HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2022 January 04.

Published in final edited form as:

Nat Med. 2021 January ; 27(1): 141–151. doi:10.1038/s41591-020-1125-8.

Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma

Ruiping Wang[1], **Minghao Dang**[1], **Kazuto Harada**[2,12], **Guangchun Han**[1], **Fang Wang**[3], **Melissa Pool Pizzi**[2], **Meina Zhao**[2], **Ghia Tatlonghari**[2], **Shaojun Zhang**[1], **Dapeng Hao**[1], **Yang Lu**[4], **Shuangtao Zhao**[1], **Brian D. Badgwell**[5], **Mariela Blum Murphy**[2], **Namita Shanbhag**[2], **Jeannelyn S. Estrella**[6], **Sinchita Roy-Chowdhuri**[6], **Ahmed Adel Fouad Abdelhakeem**[2], **Yuanxin Wang**[1], **Guang Peng**[7], **Samir Hanash**[7], **George A. Calin**[8], **Xingzhi Song**[1], **Yanshuo Chu**[1], **Jianhua Zhang**[1], **Mingyao Li**[9], **Ken Chen**[3], **Alexander J. Lazar**[6,10], **Andrew Futreal**[1], **Shumei Song**[2], **Jaffer A. Ajani**[2,✉], **Linghua Wang**[1,11,✉]

[1]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[2]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[3]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[4]Department of Nuclear Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[5]Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[6]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

✉**Correspondence and requests for materials** should be addressed to J.A.A. or L.W. jajani@mdanderson.org; LWang22@mdanderson.org.

Author contributions

L.W. and J.A.A. conceived and jointly supervised the study. S.S., K.H., M.P.P., M.Z., G.T., N.S., A.A.F.A., B.D.B. and M.B.M. contributed to sample collection and processing and collection of patient clinical information. A.J.L., J.S.E. and S.R.-C. contributed to pathology review. Y.L. reviewed the CT images. L.W. supervised the bioinformatics data analysis, data integration and interpretation. R.W. contributed to sequencing data processing, integrative analyses and generation of figures and tables for the manuscript. M.D., G.H., F.W., S. Zhang., D.H., S. Zhao., Y.W., X.S., Y.C., J.Z., M.L. and K.C. assisted with data processing and analysis. L.W., J.A.A. and R.W. wrote the manuscript. L.W., J.A.A., R.W., A.J.L., A.F., S.H., G.A.C. and G.P. revised the manuscript.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-1125-8.

Competing interests

The authors declare no competing interests.

Extended data is available for this paper at https://doi.org/10.1038/s41591-020-1125-8.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-1125-8.

Peer review information Javier Carmona was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

[7]Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[8]Department of Experimental Therapeutics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[9]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

[10]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[11]UTHealth Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[12]Present address: Department of Gastroenterological Surgery, Kumamoto University, Kumamoto, Japan

Abstract

Intratumoral heterogeneity (ITH) is a fundamental property of cancer; however, the origins of ITH remain poorly understood. We performed single-cell transcriptome profiling of peritoneal carcinomatosis (PC) from 15 patients with gastric adenocarcinoma (GAC), constructed a map of 45,048 PC cells, profiled the transcriptome states of tumor cell populations, incisively explored ITH of malignant PC cells and identified significant correlates with patient survival. The links between tumor cell lineage/state compositions and ITH were illustrated at transcriptomic, genotypic, molecular and phenotypic levels. We uncovered the diversity in tumor cell lineage/state compositions in PC specimens and defined it as a key contributor to ITH. Single-cell analysis of ITH classified PC specimens into two subtypes that were prognostically independent of clinical variables, and a 12-gene prognostic signature was derived and validated in multiple large-scale GAC cohorts. The prognostic signature appears fundamental to GAC carcinogenesis and progression and could be practical for patient stratification.

GAC remains a common and lethal disease with a poor prognosis[1]. Often diagnosed at an advanced stage, GAC is frequently resistant to therapy[2]. A common site of metastases is the peritoneal cavity (peritoneal carcinomatosis; PC) and there is a high unmet need for improved therapeutic interventions in patients with advanced GAC[3,4]. Patients with PC are highly symptomatic and can have an overall survival of <6 months. Only a small fraction of patients benefits, often only transiently, from immune checkpoint blockade[5,6] or HER2-directed therapies[7]. Molecular understanding of advanced GAC is limited. Four genotypes defined by The Cancer Genome Atlas (TCGA) were based on analysis of primary GACs[8]. The two clinically favorable subtypes, Epstein–Barr virus-positive and microsatellite instable GACs, are rare in advanced cases[9]. In the clinic, empiricism prevails as patients are not routinely stratified and rational therapeutics are exceedingly limited.

It is well recognized that GAC is endowed with extensive intertumoral heterogeneity and ITH[8,9]. ITH is fundamental for tumor cell survival as it confers therapy resistance and is a major obstacle to improving patient outcomes. However, the origins of ITH are poorly understood. Deeper understanding of the cellular/molecular basis of ITH could influence

how GACs are treated. Single-cell transcriptome sequencing (scRNA-seq) is a robust and unbiased tool to assess cellular and transcriptomic ITH[10].

