSP), in situ sequencing, and MERFISH have been developed that retain spatial architecture thereby allowing analysis of tumor-TME interactions at unprecedented depth (15-17).

Here, we performed scRNA-seq on an expanded cohort of more than thirty GC patients with different subtypes and stages, across a large number of cells (>200,000 cells) to generate a comprehensive single-cell landscape of GC encompassing both inter and intra-tumoral heterogeneity. We discovered unique and novel features of the gastric TME, including an increased proportion of ‘plasma cells’ in diffuse-type tumors, and a role for INHBA, a subunit of activin-inhibin complexes, in specific subtypes of cancer-associated fibroblasts (CAFs). We also performed spatial transcriptomic analysis on these samples to geographically validate predicted inter and intra-cell relationships in situ, and further reinforced our findings using functional assays and conventional immunohistochemistry.
Results

scRNA-seq of ~200,000 GC cells identifies diverse tissue-lineage states, cell fate trajectories, and rare cell populations

Droplet-based scRNA-seq (10x Genomics) was performed on 48 surgical resection and biopsy samples across 31 unique individuals with GC, ranging from Stage I to Stage IV, distinct histological subtypes (diffuse and intestinal), molecular subtypes (TCGA), primary tumors to distant (peritoneal) metastases, and matched normal gastric tissues (Figure 1A, Supplementary Table 1). We also performed scRNA-seq on tumor and matched normal GC patient-derived organoids (PDOs) to investigate the effects of in vitro culture on cancer-intrinsic and TME signatures (Figure 1A, Supplementary Table 1). After quality control (QC) and removal of batch effects (18), 200,954 single cells were included in the final dataset. For 13 samples (10 tumors and 3 normals; 156 regions of interest (ROIs)), we also performed spatial transcriptomics using the DSP platform (NanoString GeoMx) to gain insights into in situ geographical and spatial relationships linking discrete cell-states and molecular interactions.

We first performed dimensionality reduction on 152,423 cells derived from 40 primary samples (29 tumors; 11 normal (of which 10 were matched)). This analysis revealed 34 unique tissue-states (Figure 1B). Using tissue-type specific canonical markers defined in the literature (14) (Supplementary Figure 1), the cell-states were broadly categorized into 5 major cell-types, referred to as “meta-clusters” (myeloid, lymphoid, plasma, epithelial, and stromal) (Figure 1C, Supplementary Figure 2A). Supporting the supervised cell-type specific marker analysis, an unsupervised global clustering similarity matrix also classified the cell states into five meta-clusters (Figure 1D). Alternative clustering algorithms (e.g., integrative Non-negative Matrix Factorization) confirmed the molecular distinctiveness of each cluster, and comparison of global versus metacluster-specific clustering approaches revealed similar results (Supplementary Figure 2B-E).

Overall, immune cell populations dominated the cell-states (21 of 34 states). These included (1) a “Myeloid meta-cluster” (5 cell-states), consisting of a mast cell cluster (KIT positive: “MAST”), a dendritic cell cluster (PLD4 positive: “MYD”) and three macrophage clusters (CD163 positive: “MYM1-3”); (2) a “Lymphoid meta-cluster” (11 cell-states), consisting of six T-cell clusters, mapping to effector and
 naïve CD8 T-cells (CD8A, GZMH, GZMM, NKG7; LT1-2), naïve and helper CD4 T-cells (CCR7, STAT4; LT3-4), regulatory T-cells (Tregs) (IL2RA, STAT3; LT5) and proliferative T-cells (expressing cell-cycle genes; LT6), four NK cell clusters (KLRD1 positive: “LNK1-4”), and one B-cell cluster (MS4A1 positive: “LB”), and a (3) “Plasma meta-cluster” (5 cell-states), consisting of five mature B-cells/Plasma cell clusters (TNFRSF17 positive: “PL1-5”) (Supplementary Figure 1, Supplementary Table 2).

The “Epithelial” meta-cluster (7 cell-states; CDH1 positive) contained three distinct sub-lineages; “EpiPit”, expressing markers of mucous pit cells (MUC5AC, TFF1; EpiPit1&2), “EpiC”, expressing markers of chief cells (LIPF, PGA3; EpiC1-3) and “EpiInt”, expressing intestinal-type markers (REG4, TFF3; EpiInt1&2) (Supplementary Figure 1). EpiInt cells exhibited features of intestinal metaplasia (IM), a pre-malignant condition recognized histologically by gastric epithelial cells acquiring intestinal cell-type features (19). This was supported by the expression of CDX1 and CDX2 master transcriptional regulators, largely in EpiInt cells, which play a key role in the trans-differentiation of gastric epithelial cells to intestinal cell-type (Supplementary Figure 2F) (20). The “Stromal meta-cluster” (6 cell-states; FN1 positive) consisted of pericytes (STF2; defined by RGS5 and NOTCH3), fibroblasts (STF1 and STF3, defined by LUM and DCN), and PLVAP positive endothelial cell subclusters. The latter could be further partitioned into ACKR1-positive venular endothelial cells (STE1) or non-venular endothelial cells (STE2) (21) (Supplementary Figure 1). The remaining cluster (STF4) could not be assigned to a single cell-type and is described further in the next paragraph. The cell types varied in their relative proportions across the ~150K cells (Figure 1C; Supplementary Figure 2G). Specifically, lymphoid cells exhibited the largest proportion (39.7%; range – 19.1% for T-cells to 3.6% for B-cells) followed by epithelial cells (20%, range – 8.5% for EpiC to 5.4% for EpiPit cells). Myeloid, stromal and plasma cells constituted proportions of 11.4%, 14.4%, 14.5% respectively.

The large number of cells profiled in our study enabled us to relate distinct cell-states to one another across biological state transitions. For example, single-cell trajectory and pseudo-time analysis of the plasma cell meta-cluster demonstrated different stages of plasma cell differentiation and maturation, consistent with the known literature (22) (bootstrapping p-value <0.05 on major branching nodes) (Figure 1E). These included plasmablasts (high XBP1 and low SDC1), short-lived plasma cells (high SDC1), and long-lived plasma cells (high SDC1, STAT3, IKZF3)
Short-lived plasma cells predominated the trajectory plot, while long-lived plasma cells exhibited relatively higher expression of *IGHA1* (IgA), consistent with increased levels of IgA detected in gastric lamina propria (27) (Supplementary Figure 2H). Similarly, single-cell trajectory analysis of macrophages revealed two distinct cell-states: Pro-inflammatory “M1-like” (high *CD163* and *S100A12*) and “M2-like Tumor associated macrophages” (high *CD163* and *FOLR2*) (28,29) (Supplementary Figure 2I). The large number of analysed cells also enabled characterization of a novel and rare cell-type (STF4) within the stromal meta-cluster (0.4% of all cells, *n* = 821), expressing markers associated with both endothelial (*PLVAP*) and fibroblast (*RGS5*) lineages (Figure 1F), possibly highlighting cells undergoing endothelial-mesenchymal transition (EndoMT), a process where endothelial cells acquire a mesenchymal or myofibroblastic phenotype. We ruled out the possibility that the double-lineage nature of STF4 cells is caused by potential technical artefacts such as doublet effects (30), and orthogonally validated the existence of the *PLVAP/RGS5* double-positive population on FFPE (formalin-fixed paraffin embedded) sections using dual-color RNAScope (Supplementary Figure 2J). Taken together, these results showcase the ability of scRNA-seq data, when appropriately powered, to reveal cells exhibiting diverse tissue-lineage states, different state transitions, and rare populations associated with mixed or dual lineages.
scRNA-seq discerns cell-state-specific transcriptional patterns associated with GC

To examine tumor-associated expression programs at the single-cell level, we compared primary gastric tumor samples to normal tissues (n = 26 tumors; n = 10 normals). After random down-sampling to ensure statistical equivalency (Figure 2A), the proportion of epithelial cells in tumors was lower compared to normals (p = 0.002), while myeloid cells in tumors were higher (p = 0.05) (Supplementary Table 3). The proportions of lymphoid, plasma and stromal cells were not statistically different between tumor and normal samples.

Compared to their cognate cell-states in normal tissues, tumor-associated epithelial cells exhibited higher expression of multiple oncogenic gene signatures related to epithelial-mesenchymal transition (EMT), cell motility, and cancer-derived signatures from previous GC scRNA-seq studies (12-14) (Figure 2B, Supplementary Figure 3A-B, Supplementary Table 4) (p < 0.0001). When we assessed the expression status of individual GC oncogenes (e.g., CEACAM6, EGFR, MET, CCND1 and KRAS) among the tumor epithelial clusters, the EpiInt1 tumor epithelial cluster exhibited the largest increase (Figure 2C). We orthogonally validated the oncogenic gene signature associations in multiple ways. First, we confirmed a significant correlation between the proportions of inferred scRNA-seq tumor cells and EpiInt1 tumor cellularity proportions measured by H&E (hematoxylin and eosin) staining (R = 0.64, p < 0.0074). Second, inferring copy number variations (CNVs) from the scRNA-seq data using approaches employed in other cancer single-cell studies (31-33), we confirmed that tumor-associated epithelial cells exhibited significantly higher numbers of inferred aberrant CNVs compared to normal epithelial cells (p < 0.0001). In contrast, CNV differences were not observed between macrophage populations in tumor and normal samples (Figure 2D). Third, further supporting a substantial proportion of epithelial cells in the tumor samples as malignant, CNV profiles inferred from scRNA-seq exhibited significant concordance with CNV profiles of the same tumor samples determined by bulk whole exome sequencing (WES) (p-value range 0.022 to <0.0001, Supplementary Figure 3C). At the subcluster level, similar to their high expression of oncogenic gene signatures, EpiInt1 cells also harbored the highest proportion of CNVs (Supplementary Figure 3D).
Comparisons in a cell-state specific manner revealed that “chief cells” and “intestinal-type cells” showed the greatest differences in transcriptomic profiles between GCs and their cognate cell-state in normal gastric tissues (Figure 2E; “Chief cells”: 1232 genes/EpiC3 cluster; “Intestinal-type cells”: 775 genes/EpiInt2 cluster). Pathway analysis indicated that each tumor epithelial cluster shared both common upregulated modules relevant to cancer, but also cluster-specific tumor-associated modules. For example, all tumor epithelial clusters exhibited upregulation of immune-related pathways such as “Antigen Presentation: Folding, assembly and peptide loading of class I MHC”. However, only EpiC3 exhibited upregulation of “MHC class II antigen presentation” and downregulation of “Citric Acid (TCA) cycle and respiratory electron transport” (Supplementary Table 5). Among downregulated genes, the classical chief cell markers LIPF and PGA3 were among the top downregulated genes in tumor epithelial cells (Figure 2F) (p < 0.0001). Using spatial DSP, we confirmed the loss of LIPF in tumor epithelial cells (PanCK+) compared to normal samples (n = 13, p = 0.0035) (Figure 2G). Interestingly, we and others have previously identified LIPF as a lineage-specific target of recurrent insertion-deletion mutations in GC (34,35).