Here, we performed scRNA-seq of PC cells from ten long-term and ten short-term survivors with PC, inferred tumor cell lineages and transcriptomic states at single-cell resolution by mapping the scRNA-seq data to the Human Cell Landscape (HCL) database[11], constructed a single-cell map of malignant PC cells, comprehensively characterized ITH of PC tumor cells via integrative approaches and identified significant correlates with patient survival. This study demonstrated that the diversity in tumor cell lineage/state compositions is a key contributor to ITH. A 12-gene fundamental signature was discovered, which although derived from PC cells, retained its prognostic significance when applied to independent, localized and advanced large-scale GAC cohorts. These results provide an avenue for patient stratification and target discovery for future therapeutic exploitation.

Results

A single-cell transcriptome map of PC.

scRNA-seq was performed on cryopreserved PC cells collected from 20 patients with GAC at advanced stages, including ten long-term and ten short-term survivors (Fig. 1a). The clinical and histopathological characteristics and radiology images are summarized in the Supplementary Information (Supplementary Table 1 and Supplementary Fig. 1). All primary tumors were of diffuse type and no microsatellite instability (MSI) was observed (see the Methods). Following quality filtering, 45,048 cells were retained for the subsequent analyses. The batch effects were minimal as statistically assessed by k-BET[12] (Supplementary Fig. 2). We captured five main nonmalignant cell types: B cells, CD4 T cells, CD8 T cells, myeloid cells and fibroblasts, each defined by canonical marker genes (Fig. 1a, Extended Data Fig. 1 and Supplementary Table 2). A multistep approach was then applied to distinguish malignant PC cells and to define cell transcriptome states (see the Methods). The immune cells from different patients clustered together by cell type, whereas PC malignant cells clustered distinctly by patient (Extended Data Fig. 1a). It is evident from Extended Data Fig. 2 that tumor cell clusters from short-term survivors were relatively closer on both the uniform manifold approximation and projection (UMAP) plot and the cluster dendrogram. Consistently, the Bhattacharyya pairwise distance between clusters of long-term and short-term survivors was significantly larger than that of the background distributions, indicating distinct transcriptomic profiles associated with survival. In this study, we focused on PC tumor cells (n = 31,131). Five patients with too few tumor cells (<50) were excluded from subsequent analyses (Supplementary Tables 3 and 4). To profile the transcriptome landscape of PC tumor cells, unsupervised cell-clustering analysis was carried out and the results were illustrated using both t-distributed stochastic neighbor embedding (t-SNE) and UMAP[13], which uncovered 14 unique cell clusters, with differentially expressed genes (DEGs) specifically marking each cell cluster (Fig. 1b and Supplementary Fig. 3). Seven of 14 clusters were defined by complete patient specificity, and for the remaining seven clusters, each one was dominated mainly by cells from an individual patient (Supplementary Fig. 4). The distribution of each patient's tumor cells

across clusters was quantified using chi-squared tests (Supplementary Table 5), demonstrating a high degree of interpatient heterogeneity in PC tumor cells.

The inferred tumor cell lineages.

To map each individual tumor cell, to determine its transcriptome state and the potential cells of origin of PC, we used HCL[11], a valuable and well-annotated scRNA-seq resource for human biology, as a reference. Using a prelabeled public scRNA-seq dataset[14], we first tested the reliability of HCL and also evaluated the performance of our approach in cell lineage inference and obtained satisfactory results (Supplementary Fig. 5). We then analyzed PC cells using the same approach (see the Methods). Intriguingly, although all cases in this study were clinically diagnosed as PC from GAC, our transcriptome-based analysis revealed a high degree of cellular heterogeneity in inferred tumor cell lineages (Fig. 1c and Supplementary Table 3). Only ~70% of mapped PC tumor cells were defined as cells of stomach origin, including pit cells (47%), mucosal cells (22%) and chief cells (0.5%). However, the expression profiles of a subset of PC tumor cells (26%) transcriptomically resembled cells of other gastrointestinal (GI) organs, particularly the intestine (21%). It is unlikely that these cells represent cell doublets (Supplementary Fig. 6). DEG analysis revealed lineage-specific gene expression features across major cell lineages including colorectal-like, duodenal-like and gastric cells, and between colorectal enterocytes and goblet cells (Fig. 1d,e), which is supported by compelling evidence from the literature and public databases such as the Human Protein Atlas (https://www.proteinatlas.org; Supplementary Figs. 7 and 8).

We showed that the difference in transcriptomic features was unlikely due to dropouts or technical noise of the scRNA-seq data, as we observed a good correlation in cell lineage assignment between mapping with the raw and imputed data (Supplementary Fig. 9). In addition, we redid HCL mapping after regressing out cell-cycle-related genes and our analysis demonstrated that the cell lineage assignment was not confounded by differences in cell-cycle states (Extended Data Fig. 3). For the two cases (IP-158 and IP-010) with mixed gastric and colonic epithelial cells including colonic goblet cells, we were able to retrieve the histology images of their corresponding primary GACs and confirmed that tumors arose in the setting of gastric intestinal metaplasia, which is characterized by the presence of well-formed goblet cells in gastric mucosa (Fig. 1f). This finding is intriguing given the associated analyses showing mixed cellular populations of both gastric and colonic lineages. For case IP-070, our analysis suggested that none of the PC tumor cells was of GI origin; instead, the cells transcriptomically resembled breast luminal epithelial cells (Fig. 1c and Supplementary Fig. 10). After re-reviewing the patient's clinical records, we noted that this case was of breast cancer that metastasized to the stomach and formed PC. This vignette on the other hand reflected the accuracy of our cell lineage analysis.

The diversity in tumor cell lineage compositions links to ITH at transcriptomic, genotypic and molecular levels.

To further study transcriptomic ITH and examine its relationship with tumor cell lineage compositions, we performed unsupervised clustering analysis of PC tumor cells using Seurat[15] and Monocle[16] and colored cells on the global UMAP plots by their inferred cell

lineages (Fig. 1g,h). As expected, we observed that gastric cells were clustered distinctly from the colorectal-like cells, and this was more evident on the UMAP plots generated from subclustering analysis by patient (Fig. 1h, Extended Data Fig. 4 and Supplementary Table 6). The cluster that was highly enriched with colorectal-like cells was separated from the rest of the clusters that were mainly composed of cells of stomach origin. For case IP-009, the stomach pit cells clustered distinctly from stomach mucosal cells (Fig. 1h). We also performed unsupervised clustering analysis of tumor cells using single-cell consensus clustering (SC3)[17] at both cohort and patient levels. In line with the results of Seurat and Monocle, the independent SC3 approach grouped cells into clusters that exhibited significant differences in the compositions of different cell lineages (Extended Data Fig. 5). For example, tumor cells of IP-067 were grouped into four clusters by SC3. Cells within the cluster C1 were mainly of the stomach origin (96.5%), and only 3.5% of cells were colorectal-like, whereas the cluster C4 was mainly composed of cells of the intestine origin, with 71.4% cells being colorectal-like (two-sided proportion test, C4 versus C1, $P < 2.2 \times 10^{-16}$). Further examining DEGs between the two clusters showed that cells within C4 expressed the highest levels of marker genes of intestine origin such as *DMBT1*, *FCGBP*, *PIGR* and *WFDC2* (Fig. 1e), whereas cells within C1 had the highest expression of marker genes of the stomach origin such as *PSCA* and *TFF1*. Together, our results demonstrate that the diversity in tumor cell lineage compositions is likely a contributor to the transcriptomic ITH.

In addition, we used the Bhattacharyya distance metric to measure the similarity of gene expression distributions between inferred cell lineages. The Bhattacharyya pairwise distances between colorectal-like cells and stomach pit (or mucosal) cells were significantly larger than the distances between pairs of cells of the same lineages (Extended Data Fig. 6), indicating that the colorectal-like cells were transcriptomically distinct from cells of the stomach origin. We next quantified the extent to which lineage diversity explains variations within tumors. On average, lineage difference explains 21%, 20% and 7% of variations, respectively, in the top three PCs (principal components) of a tumor (Supplementary Fig. 11), and, overall, we observed larger distances between lineages across than within patients (Supplementary Fig. 12). Moreover, we sought to identify unsupervised factors that can explain the variances across tumors and performed SC3 unsupervised clustering of tumor cells from all patients, but detected no significant association between the clinical, histopathological or molecular variables and the SC3-defined cell clusters. We further performed pathway enrichment analyses to examine whether differences in certain molecular processes can partially explain the interpatient variances. The cells of several cases tended to cluster together and demonstrated elevated activity of the metabolic and oncogenic pathways (Supplementary Figs. 13 and 14), indicating that tumor-intrinsic signaling pathways may have contributed to the observed interpatient transcriptomic heterogeneity.

We next investigated the genotypic ITH of PC tumor cells and examined its association with inferred tumor cell lineages. Large-scale copy number variations (CNVs) were inferred from scRNA-seq as previously described[10,18,19], followed by phylogenetic reconstruction analysis (see the Methods). In a subset of patients ($n = 6$) whose genomic DNAs were available, the inferred large-scale CNVs showed an overall good correlation with the CNVs called from

bulk whole-exome sequencing (WES) data, as exemplified in Supplementary Figs. 15 and 16. We performed unsupervised clustering of inferred CNVs at both levels, by cell lineage and by patient, and observed greater intertumoral heterogeneity than ITH in the inferred large-scale CNV profiles (Supplementary Figs. 17 and 18). This observation was also consistent with the large F statistic and significant P values from one-way analysis of variance (ANOVA) of the expression profiles (the mean F statistic for the top ten principal components was 7,728.4, $P < 2.2 \times 10^{-16}$), and the larger Bhattacharyya pairwise distances between lineages across than within patients (Supplementary Fig. 12).

Some PC specimens exhibited a high level of ITH. A representative example was the case IP-067 (Fig. 2a). Phylogenetic reconstruction analysis of the inferred CNVs identified five subpopulations (B1–5) with distinct CNV profiles. The pattern of CNV subclonal structure aligned well with the inferred tumor cell lineages: the largest subpopulation, B5, that demonstrated colorectal-like transcriptomic profiles showed distinguished CNVs profiles at multiple chromosomes from the subpopulation B2 that was purely composed of cells of stomach lineage. Consistently, B5 cells were enriched in the Monocle cell cluster C1, whereas B2 cells were mainly in the cluster C3, a cluster that was clearly separated from C1 on the Monocle UMAP plot (Fig. 2a, top right). Cells from clusters B1, B3 and B4 showing shared CNV profiles with that of B2 or B5 were mainly enriched in the Monocle cell cluster C2 that links clusters C1 and C3. To further understand the ITH in this case, we performed somatic variant analysis focusing on the 3′ untranslated region (UTR) using scRNA-seq data (see Methods and Extended Data Fig. 7). Mutation overlapping analysis revealed ITH at the genomic level: 36% of somatic mutations identified in cells of C1 was not detected in cells of C2 or C3; overall, only 26% of mutations were shared among all three clusters (Fig. 2a, bottom right). However, for case IP-009 (Fig. 2b), although the Monocle cell cluster C1 that was mainly composed of stomach pit cells was separated from the cluster C2 that was mainly composed of stomach mucosal cells, we observed slight differences in their CNV profiles among the three subpopulations defined by phylogenetic reconstruction analysis. Somatic variant analysis showed that ~60% of mutations were shared between the two Monocle cell clusters (Fig. 2b, bottom right), suggesting a relatively more similar genomic background.