Among stromal meta-cluster cell-states, absolute proportions of fibroblast (clusters STF1-4) and endothelial cells (clusters STE1 and STE2) were not significantly altered between tumor and normal (Figure 2A, Supplementary Figure 3E, Supplementary Table 3). However, in the endothelial lineage, pathway analysis revealed a conserved expression response related to extracellular matrix remodelling between tumor and normal samples (Supplementary Table 6). Similarly, in the fibroblast lineage, two fibroblast sub-lineages comprising LUM associated fibroblasts (STF1 and STF3) exhibited upregulation of CAF genes such as FAP, COL8A1, THBS2, and CTHRC1, while a third sub-lineage comprising pro-angiogenic pericytes (STF2) displayed upregulation of another set of CAF marker genes such as CSPG4, PDGFA, ASPN, S100A4 (Figure 2H, Supplementary Figure 3F). Pathway analysis indicated several common (e.g., ‘ECM proteoglycans’) and specific pathways individualized to each cluster (Supplementary Table 7). For example, STF2 showed upregulation of ‘Signalling by PDGF’, while STF3 had upregulation of ‘Activation of Matrix Metalloproteinases’.

We also investigated T-cell proportions and immune checkpoints of therapeutic interest between tumor and normal (36). Consistent with the prior
literature (37), we observed a higher proportion of Tregs (p = 0.0048) and naïve CD4 T-cells (p = 0.048) in tumors compared to normal (Supplementary Figure 4A,B). T-cell receptor (TCR) sequencing revealed TCR diversity ranges of 34.3 to 994 similar to studies in other tumor types (38,39) (see Supplementary Figure 4C for further details). Taken collectively, these results suggest that in the GC tumor ecosystem, distinct cell states from both the epithelial and tumor microenvironment likely express different oncogenic transcriptomic features - each required for different cancer hallmarks and ultimately converging and intermixing to elicit a composite tumor molecular portrait (Figure 2l, Supplementary Figure 4D).
scRNA-seq reveals a KLF2-associated plasma cell program in diffuse-type GC

Lauren’s GC classification has genomic, clinical and prognostic value (40). We analyzed our scRNA-seq dataset comparing diffuse (n = 6) and intestinal-type (n = 14) tumors, after random down-sampling to achieve subtype-matching. The relative proportion of plasma cells was higher in diffuse compared to intestinal GCs (p = 0.05). In contrast, epithelial cells were lower in diffuse GCs (p = 0.04), with intestinal GCs exhibiting higher proportions of epithelial subclusters and inferred CNVs (p = 0.029) (Supplementary Figure 5A). There were no statistically significant differences in the lymphoid, myeloid and stromal cell populations (Supplementary Table 8) (Figure 3A). We confirmed an increased level of plasma cells in diffuse vs. intestinal tumors using IRF4 (a plasma cell marker) immunohistochemistry on a subset of tumors (n = 17, p = 0.036, Figure 3B, Supplementary Figure 5B). We also independently verified the increased level of plasma cells in diffuse-type GCs by re-analyzing bulk RNA-seq GC TCGA data using CIBERSORTx deconvolution (p = 0.028) (Figure 3C). Performing sub-lineage analysis in single cell data, we established that the enrichment of plasma cells in diffuse-type GC was primarily driven by plasma cell clusters PL4 and PL5 (p < 0.05) (Figure 3D). Pathway analysis comparing PL4-5 against PL1-3 in tumor samples, indicated upregulation of cytokine and interleukin signalling, and downregulation of CD22-mediated BCR regulation in PL4-5 (Supplementary Table 9).

KLF2 is a gene previously shown to regulate homing of plasma cells in multiple myeloma (41,42). We hypothesized that the increased recruitment of plasma cells in diffuse-type GCs might be mediated by KLF2 activity. To test this hypothesis, we compared KLF2 expression between diffuse and intestinal subtype GCs across the various cell states. Compared to intestinal-type cell states, plasma cells and epithelial cell clusters in diffuse-type GCs expressed increased KLF2 expression in >50% of cells with log fold difference > 0.5 (Supplementary Figure 5C). Notably, we only observed significant correlations between plasma cell proportions and KLF2-expressing epithelial cells (R = 0.69, p = 0.0008) and not between plasma cell proportions and KLF2-expressing plasma cells (R = 0.35, p = 0.13) (Figure 3E, Supplementary Figure 5D). This finding suggests that high-KLF2 expressing epithelial cells may be associated with plasma cell recruitment in diffuse-type GC. Sub-lineage analysis of diffuse-type epithelial cells revealed the highest KLF2 expression in EpiC clusters (log fold from 1.7 to 0.8, compared to EpiC intestinal-
type cells) (Supplementary Table 10). To further confirm KLF2 expression differences between intestinal-type and diffuse-type GCs, we analyzed epigenomic promoter activity (as measured by H3K27ac ChIPseq) and gene expression of KLF2 in an independent bulk-RNA dataset of 24 GCs (9 diffuse-type and 14 intestinal-type). Both KLF2 H3K27ac promoter levels and gene expression were higher in diffuse compared to intestinal GCs (p <0.05), with good correlations between promoter signals and gene expression (R = 0.55, p = 0.005) (Supplementary Figure 5E). In a second independent bulk RNA-seq dataset (GC TCGA), we further confirmed increased KLF2 expression in diffuse-type compared to intestinal-type tumors (p = 0.00013), with boundaries of variance similar to with other credentialed oncogenes such as ERBB2 and HNF4α associated with intestinal-type GC (Figure 3F, Supplementary Figure 5F)(43,44). To explore KLF2 differences in a spatial context, we then used DSP analysis to study the expression of KLF2 in epithelial cells (Pan-CK+) and confirmed higher KLF2 transcript levels in diffuse-type epithelial cells compared to intestinal-type epithelial cells (p = 0.02; n = 10) (Figure 3G). Using the same platform, we next studied KLF2 expression in epithelial cells (Pan-CK+) in the context of proximity to plasma cells (CD138+) (Epi\textsuperscript{prox} vs. Epi\textsuperscript{dist}). We observed a trend towards higher KLF2 expression in epithelial cells in close proximity to plasma cells compared to distal epithelial cells (p = 0.096) (Figure 3H, Supplementary Figure 5G).

To explore dynamic temporal interactions between KLF2 expression in tumor epithelial cells with host plasma cells, we employed a humanized-mouse cancer \textit{in vivo} model. Immune-deficient mice pups were engrafted with human umbilical cord blood CD34+ cells, and mice with post-engraftment human immune-cell reconstitution (termed “humanized-mice”) were selected for the study. We investigated KLF2\textsuperscript{pos} (GSU) and KLF2\textsuperscript{neg} (SNU1750) diffuse-type GC tumors grown in humanized-mice compared to immune deficient mice (Figure 3I). In agreement with our results in primary GCs, KLF2\textsuperscript{pos} xenografts had high expression of plasma cell markers IRF4 and SLAMF7 in the humanized-mice only, reflecting increased plasma cell recruitment associated with KLF2. To further demonstrate a causative role for epithelial KLF2 in plasma cell recruitment, we then performed \textit{in vitro} migration assays using plasma cells derived from PBMCs or KMS-11 (a multiple myeloma cell line) co-cultured with KLF2-positive GC cells. Knock-down of KLF2
expression in two independent KLF2-positive GC cell lines (GSU, LMSU) was sufficient to significantly decrease plasma cell migration \((p = 0.001;\) Figure 3J, Supplementary Figure 5H). Taken collectively, our results pinpoint specific B-cell sub-lineages exhibiting increased proportions in diffuse-type GC and highlight epithelial-cell resident KLF2 expression as a potential driver of plasma cell recruitment.

We further studied the relationship between epithelial cells from diffuse-type GCs with pathways related to immune cell biology. scRNA-seq driven pathway analysis between diffuse vs. intestinal subtype epithelial cells, revealed several upregulated genes in diffuse-type epithelial cells related to immune-signaling pathways, including numerous IG (immunoglobulin) genes belonging to both light chain and heavy chain genes (IgG/IgA) and the IG linker gene, JCHAIN (Figure 3K top panel, Supplementary Table 11). A subcluster pathway analysis of upregulated genes revealed that most of these immune-related modules were expressed in EpiC cells, similar to KLF2 (Figure 3K bottom panel, Supplementary Table 12). In contrast, downregulated genes did not show pathway enrichments related to immune signaling. Taken together, these data suggest a general transcriptional cassette expressed in diffuse-type GC epithelial cells related to engaging the tumor immune microenvironment.
scRNA-seq reveals distinct fibroblast populations and \textit{INHBA-FAP} axis as a CAF regulator

CAFs are known to influence tumor growth, migration and invasion through the regulation of extracellular matrix components in various tumor types (45,46). However, little is presently known about specific pathways of CAF regulation and heterogeneity in GC, due to limited \textit{in vitro} models of GC CAFs and the challenges of deconvoluting bulk RNA-seq data (47). Using our scRNA-seq data, we first compared the tumor fibroblast clusters (STF1-3) according to clinical stage and histological subtype. Supporting an important role for CAFs in gastric malignancy, we observed a stage-dependent increase in all three STF clusters, with STF3 as the dominant population (Figure 4A). We next closely assessed a panel of CAF canonical markers, \textit{FAP, CSPG4, ACTA2} and \textit{TAGLN} (Supplementary Table 4) in STF1-3. While STF1 and STF3 clusters exhibited upregulation of \textit{FAP, ACTA2, TAGLN}, the STF2 cluster exhibited upregulation of only \textit{CSPG4} (Figure 4B), further indicative of distinct CAF sub-lineages in GC.

TGF-\(\beta\) super family signaling has been reported to influence CAF function in other cancer types (48). We thus elected to study the Activin-inhibin signaling module, a major component of the TGF-\(\beta\) pathway, comprising 9 canonical genes (Supplementary Table 4). Among these 9 genes, only \textit{INHBA} exhibited significant upregulation in tumor-associated STF2 and STF3 clusters compared to normal (Figure 4C). As tumor STF2 (\textit{CSPG4 high}) and STF3 (\textit{FAP high}) represent distinct CAF populations, we then performed a co-expression correlation analysis between \textit{INHBA} expression and the respective cluster markers. \textit{INHBA} exhibited a significant positive correlation with \textit{FAP} in STF3 (\(R = 0.21, p < 0.0001\)), whereas no correlations were found with \textit{CSPG4} in STF2 (\(R = -0.03\)) (Supplementary Figure 6A). In TCGA bulk RNA-seq data, \textit{INHBA} was similarly correlated to \textit{FAP} (Pearson \(R = 0.59; p < 0.0001\)) but not with \textit{CSPG4} (Pearson \(R = 0.00; p = 0.964\)) (Supplementary Figure 6B). To orthogonally support the association between \textit{FAP} and \textit{INHBA}, we adopted multiple approaches. First, we analyzed spatial DSP data of \textit{FAP} and \textit{INHBA} in fibroblast regions marked by \(\alpha\)-Smooth Muscle Actin (\(n = 13\) samples: 4 ROI per sample) (Figure 4D). We observed higher expression of both \textit{FAP} and \textit{INHBA} in tumor fibroblasts compared to normal (\(p < 0.05\)) (Figure 4E-F), and within tumor fibroblasts, there was a strong correlation between \textit{FAP} and \textit{INHBA} co-
expression levels (R = 0.92, p < 0.0001) (Figure 4G). Second, in an independent cohort of 10 in vitro cultured patient-matched normal and tumor fibroblasts, isolated and purified by fluorescence-activated cell sorting from GC patients (47), we confirmed INHBA up-regulation (p = 0.023) in tumor fibroblasts, along with increased FAP (p = 0.027) (Figure 4H). Third, siRNA-mediated knockdown of INHBA in the tumor fibroblasts resulted in significant FAP gene downregulation (p < 0.001) (Figure 4I, Supplementary Figure 6C). Conversely, treatment of two normal fibroblast lines with recombinant INHBA was sufficient to increase FAP expression at 48 (p < 0.001) and 96 (p < 0.001) hours (Figure 4J). Taken collectively, these results highlight INHBA as a positive regulator of FAP in the GC STF3 fibroblast population.

We also studied correlations between INHBA expression and collagen-related gene expression (49-53), a surrogate for fibrogenic processes regulated by TGF-β signaling. Of nine collagen genes, five positively correlated with INHBA including COL1A1, COL1A2, COL6A3, COL8A1, and COL12A1 (R>0.2, p-value < 0.0001, Supplementary Figure 6D). To further verify these findings, we then analysed DSP data and again confirmed a significant positive correlation of INHBA with COL1A1, COL1A2, and COL6A3 (R > 0.69, p-value < 0.0001; Supplementary Figure 6E; COL8A1, and COL12A1 were not represented on the DSP platform). These results support a positive relationship between INHBA activity and collagen gene induction.

To explore the potential paracrine capability of INHBA (54), we then treated CAFs in vitro with recombinant INHBA for 48 hours and measured the expression of collagen genes. We observed significant increases in collagen genes including TGF-β targets, COL1A1, COL1A2, and COL6A3 in two CAF lines (Supplementary Figure 6F), consistent with a cell non-autonomous role for INHBA.

To assess the clinical relevance of our findings, we mapped FAP and INHBA expression levels across the GC cohort. We found an increasing abundance of FAP-positive and INHBA-positive cells in STF3, in a stage-wise manner from normal to stage IV (Figure 4K) (p = 0.041, after normalizing for differences in STF3 proportion). Survival analysis of TCGA samples by INHBA levels revealed a significantly poorer survival for tumor samples with high INHBA expression (HR: 0.70, 95% CI: 0.51 to 0.97, p = 0.029) (Figure 4L). This difference remained statistically significant even after adjusting for stage (HR: 0.71, p = 0.038). Similar findings was also seen in a pooled analysis of several available GC microarray data (55), (INHBA high vs. INHBA low HR 0.80, 95% CI: 0.68 to 0.96, p = 0.014) (Supplementary Figure 6G).
scRNA-seq of GC organoids support increased cancer cell transcriptional plasticity

Finally, PDOs are increasingly being used as a platform to model GC tumor behavior and drug responses (56-58). To investigate the extent to which *in vitro* organoid culture conditions affect transcriptional lineage states or overall cellular proportions compared to *in vivo* primary GCs, we derived and examined four pairs of matched normal-tumor PDOs (*n* = 48,531 cells) (<12 passages) (Methods, Supplementary Figure 7A). To perform a comparative cell-state analysis between PDOs and primary GCs, we integrated the scRNA-seq dataset from PDOs with the primary tumors. This analysis recapitulated the five major meta-clusters (Figure 5A). We performed a sub-lineage level analysis of epithelial and stromal meta-clusters in the PDOs. For the PDO epithelial meta-cluster, the proportion of the three sub-lineages were EpiC (71% in PDO vs. 43% in primary), EpiInt (15% vs. 30%) and EpiPit (14% vs. 27%). In the PDO stromal clusters, we also observed pericytes (28% vs. 14% in primary), fibroblasts (32% vs. 45% in primary), and endothelial cells (39% vs. 37%) (Supplementary Figure 7B). These findings suggest that sub-lineage heterogeneity is indeed present in PDOs, although proportions may differ compared to primary GCs.

Similar to primary GCs, tumor PDO epithelial cells exhibited upregulation of cancer-associated modules and GC-related genes compared to normal PDO epithelial cells (*p* < 0.0001) (Figure 5B). Interestingly, trajectory plot analysis of epithelial cells in both normal and tumor PDOs demonstrated that epithelial cells from tumor organoids have multiple extended branches relative to normal epithelial cells, consistent with tumor PDO epithelial cells undergoing pervasive and ongoing differentiation/dedifferentiation (Figure 5C). These findings are noteworthy given recent studies reporting that tumor-associated epithelial cells have increased transcriptional plasticity, which may drive key aspects of ITH (11). Our ability to observe similar phenomena in PDOs suggests that PDOs could be used as an *in vitro* experimental model to investigate molecular pathways governing transcriptional plasticity in GC.

Our analysis also revealed differences between primary samples and PDOs. For example, epithelial and stromal meta-clusters in PDOs were significantly enriched, relative to lymphoid and plasma cell clusters which were depleted (Figure 5D) (*p* < 0.001 for stromal, epithelial and lymphoid, *p* = 0.018 for plasma). A gene
expression comparison between PDO and primary samples indicated plasma cells as showing the greatest differences in gene expression profile in PDOs, whereas epithelial signatures were relatively more conserved (Figure 5E). In congruence with these findings, comparison of Reactome programs (59) between the PDOs and primary tumors by meta-clusters confirmed a high overlap of Reactome programs for the Epithelial (Jaccard similarity index (JSI) 0.62) and Stromal meta-clusters (JSI 0.43), with poorer overlap for the plasma (JSI 0.09) and lymphoid meta-clusters (JSI 0.15). Pathway analysis of genes unique to plasma cells, implicated “Classical antibody-mediated complement activation” as significantly downregulated (Figure 5F, Supplementary Table 13). Notably, a significant proportion of genes and pathways were also commonly upregulated in PDOs agnostic of cluster, which included “Cellular response to starvation”, highlighting umbrella culture effects (Figure 5G, Supplementary Table 14). Together, these results raise the possibility that besides epithelial cells, PDO culture may also influence the transcriptional profiles of other cell types associated with tumors.
Discussion

In this study, we applied scRNA-seq across a large number of cells to discover several novel features of GC, including rare cell populations undergoing state transitions, cell-type specific expression programs associated with GC, and distinct plasma cell and CAF sub-lineages associated with GC histological subtypes and clinical stages. Another notable aspect of our study was the application of very recently available spatial transcriptomics technologies (DSP) to orthogonally verify our key findings. Compared to earlier GC scRNA-seq studies, our experimental design and analyses is differentiated by a large cell number size (> 200,000 cells; higher than all prior GC studies combined), samples reflecting multiple clinical stages and subtypes (n = 31 patients, 48 samples), and the comparative analysis of gastric organoids (Figure 6). For example, Zhang et al, performed scRNA-seq on GC biopsy samples (n ~ 30,000 cells), from patients with premalignant lesions (atrophic gastritis, intestinal metaplasia), and one early GC sample, focusing on the evolution of epithelial cells from normal to malignancy (14). At the terminal end of the cancer spectrum Wang et al, performed scRNA-seq only on malignant ascites cells (n ~ 45,000 cells) (11). Sathe et al, studied ~55,000 cells from seven GC and one intestinal metaplasia patient, generating a receptor-ligand network of the TME components, agnostic of tumor subtypes (13).

scRNA-seq analysis of tumor-normal comparisons performed in a cell-lineage specific manner led to the emergence of a composite tumor profile assembled from distinct faulty signatures expressed by different lineages. Our results suggest that the greatest tumor-associated expression differences in the tumor epithelial component likely originate from chief cells and intestinal-type cells. Interestingly, high GC oncogene expression was observed in a specific sub-population of intestinal-type epithelial cells (Epilnt1), suggesting that Epilnt1 cells likely represent a key epithelial cell population important in the transition into malignancy from metaplasia. One potential limitation of our study is the lack of directly-inferred single-cell based DNA alterations, as currently-available platforms are unable to deliver both single-cell DNA and RNA-level alterations from the same cell on a genome-wide scale. Additionally, the global clustering methodology employed in our study does not preclude the possibility that more granular cell types may exist, which could be identified using refined local clustering (60,61). Among non-epithelial cell types, lineage analysis of fibroblasts identified discrete CAF clusters, such as CAFs that
were either LUM positive or pericyte CAFs that were CSPG positive. Both LUM and CSPG encode proteoglycans, and pericytes are being increasingly recognized as key players in tumor vessel formation and growth (62). The differences in gene expression profiles between these two fibroblast clusters may highlight distinct tumor promoting functions. For example, LUM associated CAFs may be involved in the proliferation and survival of tumor cells, while pericyte CAFs may be involved in neovascularization, thereby underscoring the functional heterogeneity of CAFs in GC. We also identified a small cluster of stromal cells (STF4) undergoing EndoMT – these cells expressed markers associated with both endothelial (PLVAP) and fibroblast (RGS5) lineages, confirmed after doublet filtering and by orthogonal RNAscope analysis. EndoMT, an embryonic process required for normal cardiovascular development (63), has been associated with TME plasticity, resistance to antineoplastic therapy and TGF-β and BMP signaling (64-66). Together, these scRNA-seq findings suggest a conceptual model wherein individual cell lineages respond distinctly, either directly or indirectly, to malignant transformation leading to a highly complex tumor ecosystem. It is reasonable to posit that these observations would have been largely obscured in bulk-RNA seq data.