We next examined ITH in the tumor cell proliferative property and its relationship with the inferred tumor cell lineages. We computationally assigned a cell-cycle state to each individual cell based on its expression profile of cell-cycle-related signature genes[20] (see the Methods and Supplementary Table 7) and compared the tumor cell proliferative property across the inferred cell lineages (Fig. 2c–e and Supplementary Fig. 19). Our analysis showed that the proliferative property of stomach pit cells was the highest among all cell lineages, indicated by high G2M and S scores (Fig. 2c) and fraction of cycling cells (Fig. 2d). On average, about 72% of stomach pit cells were cycling, and the fraction was much higher than that of stomach mucosa cells (27%) or intestinal-like cells: colon goblet cells (58%), colon enterocytes (34%), rectum (15%) and duodenum (18%). These results are in line with previous observations showing that the cellular turnover of stomach pit cells is faster than that of the mucus-secreting cells of the stomach[21,22], and with numerous reports showing that intestinal surface epithelium, including goblet cells, undergoes turnover rapidly[23,24].

Consistently, some key cell-cycle regulatory genes were differentially expressed across tumor cell lineages (Fig. 2e).

17q copy number gain is nearly exclusive to cells of stomach origin and is associated with worse survival.

We analyzed the inferred CNVs from all cases together and discovered 17q copy number gain as a unique event that was prevalent in tumor cells of stomach origin and only present in tumor cells from the short-term survivors (Fig. 3a). By integrating genotypic and transcriptomic profiles, we identified a list of genes upregulated on 17q in tumor cells from cases with evident 17q gain (versus cases without) and associated with patient survival (Fig. 3b,c and Supplementary Table 8). Some of these upregulated genes involved in key signaling pathways (for example, PI3K/AKT/mTOR, mTORC1, MYC) are also potential therapeutic targets (for example, *HN1*, *GRB2*, *PSMB3*), with a number of compounds being screened as active[25]. However, as the CNV profiles were inferred from scRNA-seq data the analysis was limited due to the low resolution. We performed validation analysis in two independent GAC cohorts: the TCGA primary GAC cohort ($n = 411$; Fig. 3d) and a cohort of metastatic GAC ($n = 45$; Fig. 3e). The data from both validation cohorts showed that patients with 17q gain in their tumors had significantly worse survival.

Cell signaling heterogeneity correlates with tumor cell lineages/states.

To examine the molecular consequences of transcriptomic and genotypic alterations described above and to better understand the biological programs associated with patient survival, we performed pathway enrichment analysis of >900 curated gene sets (see the Methods). Among them, 80 pathways were differentially expressed across the inferred tumor cell lineages (Fig. 4a), and, of these, 37 were also strongly associated with patient survival (Fig. 4b and Supplementary Fig. 20). These pathways were categorized into five major classes based on their biological functions: oncogenic signaling, cell cycle, DNA repair, metabolism and immune signaling. Pathway interaction analysis revealed that these biological processes are functionally connected (Fig. 4c).

Pathways that were significantly enriched in tumor cells of stomach origin and associated with shorter survival included cell cycle, DNA repair, PI3K/AKT/mTOR, mTORC1, Wnt, NF-κB and metabolic reprogramming, which are predominantly oncogenic. The pathways that were enriched in colorectal-like tumor cells and associated with longer survival included defensins, IL-7 signaling, complement cascade, IL6/JAK/STAT3 signaling and interferon alpha/gamma, which are all immune related (Fig. 4a,b). These results indicated that different biological processes might have been implicated in tumor cells with different lineages or transcriptome states, contributing to their distinct molecular consequences and patient survival.

To assess whether the cellular composition of the tumor immune microenvironment differed between tumors with gastric-dominant features and those with GI-mixed features, we performed immune deconvolution analysis of the bulk expression data using public datasets (see the Methods). Our results in Fig. 4d and Supplementary Fig. 21 show that the abundance scores of B cells increased significantly in tumors with GI-mixed features (versus

those with gastric-dominant features), and this observation was replicated in three independent cohorts. In addition, the fractions of M1-like macrophages (pro-inflammatory) were higher, and M2-like macrophages (anti-inflammatory) were lower, in tumors with GI-mixed features. There was also a significant difference in the abundance scores of cancer-associated fibroblasts, which were lower in tumors with GI-mixed features. Together, our analysis suggests that tumors with GI-mixed features are immunologically more active.

Single-cell analysis of tumor cell lineage compositions classified PC cases into two subtypes with significant survival difference.

Based on tumor cell lineage compositions, we classified PC samples into two main subtypes: gastric-dominant (mainly gastric cell lineages) and GI-mixed (with mixed gastric and colorectal-like cells) (Fig. 1c), and performed correlation analysis with the clinical/histopathological variables and patient survival (Extended Data Fig. 8 and Supplementary Fig. 22). No significant difference was observed in the histopathological features between these two subtypes. Notably, 17q gain was highly enriched in the gastric-dominant group in the GAC-PC (PC specimens from patients with metastatic GAC) validation cohort (Supplementary Fig. 23); the cell-of-origin-based classification of PC showed a strong correlation with patient survival (Fig. 1c): all six cases with a GI-mixed phenotype were long-term survivors, whereas six of eight cases with a gastric-dominant phenotype were short-term survivors (Fisher's exact test, $P = 0.0097$, log-rank $P = 0.05$; Fig. 5a). Currently, a validated and practical molecular signature for PC is lacking. These results suggest that the transcriptomic features of PC tumor cells could prognosticate patient survival.

Generation and validation of a 12-gene prognostic signature.

We next sought to generate a gene expression signature that could be practical. We performed single-cell DEG analysis on PC tumor cells between the gastric-dominant and GI-mixed subtypes, followed by filtering the DEGs list to identify the most significant DEGs, screening each of the DEGs based on their statistical correlation with patient survival and testing gene combination using a forward selection method (Fig. 5b and see the Methods). After a multistep process, a 12-gene signature was derived (Fig. 5c).

We then validated this signature in an independent GAC-PC cohort ($n = 45$). For each tumor sample, a signature score was computed using bulk RNA-sequencing (RNA-seq) data, and based on which the sample was categorized into either the gastric-dominant or the GI-mixed group for subsequent analysis (Supplementary Fig. 24 and see the Methods). The signature demonstrated an excellent power to prognosticate patient survival, and consistently, patients whose PCs were in the gastric-dominant group survived significantly shorter (7.8 versus 24.5 months) than those whose PCs were in the GI-mixed group (Fig. 5d, left). Multivariate Cox regression analysis showed that this signature was a strong prognosticator of short survival, with a hazard ratio of 12.7 (95% confidence interval, 3.2–51.0, $P = 3.3 \times 10^{-4}$; Fig. 5d, right), and it was independent of clinical/histopathological variables (Supplementary Fig. 22).

We also evaluated its prognostic significance in four other large-scale localized GAC cohorts[19,26–28], totaling 1,336 patients. Notably, although this signature was derived from an

advanced GAC cohort, it retained its prognostic significance in all four validation cohorts of localized GACs (Fig. 5e–h and Extended Data Fig. 8a). Intriguingly, this signature was independent of other molecular and clinical subtypes (Extended Data Fig. 8b). The multivariate Cox proportional-hazards model analysis revealed that its prognostic value was preserved after accounting for the previously defined molecular subtypes, including MSI and EMT (epithelial–mesenchymal transition) signatures by Cristescu et al.[28]; metabolic, proliferative and invasive signatures by Ooi et al.[27]; and other clinical and histopathological variables including age, sex and histology types (Extended Data Figs. 9 and 10). In addition, it correlated strongly with the risk of local recurrence/distant metastasis among the TCGA[8] and Cristescu cohorts[28], where both the expression and outcome data were available (Fig. 5e, Extended Data Fig. 8c and Supplementary Fig. 25). These results further highlighted the value of this prognostic signature and its robustness in prognosticating patient survival.

Discussion

The progress against GAC has lagged behind other GI tumor types. Therapy resistance and the lack of rational therapeutic targets represent the major obstacles in improving survival of patients with advanced GAC[29]. It is widely appreciated that ITH is a fundamental property of cancer contributing to therapeutic failure, development of distant metastases[30] and hindrance to biomarker/target discoveries[31]. Studies of localized and advanced GACs identified multiple molecular subtypes and revealed a high degree of ITH, which is associated with poor clinical outcomes[9,32,33]. Therefore, deeper dissection of ITH is critical for understanding the mechanisms driving poor prognosis of GAC and for overcoming therapeutic resistance. In this study, we dissected, at single-cell resolution, the cellular and transcriptomic ITH of PC tumor cells using the cutting-edge scRNA-seq technology, in combination with integrative computational analyses.

A key finding of this study is that the diversity in tumor cell lineage/state compositions appears to mirror and may even dictate the inherent ITH of PC tumor cells at multiple levels. The origins of ITH have been a subject of discussion, with multiple models being proposed[34,35]. The peritoneal cavity is a unique microenvironment where tumor cells can be in suspension in the peritoneal fluid as opposed to being localized in solid tumor tissues, and thus the PC cells we have sequenced may be a better representation of ITH. We observed that more than one transcriptomically distinct tumor cell subpopulation co-existed in most of the PC cases analyzed and could be distinguished by the inferred cell lineage characteristics. We discovered that 6 of 14 (43%) cases in our discovery cohort had a considerable fraction (~26%) of PC tumor cells that transcriptomically resembled cells of nonstomach GI lineages, particular the intestine. Notably, ITH defined by single-cell lineage/state compositions is perpetuated at transcriptomic, genotypic, cell-cycle state, molecular signaling and phenotypic levels and strongly associated with patient survival. Tumor cell transcriptomic profiles and proliferative property also significantly differed across the inferred tumor cell lineages/states, as did the molecular signaling, suggesting that treatment strategies could potentially be tailored to these molecular features. It appears that the contributors to ITH are likely diverse and more complicated than original thought, and varied biological programs (for example, genomic/epigenomic) might have been engaged early in tumorigenesis. It is important to note that the different tumor cell lineages were

inferred by mapping scRNA-seq data to HCL and thus may only reflect changes at transcriptome level, instead of developmental lineages. Although we are unable to discern the precise cells of origin of each PC, we believe that the insights shared by this report will stimulate further studies in the field focusing on tumor cell of origin and lineage diversity/plasticity analyses of both gastric and other cancer types, to better elucidate the regulatory mechanisms, possible effects on tumor progression and therapy responses.