CAFs, representing a predominant stromal cell population, have been shown to play cardinal roles in shaping tumor growth and metastasis in various tumor types (47,67,68). CAFs are thought to largely affect tumor behavior via ECM modelling, secretion of soluble factors and promoting angiogenesis (67). In the case of GC, whilst some studies have investigated CAF-associated signaling pathways in tumor cells, molecular events underpinning CAF function and heterogeneity remain poorly defined. In our dataset, we found that GC CAFs have at least 3 distinct subtypes (STF1-3), each expressing distinct subsets of canonical CAF markers. Moreover, FAP-high STF3 cells also exhibited high INHBA co-expression, a correlation we confirmed in multiple orthogonal settings inclusive of a clinical stage-wise gradation. Functionally, silencing of INHBA impacted FAP levels in GC CAFs, implying a potentially direct role for INHBA in regulating FAP expression, and consistent with TGFβ signaling as a mediator of INHBA (69). A functional role for INHBA in GC CAFs has also been demonstrated by a recent study in GC that used bulk RNA-sequencing of LCM-derived CAF samples (70). Clinically, patients with high INHBA-expressing tumors exhibited poorer survival outcomes in multiple GC cohorts.
consistent with a tumorigenic function for INHBA (71-73). INHBA encodes a subunit of activin and inhibin, which have been reported to play opposing roles in many facets of normal biology and disease (74). It is possible that high INHBA expression levels may facilitate the formation of INHBA homodimers, otherwise known as Activin A, leading to the activation of activin receptors such as ACVR1 which has established roles in cancer (75). In other cancer types, aberrant increases in INHBA expression have been reported in both the epithelial and CAF components, involving autocrine and paracrine functions (76-79). Our observations put forth the INHBA pathway as a potential target to disrupt CAF function, and warrant testing of drug modalities in appropriate model systems (80).

In most parts of the world, Lauren’s histopathological subtypes is frequently employed as a reference standard in gastric clinical pathology. Evolution of intestinal-type GCs is usually ascribed to the Correa cascade, and the etiology and molecular features of diffuse-type GC are not well understood (81). In our study, scRNA-seq enabled lineage-based comparisons of the TME between these histological subtypes, leading to the discovery of increased plasma cells in diffuse-type GCs. Compared to T-cells, there are fewer studies on B-cell and plasma cell populations in GC (82,83). Derks et al, reported a higher proportion of tertiary lymphoid structures in GS tumors with enrichment of B-cells and CD4 T-cells (84). One report by Katoh et al, indicated a finding of increased B-cell lineages (which includes plasma cells) in diffuse-type and genome-stable GCs using the TCGA dataset and identified sulfated glycosaminoglycan as a key functional B-cell antigen in these tumors (85). Notably, B-cell/plasma cell infiltration has been associated with pro-tumor (i.e. immune suppressive) and anti-tumor (i.e. immune active) growth in other cancers (86). For GC, we speculate a ‘pro-tumor’ function for plasma cell infiltration, since increased plasma cells are associated with diffuse-type GCs that respond poorly to ICIs (87,88) and exhibit classical “immune-suppressive” features (89). To identify mechanisms responsible for the plasma cell increase, we investigated KLF2, a transcription factor previously shown to modulate multiple myeloma cell adhesion and homing of plasma cells (41). We observed that KLF2 expression in diffuse-type epithelial cells (EpiC cluster) was positively correlated with plasma cell proportions, and also increased KLF2 expression in diffuse-type and genome-stable GCs (Supplementary Table 15) in the TCGA cohort were consistent with a regulatory role for epithelial KLF2 in shaping plasma cell populations. We
functionally explored this possibility in “humanized mice”, by xenografting human derived diffuse-type GC cell lines that were KLF2-positive and negative. We found that only tumors with KLF2 had human plasma cell markers, implying a more casual role for epithelial KLF2 expression in plasma cell recruitment. Using DSP, we also gained spatial insights of KLF2 expression on primary samples and observed a trend for higher KLF2 expression in epithelial cells proximate to plasma cells, further suggestive of a functional interaction between these cell types. The precise mechanisms underlying the plasma cell enrichment remain undefined. Earlier studies have reported that KLF2 can modulate pro-inflammatory gene expression in monocytes and endothelial cells (90-92), and our findings suggest a role for KLF2 in driving tumor epithelial cell programs possibly involving paracrine cell signaling pathways (e.g. exosome-derived miRNAs (93-96)). Interestingly, besides KLF2, we also found that epithelial cells in diffuse-type tumors (EpiC cluster) displayed a striking upregulation of multiple immune-signaling pathways, including immunoglobulin genes and JCHAIN, a gene responsible for the secretory form of immunoglobulin, mimicking gene expression signatures of immune cells. This observation of an ‘epithelial-immune cell state’ has been described in other cancers (97,98). We report the existence of this phenomenon for the first time in diffuse-type tumors, adding a novel and singular dimension to gastric epithelial cell behavior. Probing the mechanisms underlying epithelial-plasma cell cross talk in diffuse-type GCs may pave way for new intervention strategies to modulate TME for this recalcitrant subtype.

Amongst the platforms to study GC, PDOs have recently emerged as a promising system for ex vivo testing of therapeutic agents, precision oncology and assessing driver gene function (99,100). Our comparison of PDO-primary samples using scRNA-seq revealed that PDOs indeed maintained most major cell types, except for a depletion in lymphoid and plasma cell lineages. However, it should be noted that these PDOs represent a single snapshot and assessment of organoid cellular compositions over time represents an area of future research. Global gene expression analyses indicated greatest changes in plasma cells compared to epithelial state. This suggests that PDOs may serve well to study gastric epithelial biology, whereas immune cells may be more affected by the process of culturing. These factors must be considered when using PDOs for personalized ex vivo drug testing, especially those involving immunological pathways or characterizing TME.
differences. Interestingly, our data reconfirms the association of ARID1A-loss with the induction of mucinous phenotype, an observation first made by Lo et al (100) (Supplementary Figure 7C).

In conclusion, in one of the largest single-cell analyses of GC performed to date, our study forms a unique resource for generating novel biological insights on tumor cell types, subtype-based TME compositions and cell-cell interactions in gastric tumors. In terms of future directions, our data supports the need for further in-depth studies on plasma cell homing biology guided by epithelial-KLF2, and the potential clinical implications of perturbing these interactions. The role of INHBA, and the TGF-ß super family, in the regulation of CAFs also carries potential for therapeutic target and/or predictive biomarker discovery. We anticipate future work to utilize combinatorial single cell approaches, including epigenetic, genetic, and transcriptional layers and spatial context to enhance our understanding of the gastric tumor architecture.
Methods

Ethics declaration. The study was approved by the local ethics board (National Healthcare Group, Domain Specific Review Board Ref No: 2005/00440, 2016/00059). All animal experiments and procedures were approved by the Institutional Animal Care and Use Committee (IACUC; IACUC# 191440) of A*STAR in accordance with guidelines of Agri-Food and Veterinary Authority and the National Advisory Committee for Laboratory Animal Research of Singapore. Primary CAFs and NFs were isolated from patients with gastric cancer who underwent gastrectomy without preoperative treatment at Kumamoto University after written informed consent was obtained from each patient. The study was approved by the Medical Ethics Committee of Kumamoto University (Approval Number: 1277). PBMCs were isolated from human blood obtained from healthy volunteers with written informed consent. Study protocols were approved by the SingHealth Centralised Institutional Review Board (CIRB reference number: 2017/2806). Protocols were performed in accordance with the Declaration of Helsinki for Human Research.

Sample acquisition and tissue processing. Patients diagnosed with gastric adenocarcinoma and undergoing surgical resection or endoscopy at the National University Hospital, Singapore were enrolled after obtaining written informed consent. On-table endoscopic biopsies or surgical resection samples were harvested. For surgical samples, matched normal gastric tissues from sites displaced at least several centimetres from the tumor were used. Tissue from distant metastases to the peritoneum were taken during diagnostic laparoscopy. Tissues were collected in MACS tissue storage buffer (Miltenyi Biotec, Germany) immediately after biopsy or resection and stored on ice. Samples were processed using enzymatic and mechanical dissociation using a human tumor dissociation kit and the Gentle MACS Octodissociator (Miltenyi Biotec, Germany) following the manufacturer’s ‘37_h_TDK_2’ program. Dissociated cells were passed through a MACS smartstrainer (70um) and incubated with RBC lysis buffer for 5 minutes followed by PBS neutralization. All centrifugation steps were carried out at 300g for 7 minutes. Dissociated cells were washed twice in PBS + 1% BSA, and filtered through a 40 μm smartstrainer. Live cell counts were obtained by manual cell counting using 1:1 trypan blue dilution. Cells were concentrated to 800-1200 live cells/μl and then processed for single-cell analysis.
Single-cell sequencing. Samples from each patient were processed in a single batch for library preparation. The Chromium Single Cell 5′ Library & Gel Bead Kit (10x Genomics) was used according to manufacturer protocols. Briefly, gel bead-based emulsions (GEMs) were generated by combining barcoded Single Cell 5’ Gel Beads, cells, and partitioning oil. 10x Barcoded, full-length cDNAs generated from GEMs were amplified by PCR. Enriched libraries were enzymatically digested, size-selected and adapter ligated for sequencing. To obtain TCR repertoire profile, VDJ enrichment was carried out as per the Chromium Single Cell V(D)J Enrichment Kit, Human T cell #1000005 (10x Genomics) using the same input samples. Sequencing libraries were generated with unique sample indices for each sample and quantified using Kapa library kit. Quantified libraries were sequenced on an Illumina Hiseq4000 sequencer.