It is noteworthy that we discovered 17q gain as a nearly exclusive event associated with PC cells of gastric lineage. The 17q region harbors multiple potential therapeutic targets and, interestingly, all patients whose tumors had 17q gain in our discovery cohort were short-term survivors, and the association of 17q gain and inferior survival was validated in both the localized and advanced GAC cohorts. Our discovery of the intimate link between tumor cell lineage compositions and genomic ITH at single-cell resolution could be generalized to other cancer types and broaden our understanding of cancer biology in general.

High genomic ITH in most cancers is associated with worse survival. We observed an opposite phenomenon in this study: patients with the GI-mixed molecular features in their PC tumor cells survived significantly longer than those with the gastric-dominant features. We remain uncertain of the detailed mechanisms. A possible explanation is that the intestinal-like cells in the GI-mixed tumors could be acquired from a process called intestinal metaplasia, which is the main precancerous lesion of the stomach. Intestinal metaplasia is characterized by the presence of differentiated epithelium that resembles the small intestine (partial or complete transformation of gastric gland epithelial cells into the intestinal type) on the basis of ultrastructural morphology, mucin patterns and enzyme histochemistry[36]. Consistent with this, two patients (IP-010 and IP-158) in this study had intestinal metaplasia confirmed in their primary tumors (diffuse type), although intestinal metaplasia is thought to be mainly associated with GAC of the intestinal type. In line with this, it is generally recognized that patients with intestinal-type GAC have better prognoses than those with a poorly differentiated or diffuse type histology[37–39]. In addition, immune deconvolution analysis using public datasets suggested that the better clinical outcomes of GI-mixed tumors could be partially associated with a more engaged and effective immune response against the tumor, including higher levels of B cells (which are known to be associated with a protective immunity and better clinical outcomes)[40–42] and M1 polarization[43,44], lower levels of fibroblasts[45–47] and M2-like macrophages, and elevated cytolytic activity. Nevertheless, further investigations are needed to elucidate the underlying mechanisms.

Most intriguingly, based on tumor cell lineage/state compositions, PC cases were classified into two cellular subtypes that were prognostic independent of histopathological features. Further analyses led us to discover a 12-gene signature that appears to be fundamental to GAC carcinogenesis/propagation as it was not only highly prognostic in GAC-PC validation cohort but performed just as robustly in several large-scale localized GAC cohorts. Currently, to our knowledge there is no such signature in clinical use, and thus it has a high potential to stratify patients for more effective therapies as they become available.

Methods

Patient cohort, clinical characteristics and sample collection.

A total of 20 patients with GAC with malignant ascites (PC) were included in this study. Based on the Lauren classification, all of the primary GACs were of diffuse type. The primary GAC diagnosis was confirmed through an endoscopic biopsy. Pathology results were verified independently by two experienced GI pathologists. In addition, computed tomography images of all patients were re-reviewed by an experienced imaging physician. Our reviews of pathology and radiology results confirmed that all cases except IP-070 (breast cancer metastatic to stomach determined by pathology and profiling) represented primary gastric cancer. None of the patients had a history or documented diagnosis of primary colon cancer, or imaging findings of colon cancer. The detailed clinical and histopathological characteristics are described in Supplementary Table 1, and representative computed tomography images are shown in Supplementary Fig. 1. GACs were staged according to the American Joint Committee on Cancer Staging Manual (8th edition)[48,49]. PC was confirmed by cytologic examination. This cohort included ten long-term survivors and ten short-term survivors. The long-term survivors were patients who survived more than 1 yr after the diagnosis of PC and the short-term survivors were patients who died within 6 months after the diagnosis of PC. Of the 20 patients, 16 had signet-ring cell carcinoma. Her2 staining was performed and all tumors were Her2 negative. PC specimens were collected at The University of Texas MD Anderson Cancer Center (Houston, USA) under an Institutional Review Board-approved protocol (no. LAB01–543) after obtaining written, informed consent from each participant. Patients with diagnosed GAC-PC with ascites were approached when they required a therapeutic paracentesis. No other selection criteria were applied. This project was in accordance with the policy advanced by the Helsinki Declaration of 1964 and later versions. PC specimens were spun down for 20 min at 2,000*g* and pelleted cells (PC cells) were isolated, cryopreserved at −80 °C and used for scRNA-seq. All samples were processed using the same protocol and by the same research assistant.

MSI testing.—The MSI test was not routinely done in clinic for gastric cancer when these patients were enrolled. For this study, only three patients were assessed for MSI test and reported as ‘microsatellite (MS) stable’. In six patients (all expired) there were no residual tissues for additional testing. For the remaining 13 patients, WES was performed on the same ascites samples or the ascites samples collected at a similar timepoint as the sample used for scRNA-seq, and was thus used for an unbiased genomic analysis of microsatellites. WES data were processed and mapped to human reference genome as previously described[9]. MSIsensor[50], a validated algorithm for deriving MSI status from genomic sequencing data, was applied to the aligned BAM files for detection of somatic microsatellite changes. The MSIsensor score was below the suggested threshold (score > 10 for matched tumor–normal pair, score > 20 for tumor-only sample)[50] and all 13 patients were designated as MS stable.

scRNA-seq library preparation and sequencing.

Chromium single-cell sequencing technology from 10x Genomics was used to perform single-cell separation, complementary DNA amplification and library construction following the manufacturer’s guidelines. Briefly, the cellular suspensions were loaded on a 10x

Chromium Single Cell Controller to generate single-cell gel bead-in-emulsions. The scRNA-seq libraries were constructed using the Chromium Single Cell 3′ Library & Gel Bead Kit v.2 (PN-120237, 10x Genomics). The HS dsDNA Qubit Kit was used to determine the concentrations of both the cDNA and the libraries. The HS DNA Bioanalyzer was used for quality-tracking purposes and size determination for cDNA and lower-concentrated libraries. Sample libraries were normalized to 7.5 nM and equal volumes were added of each library for pooling. The concentration of the library pool was determined using the Library Quantification qPCR Kit (KAPA Biosystems) before sequencing. The barcoded library at the concentration of 275 pM was sequenced on the NovaSeq6000 (Illumina) S2 flow cell (100 cycle kit) using a 26 × 91 run format with 8 bp index (read 1). To minimize batch effects, the libraries were constructed using the same versions of reagent kits and following the same protocols, and the libraries were sequenced on the same NovaSeq6000 flow cell and analyzed together.

scRNA-seq data processing and analysis.

Raw sequencing data processing, quality check, data filtering, doublets removal, batch-effect evaluation and data normalization.—The raw scRNA-seq data were preprocessed (demultiplex cellular barcodes, read alignment and generation of gene count matrix) using the Cell Ranger Single Cell Software Suite provided by 10x Genomics. Detailed quality-control metrics were generated and evaluated. Genes detected in fewer than three cells and cells with low-complexity libraries (in which detected transcripts were aligned to less than 200 genes) were filtered out and excluded from subsequent analysis. Low-quality cells where >15% of transcripts derived from the mitochondria were considered apoptotic and also excluded. Following the initial clustering, we removed likely cell doublets from all clusters. Doublets were identified by the following methods: (1) Library complexity: cells are outliers in terms of library complexity. Cells in the top 1% of the distribution of genes detected per cell were removed. (2) Cluster distribution: doublets or multiplets likely form distinct clusters with hybrid expression features and exhibit an aberrantly high gene count. (3) Cluster marker gene expression: cells of a cluster express markers from distinct lineages (for example, cells in the T cell cluster showed expression of epithelial cell markers; cells in the B cell cluster showed expression of myeloid cell markers). We carefully reviewed canonical marker gene expression on UMAP plots and repeated the steps above a couple of times to ensure that we had filtered out most of the barcodes associated with cell doublets.

Following removal of the poor-quality cells and doublets, a total of 45,048 cells were retained for downstream analyses. Library size normalization was performed in Seurat[15] on the filtered gene–cell matrix to obtain the normalized UMI (unique molecular identifier) count data as previously described[51]. Statistical assessment of possible batch effects was performed using the R package k-BET (a robust and sensitive *k*-nearest neighbor batch-effect test)[12]. k-BET was run on major immune cell types including B, myeloid, CD4 and CD8 T cells separately with default parameters. A control dataset with known significant batch effects was included to assist with data interpretation. We chose the *k* input value from 1% to 100% of the sample size. In each run, the number of tested neighborhoods was 10% of the sample size. The mean and maximal rejection rates were then calculated based on a

total of 100 repeated k-BET runs. A low rejection rate indicates homogeneous mixing of samples from different batches. k-BET results suggested minimal batch effects in this dataset (Supplementary Fig. 2).

Unsupervised cell clustering, dimensionality reduction and cluster relationship analysis.—Seurat[15] was applied to the normalized gene–cell matrix to identify highly variable genes for unsupervised cell clustering. To identify highly variable genes, the *MeanVarPlot* method in the Seurat[15] package was used to establish the mean–variance relationship of the normalized counts of each gene across cells. We then chose genes whose log-mean was between 0.0125 and 3 and whose dispersion was above 0.5, resulting in 3,018 highly variable genes. The elbow plot was generated with the *PCElbowPlot* function of Seurat[15], and based on which the numbers of significant principal components (PCs) were determined. Different resolution parameters for unsupervised clustering were then examined to determine the optimal number of clusters. For this study, the first ten principal components and the highly variable genes identified by Seurat[15] were used for unsupervised clustering with a resolution set to 0.6, yielding a total of 20 cell clusters (Extended Data Fig. 1a). For visualization, the dimensionality was further reduced using either the *t*-SNE or UMAP[13] methods with *Seurat* functions *RunTSNE* and *RunUMAP*, respectively. The principal components used to calculate the embedding were the same as those used for clustering.

In addition, *Monocle 3* alpha (http://cole-trapnell-lab.github.io/monocle-release/monocle3/)[16] was applied as an independent tool for unsupervised clustering analysis (function *cluster_cells*) focusing on tumor cells, and UMAP was used by default with the *Monocle* functions *reduce_dimension* and *plot_cells* for dimensionality reduction and visualization of the *Monocle* clustering results. *Monocle 3* alpha was also used to construct the single-cell trajectories. The function *learn_graph* was run with default parameters. Moreover, we applied an additional unsupervised clustering approach, the single-cell consensus clustering (SC3) analysis[17], on tumor cells from all patients and on tumor cells from each individual PC specimen. SC3 was run with default parameters and independent of cell lineage annotation. Furthermore, to study the hierarchical relationships among tumor cell clusters, we performed unsupervised cluster analysis. The dendrogram was drawn using Pearson correlation coefficient (PCC) with average principal component analysis (PCA) space (*Seurat* function *RunPCA*) for each tumor cell cluster with the R package denextend[52].