Bioinformatic QC, normalization, clustering and differential gene expression of single-cell data. Cellranger v3.0 (https://support.10xgenomics.com/single-cell-gene-expression/software/) was used to align FASTQ sequencing reads to the hg38 reference transcriptome, generating single cell feature counts for each sample. Using Seruat version 3.0 (101), each sample was considered for genes/features shared by 3 or more cells, and cells showing 500 or more features and less than 6000 features. Cells with mitochondrial RNA percentages of >20 were filtered out. SCtransform normalization was performed on each sample dataset separately, along with regression of mitochondrial RNA as a variable. Single cell datasets were combined by first developing an initial reference dataset of 18 gastric samples with high cell numbers encompassing different characteristics of the primary and peritoneal metastasis samples for identifying anchors between pairs of samples, and subsequently integrating the remaining samples as ‘query’ datasets (18). Subsequently, PrepSCTIntegration was run to select features for downstream integration, and FindIntegrationAnchors to identify anchor genes. The integrated data were scaled, and Principal Component Analysis was performed. Data were visualized using Uniform Manifold Approximation and Projection (UMAP) software. Cell clusters were identified by a shared nearest neighbor (SNN) modularity optimization based clustering algorithm set at a resolution of 1. To identify differentially expressed genes for cluster demarcation, FindAllMarkers module was used and genes expressed in more than 25% of the cells in each cluster were
selected. Differentially expressed genes from each cluster were compared to sets of previously described cell type markers to assign cell identities, lineages, and sub-lineages. The Addmodulescore function was used to calculate composite module scores for different gene expression programs in each single cell. These scores were calculated by averaging the expression levels of all the genes in each program at the single cell level, subtracted by the aggregated expression of randomly selected control feature sets. Supplementary Table 2 lists the genes used for the various gene expression programs/modules. For organoid analysis, the reference 18 scRNA sample dataset (above) was integrated with the remaining primary, peritoneal and organoid datasets, resolving cell clusters by SNN clustering at a resolution of 1.25. Organoid data were integrated with primary samples datasets using canonical correlation analysis (101). Pathway analysis was performed using the REACTOME database (https://reactome.org/) separately for differentially upregulated and downregulated genes (Adjusted p-value <0.05) for each cluster. Pathways with entities p-value and FDR < 0.05 were investigated. Using these pathways, an overlap analysis was performed between epithelial meta-cluster and sub-lineages. For individual re-clustering of metaclusters we used LIGER (102) which relies on integrative non-negative matrix (iNMF) factorization to identify shared and dataset-specific factors. Briefly, we first selected variable genes, scaled them without centring followed by iNMF and quantile normalisation of factors using the selectGenes, scaleNotCenter, optimizeALS (k = 30) and quantileAlignSNF modules of LIGER. These iNMFs were added to Seurat for UMAP and clustering visualization.

**Doublet analysis.** scRNA-seq data are commonly affected by technical artifacts known as “doublets,” which limit cell throughput and lead to spurious biological conclusions such as discovery of mixed linages. DoubletFinder was used to identify doublets using gene expression data (30). DoubletFinder predicts doublets according to each real cell’s proximity in gene expression space to artificial doublets created by averaging the transcriptional profile of randomly chosen cell pairs. After identifying doublet cells, these were removed from the dataset.

**Molecular subtyping:** TCGA molecular subtyping was performed as previously described (103). Briefly, GCs were first classified as EBV or MSI using EBV-encoded
RNA in-situ hybridization (EBER–ISH) and mismatch repair (MMR) protein immunohistochemistry (IHC).

MMR IHC: Tissue were cut into 4-µm thick sections. The MMR panel consisted of monoclonal antibodies – mouse anti-MLH1 (Ventana, M1), mouse anti-PMS2 (Ventana, A16-4), mouse anti-MSH2 (Ventana G219-1129) and rabbit anti-MSH6 (Ventana, SP93). The panel is performed using Ventana Optiview DAB Detection kit, Ventana Optiview Amplification Kit and ancillaries on the fully-automated Roche Ventana Ultra instrument. MMR loss was determined when the tumor showed loss of expression for the examined MMR proteins. Normal tissue adjacent to the tumor was used as a positive internal control.

EBER–ISH: Tissue were cut on 4-µm thick sections. The EBER Probe is a fluorescein-conjugated oligonucleotide Ready to use ISH probe from Leica BOND, and with hybridization for 2 hours. Detection was performed with anti-FITC antibody, Leica Bond DAB polymer kit and ancillaries on the fully-automated Leica Bond III Biosystem. EBER positivity was determined by localization of the EBER signal within the nucleus of the tumour cells.

Whole exome sequencing (WES) of remaining samples performed as previously described (104). Briefly 100ng of DNA from each sample were sheared, adapter ligated, exome captured using the Agilent Sureselect Human all exon kit v6, and sequenced using Illumina Hiseq4000 platform. Tumors then classified as either CIN or GS based on the copy number profiles inferred from WES using GISTIC2 (105) or scRNA-seq. Tumors were clustered with a uniformly processed in-house dataset of >200 GCs, using Euclidean distance and Ward’s method, based on thresholded copy numbers from significantly altered sites identified by GISTIC2 (q<0.05). GC samples exhibiting elevated copy number changes were assigned as CIN. We confirmed significant associations between GS tumors with diffuse histological subtype (p = 0.03; Chi-square test), increased T-cell infiltrates in MSI-positive tumors (p = 0.021). Samples with whole genome doubling predicted by WES (thus classified as CIN) also exhibited high CNVs measured by scRNA-seq.

**Trajectory analysis.** Trajectory analysis was performed using Monocle 3.0 using two approaches (106,107). For plasma cell pseudo-temporal analysis, plasma cells
were extracted from the dataset. Monocle was used to preprocess, align, perform UMAP directionality reduction, cell clustering, and to develop learned graphs and cell orders. To compare trajectories between normal and tumor cells, the Seurat wrapper for Monocle 3.0 was used to cluster cells and develop learned graphs. To assess the robustness of inferred trajectories, we performed bootstrapping by randomly down-sampling samples by 10 percent and computing binomial P-values for each major branching node.

**CNV analysis.** To identify large-scale chromosomal copy number variations, InferCNV (https://github.com/broadinstitute/inferCNV) was used to explore tumor scRNAseq compared with normal scRNAseq datasets, and between diffuse and intestinal-type scRNAseq datasets. Expression intensities in epithelial cells of genes across positions of the genome were compared to a set of reference T and B-cells. Genes with a mean count number of 0.1 or less across cells were excluded. In addition, the hidden Markov method (HMM; ‘type i6’) was used, applying noise-filters and ‘subcluster’ analysis. For each sample, the gene expression of cells were re-standardized and values were limited from −1 to 1. Since HMM model uses values which are not centred on 1 but 6 states, to make gains and losses symmetrical we merged all ‘state 6’ states with ‘state 5’ before re-standardisation. The final CNV score of each cell was calculated as the sum of the absolute values of each CNV region. We also used CONICSmat for CNV calling (108). Briefly, after filtering out uninformative genes a normalisation factor was calculated for each cell and the average expression in each cell was centered using the calculated normalization factor. The z-score of the centered gene expression across all cells was calculated. Based on these z-scores, a Gaussian mixture model was calculated to determine regions showing an average gene expression bimodal distribution across cells. Only results for regions harboring more than 100 expressed genes were calculated to ensure the predictions were not influenced by a few differentially expressed genes in a small region. To avoid batch effects, we called CNV for each sample separately on the gene expression matrix on all cells.

**TCR analysis.** Cellranger v6.0 was used (using mkfastq module) to align FASTQ sequencing reads to the hg38 reference transcriptome. To generate single-cell V(D)J sequences and annotation for each sample we used cellranger vdj module. Basic
statistics and diversity index was estimated using immunarch (v0.6.6) (https://immunarch.com/index.html).

**Digital Spatial Profiling and analysis.** Gastric cancer and normal FFPE tissues were mounted on super frost slides to validate the spatial profiling of RNA by the NanoString GeoMx DSP platform. For H&E staining, FFPE slides were deparaffinized with histoclear, rehydrated and stained with Hematoxylin Solution. Slides were counterstained with Eosin and mounted with mounting media. FFPE slides were subjected to conventional tissue pre-processing (deparaffinization and rehydration). For RNA profiling, the transcriptomics cancer atlas with a 1800-plex RNA probe set was used along with an additional set of 12 genes (Supplementary Table 16). Standard fluorescence-labelled morphology marker panel consisting of Pan-CK for epithelial regions, CD138 for plasma cells, α-Smooth muscle actin (SMA) for fibroblast and nuclear stain were used as ROI selection references. 12 ROIs (4 for each region) measuring 300um diameter on each slide were drawn and selected. All oligos from the selected ROI were collected into 96-well plates using the proprietary UV-guided technology in the DSP approach. Resultant oligos representing individual targets for individual ROIs were sequenced using Illumina sequencing. Data was analysed by uploading the counts dataset from the Illumina run into the GeoMx DSP analysis suite. Biological probe quality control was performed using default settings. Scaling was performed using geometric means and normalization using Q3 averages of housekeeping genes.

**Humanized mouse model.** Details of the humanized mouse model experiment have been described previously (109). One to three day-old NSG pups were sub-lethally irradiated at 1 Gy and engrafted with $1 \times 10^5$ human CD34$^+$ cordblood cells (HLA-A24:02; Stemcell Technologies) via intrahepatic injection. Mice with more than 10% human immune cell reconstitution (calculated based on the proportion of human CD45 relative to the sum of human and mouse CD45) were included in the study. In total 2 diffuse-type GC cell lines of HLA-A24:02 subtype were selected for the experiment (KLF2 positive - GSU and KLF2 negative – SNU1750)(110,111). For each cell-line, 5 humanized-mice and 5 NSG mice were injected with the tumor cells and observed for one month. Mice were sacrificed at the end of one month, necropsies performed, and tumors harvested and sequenced (bulk RNA-Seq).
**Immunohistochemistry for IRF4.** Immunohistochemistry analysis was performed to evaluate the expression of IRF4 in diffuse and intestinal GC samples using MUM1 (IRF4) primary antibodies (Clone MUM1p, DAKO, IS64430). Immunohistochemical stains of 4 μm paraffin sections of patient samples were performed using a Bond Max automated immunostainer (Leica Biosystem, Melbourne, using antibodies for IRF4 (MUM1p, 1/500; Dako). Briefly, sections were deparaffinized and rehydrated followed by antigen retrieval at pH 9.0 (Tris buffer) for 36 minutes. Sections were treated with peroxidase solution followed by incubation with MUM1 primary antibody for 60 min according to the manufacturer's protocol. Tissue sections were further incubated with horseradish peroxidase-labeled polymer secondary antibody and localization was performed with horse radish peroxidase-labeled polymer with DAB using a bond polymer refine detection kit (Leica Bondmax, Leica Biosystems, Nussloch GmbH) according to standard protocols. Appropriate positive controls were immunostained in each batch of IHC. Gastric epithelial cells served as an internal negative control. MUM1 expression was predominantly observed in the nuclei of plasma cells along with some weak expression in cytoplasmic regions Strong nuclear expression was considered as positive staining whereas standalone cytoplasmic staining was considered negative. For each case, 10 fields (at 20x magnification) were selected randomly and MUM1 positive cells were counted in each field using the ImageJ software. A mean value of positive MUM1 expression was calculated for each case.

**Fibroblast cell culture experiments.** Human gastric fibroblast cell lines were derived from surgical gastric tissue of patients with GC who underwent gastrectomy without preoperative treatment with written informed consent from each patient. CAFs were established from the tumoral gastric wall, and normal fibroblast (NF’s) were from non-tumoral gastric wall according to previous established protocols (112,113). Fibroblast cell lines were cultured in RPMI 1640 containing 10% FBS (normal medium) and incubated at 37°C with 5% CO2. INHBA expression was downregulated by transfecting cells with predesigned Silencer Select siRNAs directed against INHBA (#1, s7434; #2, s7436; Cat# 4427037, Thermo Fisher Scientific) and a non-targeting siRNA (#4390843, Thermo Fisher Scientific) was used as the negative control. CAFs were transfected with annealed siRNAs for 48
hours (5 μM) using Lipofectamine RNAiMAX (#13778-150, Thermo Fisher Scientific).

Real time PCR primers for

**INHBA:** F 5’ AGCTCAGACAGCTTTACCACA 3’; R 5’ TTTTCTTCTCCTCTTCAGCA 3’

**FAP:** F 5’ TGGCGATGAAACAATATCTAGA 3’; R 5’ ATCCGAACAACGGGATTCTT 3’

NFs were seeded and incubated at 37°C with 5% CO2. After 24 hrs seeding, the cultured cells were treated with 100 ng/ml recombinant human Activin A (Biolegend) for 48 and 96 hrs respectively. Collagen genes were measured in recombinant INHBA-treated and control fibroblast lines using the following primers:

**COL1A1:** F 5’ AGACGAAGACATCCCACCA 3’; R 5’ GTCATCGCACAACACCTTG 3’

**COL1A2:** F 5’ CTGGAGAGGCTGGTACTGCT 3’; R 5’ GAGCACAAAGAAGACCCTGA 3’

**COL6A3:** F 5’ ACCGTCCAACGGTACTCCT 3’; R 5’ CTCTTGCCACCACACCTGG 3’

All experiments were performed in triplicate. Data are presented as mean ± standard deviation (SD). Statistical significance was determined using a 2-tailed Student’s t-test

**KLF2 knockdown model.** LMSU cells were transfected with pooled siRNAs for KLF2 (Dharmacon, L-006928-00-0020 ON-TARGETplus Human KLF2 (10365) siRNA -SMARTpool) or negative control siRNAs (Dharmacon, D-001810-10-20 ON-TARGETplus Non-targeting Control Pool) at 100nM using Lipofectamine 2000 reagent (Thermo Fischer Scientific, 11668027). To obtain stable knockdown of KLF2 in GSU cells, cells were plated at a concentration of 500,000 cells in a 10 cm dish and infected with MISSION lentiviral Transduction particle encoding PLKO.1-PURO-CMV- tGFP -KLF2 (clone ID TRCN000418423) or GFP (as non-targeting control). After 48hrs, transfected cells were selected with 2 μg/ml puromycin. The transfected cells were then expanded and selected in culture medium plus puromycin (1μg/ml) for 3 weeks to obtain stable KLF2-knockdown cells. Total RNA was extracted using the Qiagen RNAeasy mini kit according to manufacturer’s instructions. RNA was converted to cDNA using Improm-ITM Reverse Transcriptase (Promega) or iScript CDNA synthesis kit (Biorad). Quantitative PCR was performed in triplicate using Quantifast SYBR Green PCR kit (Qiagen) on an Applied Biosystems HT7900 Real Time PCR System for 40 cycles using the following primers with an annealing temperature of 55°C:

**KLF2:** F 5’CGGCAAGACCTACCAAGAGT3’; R 5’CGCAGATGGCACTGGAATG3’
**ACTB:** F 5' CATGTACGTTGCTATCCAGGC 3'; R 5' CTCTTAATGTCACGCACGAT 3'

Fold change was calculated using the Delta-Delta Ct method. Total intracellular protein was extracted with RIPA lysis buffer (Thermofisher). Protein concentration was determined using a bicinchoninic acid (BCA) assay kit (Thermofisher).

**Western blotting.** Equal amounts of protein from each sample were separated on 4-12% sodium dodecyl sulfate polyacrylamide gel (SDS-PAGE) and electrotransferred onto polyvinylidene difluoride (PVDF) membranes. Membranes were blocked with 5% fat-free dry milk in 0.1 M Tris-buffered saline-0.1% Tween-20 buffer (TBST) for 1 h. The membranes were sequentially incubated with a KLF2 primary antibody (Affinity Bio, dilution of 1 in 1000) or GAPDH primary antibody (Proteintech, dilution of 1 in 3000) overnight and horseradish peroxidase-conjugated secondary antibody (anti-rabbit and anti-mouse, Santa Cruz Biotechnology) for 1 h. Blots were developed with an enhanced chemiluminescence reagent (Amersham Biosciences, USA) and quantified by densitometric scanning and analyses using a ChemiDoc system (BioRad).

**Transwell migration assay.** Transwell migration assays were performed with human plasmablasts/plasma cells isolated from primary human peripheral blood mononuclear cells (PBMCs) or KMS-11 cells, a human multiple myeloma cell line. Briefly, human plasmablasts/plasma cells from PBMCs were isolated using the Plasma Cell Isolation Kit II according to manufacturer’s instructions (Cat No: 130-093-628, Miltenyi Biotec). The purity of isolated plasmablasts/plasma cells was assessed by flow cytometry using CD19 and CD38 antibodies (BD Biosciences) whereby peripheral blood plasmablasts/plasma cells were defined as CD19+ CD38+ cells. We loaded 2 × 10^4 cells per well into the top chamber, and for KMS-11 cells, 2.5 × 10^5 cells were resuspended in 200 µL of the migration buffer and loaded onto the upper chamber of transwell inserts (Transwell Permeable Support with a 5.0-µm polycarbonate membrane, 6.5-mm insert; 3421, Corning). GC cells were resuspended in 600 µL of migration buffer (0.5% BSA-RPMI 1640) and seeded into the bottom chamber of 24-well transwell plate (Cat No: 3421, Corning) and allowed to settle for 4-6 hours. Recombinant human CXCL12/SDF1a (350-NS-010-CF, R & D systems) resuspended in 600 µL migration buffer (Final concentration of 200 ng/mL) or migration buffer alone were used as positive and negative controls respectively. 20-24 hours later, unmigrated cells in the top chamber were counted. Unmigrated...
cells were subtracted from total cells seeded and the proportion of migrated cells were calculated. Pair-wise comparisons were performed using the Mann-Whitney test.

**RNAScope.** RNAScope in situ hybridization (ISH) in gastric cancer samples was performed according to manufacturer protocols (114). Briefly, FFPE slides were baked at 60 °C for 1 hour before being deparaffinised in xylene and 100% ethanol. After drying slides for 5 min at RT, H₂O₂ was added for 10 min at RT. For antigen accessibility, slides were incubated in boiling antigen retrieval solution (<98 °C) for 15 min, washed in water twice, dehydrated in 100% ethanol and finally treated with Protease Plus for 30 min at 40 °C. Probes were then hybridized for 2 h at 40°C followed by RNAScope amplification and chromogenic detection. RNA scope 2.5 HD Duplex detection kits were used for simultaneous visualization of two RNA targets in single cell resolution using Hs-PLVAP HRP-GREEN (437461) and Hs-RGS5-C2 AP-FAST RED (533421, ACD bio). Sections were counterstained with hematoxylin and mounted with Vectamount. Duplex probes targeting CI-PPIB and C2 -POLR2A (322435) were used as positive control probes, and dihydrodipicolinate reductase (dapB), a bacterial gene (310043) as a negative control probe.

**Analysis of bulk-RNA seq data.** To assess cellular abundances in bulk tissue transcriptome profiles, CIBERSORTx was used to estimate cellular abundances of B-cells and plasma cells in intestinal and diffuse TCGA-STAD datatsets downloaded from FireBrowse (9,115). For reference signature gene matrices, NSCLC PBMC scRNAseq datasets provided with the CIBERSORTx suite were used. To compare single cell RNA data with gastric bulk-RNA data, combined TCGA and GTEx gene expression data sets were used from the Xena browser (116). For survival analysis of INHBA, pooled survival analysis of several available GC microarray datasets was performed using the Kaplan Meier plotter tool, GC subgroup (https://kmplot.com/analysis/index.php?p=service&cancer=gastric). Both INHBA probes (Affy ID: 204926_at and 210511_s_at) were pooled and the mean expression of both probes were used to generate INHBA high and low subgroups by dividing samples at the median. For the survival analysis of INHBA in TCGA, samples were divided in INHBA high and low at the median value and adjustment of stage was performed using Cox regression.
Analysis of **KLF2** cis-regulation. H3K27ac ChIP-seq and RNA-seq of 24 primary GC samples were performed and analyzed as previously described (117). To explore the **KLF2** locus in greater detail, we first identified **KLF2** promoter regions using an in-house GC promoter catalogue derived from the H3K27ac ChIP-seq of 26 in-house GC cell lines. For each primary GC sample, the input subtracted H3K27ac signal at the **KLF2** promoter region was then computed using bigWigAverageOverBed to yield reads per kilobase per million (RPKM).

**Generation, maintenance and single-cell sequencing of GC PDOs.** Human gastric tissues were biopsied from tumor and matched adjacent normal sites of each patient during surgical intervention. Tissues were processed as previously described (118). Briefly, tissues were minced and washed in Dulbecco’s phosphate-buffered saline (DPBS; Thermo Fisher Scientific, Waltham, MA), and digested in DPBS containing 1 mg/ml collagenase (Sigma-Aldrich, Saint Louis, MI) and 2 mg/ml bovine serum albumin (BSA; Sigma-Aldrich) for 30 minutes at 37°C. Digested tissues were passed through 30 μm filters (Miltenyi Biotec, Germany). Filtered cells were pelleted at 300 × g for 5 minutes, resuspended in Matrigel (Corning Life Sciences, USA), and seeded into multiwell plates (Thermo Fisher Scientific, USA). Cultures were maintained in custom gastric PDO culture medium at 37°C in 5% CO₂ and monitored daily for organoid generation. The culture medium in each well was replaced with fresh medium on alternate days. PDOs were passaged once every 7-10 days at a 1:3 ratio. The median establishment time to the respective passage at time of sequencing was 17 weeks (range 17 to 30 weeks; passage numbers: 9 to 11). Gastric organoids were harvested from gel matrices by washing briefly with DPBS, incubating with Trypsin-EDTA at 37 °C for up to 30 min, and pelleted at 300 g for 5 min. Supernatants were discarded and cell pellets were washed twice with 10 mL DPBS each and filtered through cell strainers (mesh size: 30 μm). After centrifugation at 300 g for 5 min, the supernatant was discarded, and cells were washed with 1X DPBS and then re-suspended at ~1000 cells/μL in 1X PBS containing 0.4% BSA. Organoid scRNA-seq libraries were prepared using the 10X Genomics Single Cell 3’ Gel Bead and Library Kit.

**Statistical analysis.** All analyses were done using R (V.4.0.3) with statistical significance set at p<0.05 adjusted for multiple testing. Wilcoxon-rank sum test was used to evaluate associations with continuous variables. Student’s t-test was used to
evaluate associations with parametric continuous variables. Bi-variate correlation analysis was performed using Pearson or Kendall’s Tau (clinical stage correlation). Significance of overlapping CNVs called by WES and scRNA-seq was assessed using hypergeometric distribution tests by phyper (lower.tail = FALSE) in R (V.4.0.3). Kaplan-Meier curves with log-rank statistics were used to compare overall survival (OS). Pearson correlation analysis for TCGA bulk-RNA seq was performed using cbioportal. We computed Jaccard Similarity Indexes (JSI) for the top 30 significant Reactome programs (59) between PDOs and primary tumors by their meta-clusters.

Data availability

scRNA-seq data has been uploaded to the GEO repository:

Acknowledgements

We thank NanoString Technologies USA for Digital Spatial Transcriptomics Platform expertise, Dr. Anita Chan from Singapore Eye Research Institute, Dr. Anjali Verma, Dr. Amrita Roy for assistance with RNAscope, and the Duke-NUS Genome Biology Facility (DGBF) for assistance with sequencing. Certain figures were created using BioRender.com, with relevant licenses obtained.

Author contributions

Funding acquisition: P.T, R.S, J.S, Y.W.P
Methodology: V.K, K.R, P.T
Project administration: K.R, A.L.K.T
Supervision: PT
Writing – original draft & Visualization: V.K, K.R, R.S, N.P
Writing – review & editing: PT
Approval of final version of manuscript: All authors
References


78. Cangkrama M, Wietecha M, Mathis N, Okumura R, Ferrarese L, Al-Nuaimi D, et al. A paracrine activin A–mDia2 axis promotes squamous carcinogenesis via fibroblast...


**Figure Legends**

**Figure 1.** scRNA-seq of gastric tumor and normal samples defines 34 cell states including rare cell populations.

A. Schematic representation of experimental design and techniques used in this study. 31 unique GC patients undergoing surgical resection or endoscopy had tumor samples (n = 31) and adjacent normal (n = 11) harvested for analysis. Tumors ranged from Stage I to Stage IV and included samples of both primary tumors, distant (peritoneal) metastases, and matched normal gastric tissues. 29 tumors had scRNA-seq performed using the 10X platform (along with 11 adjacent normal). Four patients had PDOs generated from their tumors (4 tumors + 4 adjacent normal), which were also sequenced by 10X scRNA-seq. A subset of 13 samples also had DSP performed using the NanoString GeoMx platform (10 tumor + 3 normal). In total, more than 200,000 cells were sequenced in this study.

B. UMAP of 152,423 cells representing 34 unique cell-states color-coded by their corresponding cell lineage or subtype. Each dot in the UMAP represents a single cell.

C. Cell lineage compositions of GC and normal samples inferred by scRNA-seq data. The middle panel (bubble plot) shows the cell subclusters (rows) by: tumor vs normal, by stage, and GC histological subtype (diffuse vs. intestinal). The size of the circle represents the cell proportion of each specific cell lineage/type. The circles are color-coded by defined cell lineages/types as shown in Figure 1B. The stacked bar graph on the top shows the number of cells in each meta-cluster for each category. The histogram on the right shows the absolute cell numbers in each subcluster.

D. Cluster-cluster heat map of gene expression data of all 34 cell-states across all samples using Pearson correlation matrix. Darker colors correspond to higher correlation.

E. Pseudo-time analysis of plasma meta-cluster generated using Monocle. The trajectory was rooted against the plasmablasts. Pseudo-time analysis demonstrates different stages of plasma cell differentiation and maturation including plasmablasts, short-lived plasma cells and long-lived plasma cells.

F. Expression of PLVAP and RGS5 in endothelial (STE2) and fibroblast clusters (STF2, STF4). Doublets were identified and filtered out using DoubletFinder. PLVAP+ RGS5- cells are predominantly present in endothelial cluster (STE2).
PLVAP- RGS5+ cells are predominant in the fibroblast cluster (STF2). The STF4 cluster shows cells expressing PLVAP+ RGS5+ suggestive of a rare mixed lineage population.
Figure 2. scRNA-seq deconvolutes gastric tumor programs associated with distinct cell-states.
A. Density plot of UMAP representation comparing normal and gastric tumor samples after random down-sampling to ~30,000 cells each to allow statistical equivalence. Each dot represents a single cell. Dashed-lines highlight higher proportions of epithelial cells in normal samples and myeloid cells in tumor samples.
B. Split violin plot of epithelial-mesenchymal transition (EMT) oncogenic gene signature score in normal and tumor cells showing a significantly higher score in tumor cells.
C. Bubble plot depicting the expression of GC oncogenes in tumor epithelial cell clusters. The size of circle represents the percentage of cells expressing the gene in that specific epithelial cell cluster, while the color represents the average expression of the gene.
D. Box plot depicting CNV scores for epithelial (green) and macrophage cells (blue) in normal and tumor samples. CNV scores were computed using InferCNV. P values were computed using Wilcoxon rank-sum test.
E. Bar graph depicting differences in transcriptomic profiles between tumor and normal tissue by number of upregulated and downregulated genes in epithelial cell clusters.
F. Heatmap of LIPF and PGA3 gene expression (classical chief cell marker genes) in tumor versus normal samples. Darker color signifies higher expression.
G. Gene expression of LIPF by DSP analysis, in tumor epithelial cells (PanCK+) compared to normal samples (n = 13).
H. Bubble plot depicting sub-lineage specific expression of CAF marker genes FAP, CSPG4, PDGFA, ASPN, S100A4, COL8A1, THBS2 and CTHRC1 in fibroblast clusters. STF1 and STF3 are LUM associated fibroblasts, while STF2 comprises pro-angiogenic pericytes. The size of circle represents the percentage of cells expressing the gene in that specific fibroblast cell cluster, while the color represents the average expression of the gene.
I. Bubble plot depicting significant log fold differences in expression of genes between tumor and normal by meta-cluster mapped against the bulk-RNA seq data (five genes per meta-cluster are shown). The size of the circle represents the log fold change in the expression of specific genes.
**Figure 3.** Differential TME analysis between histological subtypes identifies increased plasma cells in diffuse-type tumors.

A. Density plot of UMAP representation comparing diffuse and intestinal GC samples after random down-sampling to ~25,000 cells each. Each dot represents a single cell. Dashed-lines highlight a higher proportion of plasma cells in diffuse GC.

B. Immunohistochemical staining of IRF4 expression in diffuse (n = 5) and intestinal (n = 12) GC samples (scale bar 400um). Bar graph showing significantly higher average IRF4 IHC score in diffuse compared to intestinal GC.

C. Plasma cell proportions deconvoluted by CIBERSORTX in diffuse and intestinal GC in the TCGA dataset (diffuse n = 63, Intestinal n = 167).

D. Bar graph showing enrichment of plasma cell proportions in PL4 and PL5 plasma subclusters in diffuse vs. intestinal GC single-cell samples.

E. Pearson correlation plot showing significant positive correlation of plasma cell proportion to average KLF2 expression in the epithelial meta-cluster.

F. Bee swarm plot showing increased KLF2 expression in diffuse vs. intestinal GC samples in the bulk RNA-seq TCGA STAD dataset.

G. Bar graph showing increased KLF2 expression in PanCK+ epithelial morphological regions of diffuse vs. intestinal GC by DSP (n = 10).

H. DSP analysis depicting epithelial ROIs proximal (top) and distal to plasma cells (bottom). Analysis is based on immunofluorescence staining for PanCK (epithelial, Green), CD138 (Plasma, pink), SMA (Fibroblast, Cyan), and DAPI (Blue). Each circular ROI is 300μm in diameter.

I. Bar graph showing expression of KLF2, IRF4 and SLAMF7 genes in GSU humanized and non-humanized mice against SNU1750 humanized and non-humanized mice.

J. Western blot (top panel) showing stable knockdown of KLF2 in GC cell line GSU (shKLF2) compared to shNT [non-targeting] control. Loss of KLF2 in GSU significantly reduces migration of plasma cells derived from peripheral blood mononuclear cells [PBMCs] and multiple myeloma cell line KMS-11 (N = 17) (bottom panel).

K. Upregulation of immunoglobulin genes in diffuse vs intestinal epithelial meta-cluster (top panel). Tree map shows the overlap of upregulated pathways in
epithelial meta-cluster vs. subclusters, with EpiC cluster showing the greatest overlap (bottom panel).

**Figure 4.** scRNA-seq enables identification of distinct GC fibroblast subtypes and INHBA-FAP axis as a CAF regulator.

A. Bubble plots demonstrating stage dependent increases in the proportion of fibroblast cells with STF3 being the dominant subcluster. The size of the circle represents the proportion of cells expressing subcluster specific genes.

B. Bubble plots showing fibroblast subclusters (STF1-3) expressing distinct CAF canonical markers (*FAP, CSPG4, ACTA2* and *TAGLN*). The size of the circle represents the proportion of cells expressing different genes.

C. Violin plot showing the expression of *INHBA* in STF2 and STF3 fibroblast clusters with negligible expression in the STF1 fibroblast cluster.

D. Fibroblast ROIs captured by DSP analysis based on immunofluorescence staining for PanCK (epithelial, Green), CD138 (Plasma, pink), SMA (Fibroblast, Cyan), and DAPI (Blue). Each circular ROI is 300μm in diameter.

E. Beeswarm plot showing differential expression of *FAP* in fibroblast ROIs of normal and tumor samples by DSP (*n* = 13).

F. Beeswarm plot showing differential expression of *INHBA* in fibroblast ROIs of normal and tumor samples by DSP (*n* = 13).

G. Pearson correlation graph demonstrating strong positive correlations between *INHBA* and *FAP* gene expression in fibroblast ROIs using DSP.

H. Bar graph showing significant expression of *FAP* and *INHBA* gene in flow sorted tumor fibroblasts compared to matched normal fibroblasts (*n* = 10 each).

I. Bar graph showing significant reduction in *FAP* gene expression after siRNA mediated *INHBA* knockdown in tumor fibroblast lines. Two independent siRNAs were used.

J. Bar graph showing significant increases in *FAP* gene expression in two normal fibroblast lines after treatment with recombinant INHBA for 48 and 96 hours respectively.

K. Bubble plot depicting stage dependent increases of *FAP*+ and *INHBA*+ cells in fibroblast cluster STF3 (*p* = 0.041). The circle sizes represent the relative proportion of cells expressing these genes. P values were computed using Kendall’s τ method.
L. Kaplan–Meier survival curves of TCGA STAD data showing significant differences in overall survival between INHBA high and low samples. P values were computed using log rank tests.

**Figure 5.** Comparative analysis of primary and organoid samples.
A. UMAP representation of ~200,000 cells (~48,000 cells from tumor PDOs with matched normal PDOs, combined with primary samples (~152,000 cells)) recapitulating the major five meta-clusters color-coded by their cell types. Each dot in the UMAP represents a single cell.
B. Violin plot showing the expression of GC gene module scores in tumor PDOs compared to matched normal.
C. Trajectory plot analysis of epithelial cells from tumor and normal PDOs demonstrating the expression of cellular differentiation gene programs in tumor PDOs depicted by long multiple branches.
D. Density plot of UMAP representation comparing PDO and primary gastric samples demonstrating enrichment of lymphoid and plasma meta-clusters in primary samples compared to PDOs.
E. Graph showing the number of upregulated and downregulated genes in PDOs vs primary samples in the five meta-clusters. The plasma meta-cluster shows the highest number of differentially expressed genes as compared to other meta-clusters.
F. Volcano plot of upregulated and downregulated genes in the plasma meta-cluster between PDOs and primary samples, showing significant downregulation of antibody mediated complement factor genes in PDOs. X-axis shows the -log10 adjusted p value and y-axis log2 fold change in gene expression.
G. Top common upregulated pathways in PDOs vs. primary samples across all meta-clusters.

**Figure 6.** Comprehensive single-cell atlas of gastric cancer.
This study included more than 200,000 cells from 31 primary gastric tumor samples. In total, 34 distinct cell-lineage states were identified, related by developmental trajectories and previously unreported rare cell populations. An increase in plasma
cell proportions were observed as a feature of diffuse-type tumors associated with epithelial-resident *KLF2*. A stage-wise accrual of novel cancer-associated fibroblast sub-populations marked by high *INHBA* and *FAP* co-expression. Findings were complemented using digital spatial transcriptomics and RNAscope. Our results provide a high-resolution molecular resource for GC translational studies, identifying intra- and inter-patient lineage-states across distinct GC subtypes.
Patients with Gastric Cancer

Stage I - IV
Intestinal $n = 14$
Diffuse $n = 6$
Mixed $n = 9$

Adjacent Normal $n = 11$

$n = 31$

$>200,000$ cells

10X scRNA-seq
Organoid generation

NanoString GeoMx DSP
(overlapping samples)

Plasma cell
B-cell
T-cell
NK cell
Mast cell
Dendritic cell
Macrophage
Fibroblast
Endothelial
EpiC
EpiInt
EpiPit

Number of cells
30,000
60,000
90,000

Cell Proportion

Meta-cluster
Cell Proportion

Subcluster

Plasma cell
B-cell
T-cell
NK cell
Mast cell
Dendritic cell
Macrophage
Fibroblast
Endothelial
EpiC
EpiInt
EpiPit

Number of cells
10,000
20,000
30,000

Adjacently Normal $n = 11$

Cancer Research.
Figure 3

A

Intestinal

Epithelial

Plasma

Stromal

Lymphoid

Diffuse

Intestinal

Epithelial

Plasma

Stromal

Lymphoid

B

Intestinal

IRF4

Diffuse

IRF4

Plasma cell proportion

R = 0.69, p = 0.0008

E

Average KLF2 expression (epithelial)

Plasma cell proportion

R = 0.69, p = 0.0008

C

Intestinal

Diffuse

Percentage of plasma cells

TCGA

p = 0.028

D

Intestinal

Diffuse

Proportion of plasma cells

PL1

PL2

PL3

PL4

PL5

p = 0.049

p = 0.036

EpiC1

EpiC2

EpiC3

EpiInt1

EpiInt2

EpiPit1

EpiPit2

Chief

Intestinal

Overlap

Upregulated pathways in epithelial meta-cluster (diffuse)

Log fold change

Overlap

Upregulated pathways in epithelial meta-cluster (diffuse)

K

Top upregulated genes in diffuse vs. intestinal epithelial meta-cluster

Expression of KLF2

Expression of SLAMF7

Expression of IRF4

% of plasma cells migrated

PBMCs

KMS-11

shNT

shKLF2

KLF2 (37 kDa)

GAPDH (37 kDa)

GSU

CD138

PanCK

SMA

Epithelial cells in proximity to plasma cells

CD138

PanCK

SMA

Epithelial cells distant to plasma cells

CD138

PanCK

SMA

TCGA

KLF2 gene expression

p = 0.00013

TCGA

p = 0.028

GSU Humanized

SNU1750 Humanized

GSU Non Humanized

SNU1750 Non Humanized

Expression of KLF2

Expression of SLAMF7

Expression of IRF4

% of plasma cells migrated

p = 0.001

GSU

K

Cancer Research.
**Figure 5**

**A** UMAP visualization showing distinct clusters for different cell types (Plasma, Epithelial, Myeloid, Stromal, Lymphoid) in PDO and Primary samples.

**B** Gastric cancer gene score distribution with a significant p-value of <0.0001.

**C** Comparison of cell distribution between normal and tumor tissues.

**D** Heatmaps showing cell proportions for PDO vs. Primary samples, with significant p-values indicated for each comparison.

**E** Bar chart illustrating the number of upregulated and downregulated genes in PDO vs. Primary samples.

**F** Heatmap and volcano plot showing log2 fold change and adjusted p-values for upregulated and downregulated genes in PDO vs. Primary samples.

**G** Top common upregulated pathways in PDO vs. Primary samples (FDR < 0.05):
- Cellular responses to stress
- Translation
- Cellular response to starvation
- SRP-dependent protein targeting membrane
- Response of EIF2AK4 to amino acid deficiency
- Selenocysteine synthesis
- Nonsense Mediated Decay (NMD)
- Peptide chain elongation
- Regulation of cholesterol biosynthesis
- Activation of gene expression by SREBF
Single-cell atlas of lineage states, tumor microenvironment and subtype-specific expression programs in gastric cancer

Vikrant Kumar, Kalpana Ramnarayanan, Raghav Sundar, et al.

Cancer Discov  Published OnlineFirst October 12, 2021.

Updated version  Access the most recent version of this article at: doi:10.1158/2159-8290.CD-21-0683

Supplementary Material  Access the most recent supplemental material at: http://cancerdiscovery.aacrjournals.org/content/suppl/2021/10/12/2159-8290.CD-21-0683.DC1

Author Manuscript  Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.