Sample distribution analysis.—To quantify the distribution of each patient's tumor cells across Seurat-defined cell clusters, we used the ratio of the observed to expected cell numbers in clusters to measure the enrichment of cells within a sample (tumor) across different cell clusters as previously described[53]. Given a contingency table of samples by clusters, we first applied the chi-squared tests to evaluate whether the distribution of cells of a sample across clusters significantly deviates from random expectations. We then calculated the $R_{o/e}$ for each combination of samples and clusters as follows:

$$R_{o/e} = \frac{\text{observed}}{\text{expected}}$$

where $R_{o/e}$ is the ratio of the observed cell number to the expected cell number of a given combination of cluster and sample. The expected cell numbers for each combination of clusters and samples were obtained from the chi-squared test. Different from the χ^2 values, which are defined as $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ and could only indicate the divergence of observations from random expectations, $R_{o/e}$ could indicate whether cells of a certain sample are enriched in a specific cluster. For example, if $R_{o/e} > 1$, it suggests that cells of the sample are more frequently observed than random expectations in a specific cluster; that is, enriched. If $R_{o/e} < 1$, it suggests that cells of a given sample are observed with less frequency than random expectations in a specific cluster.

Determination of major cell types and cell states.—To define the major cell type of each single cell, DEGs were identified for each cell cluster using the *FindAllMarkers* analysis in the Seurat[15] package and the top 20 most significant DEGs were carefully reviewed. In parallel, feature plots were generated for the top 20 DEGs and a suggested set of canonical immune and stromal cell markers (Supplementary Table 2), a similar approach as previously described[54,55], followed by a manual review process. Enrichment of these markers (for example, *EPCAM* for epithelial cells, *PTPRC* for immune cells, *CD3D/E* for T cells, *CD8A/B* for CD8 T cells, *IL7R/CD4/CD40LG* for CD4 T cells, *CD19/MS4A1/CD79A* for B cells, *COL1A1/COL1A2* for fibroblasts, and so on) in certain clusters was considered a strong indication of the clusters representing the corresponding cell types (Extended Data Fig. 1). The two approaches were combined to infer major cell types for each cell cluster according to the enrichment of marker genes and top-ranked DEGs in each cell cluster, as previously described[55].

Inference of large-scale CNVs, phylogenetic tree construction and correlation analysis.—The tool inferCNV (https://github.com/broadinstitute/inferCNV) was applied to infer the large-scale CNVs from scRNA-seq data, and monocytes from this dataset were used as a control for CNV analysis. Initial CNVs were estimated by sorting the analyzed genes by their chromosomal locations and applying a moving average to the relative expression values, with a sliding window of 100 genes within each chromosome, as previously described[10,19]. In a subset of cases ($n = 6$) whose WES data were available, the true CNVs called from WES were used as the positive control to assess the performance of the inferCNV analysis. Finally, malignant cells were distinguished from normal cells based on the information integrated from multiple sources, including cluster distribution of the cells, marker gene expression, inferred large-scale CNVs and aneuploidy status.

To construct a phylogenetic tree from the CNV calls in each tumor cell of a specific sample, the relative CNVs were calculated using the inferCNV outputs and the average CNV values were computed for nonoverlapping genomic bins, each consisting of 30 genes. For each cell within a bin, we calculated an integer copy number by multiplying relative CNV value by 2 (diploid) and rounding the results off to the closest integers. The R package phangorn was then used to construct the phylogenetic maximal parsimony tree. The integer copy number profiles were re-segmented by the collection of breakpoints detected in each cell, so that each column in the data matrix corresponds to the longest interval uninterrupted by any

variations across the cell population. The breakpoints in individual cells were determined by the R package copynumber[56] under default parameters.

Correlation analysis of the CNVs inferred from scRNA-seq data and that identified using bulk WES.—Copy number analysis using scRNA-seq data is limited due to the low resolution; therefore, only the arm-level events were included. The arm-level copy number ratio was calculated as the weighted average of copy number ratio of genes (scRNA-seq data, inferCNV output) or segments (WES data, the copy number segments) as follows:

$$CN = \frac{\sum CN_i \times L_i}{\sum L_i}$$

where CN_i means the copy number ratio of the ith gene or segment. L_i means the length of the ith gene or segmentation.

The Pearson correlation and Spearman correlation analyses were then conducted on the resultant arm-level copy number ratio.

Cell-of-origin inference.—Origins of tumor cells were inferred by mapping our scRNA-seq data to the well-annotated single-cell database of HCL (http://bis.zju.edu.cn/HCL)[11] using the R package scHCL (https://github.com/ggjlab/scHCL). First, gene expression normalization was performed using the following formula:

$$E = \frac{\text{Count}}{\text{sum(Count)}} \times 10^5$$

where E denotes the normalized gene expression value; ‘Count’ denotes the raw UMI counts; and ‘sum (Count)’ is the sum of all raw UMI counts in one cell. The PCC between the normalized expression profile of each query cell and the expression profile of each annotated cell type from the HCL reference dataset was calculated using the scHCL software package. The PCC was estimated using 6,075 signature genes provided by scHCL[11]. Cell lineage/type was subsequently assigned for each query cell based on the following criteria: PCC > 0.3; and the best-matched stomach-derived cell lineage/type or the best-matched cell lineage/type if there was no stomach-derived cell lineage among the top five hits. Cells that did not have a good correlation coefficient (PCC < 0.3) with any cell lineage/type in the HCL database were classified as ‘other’. To examine the similarity between colorectal-like cells and cell doublets, we applied a doublet simulation approach. Therein, a random sampling of cells (excluding intestinal-like cells) was taken to generate 500 simulated doublets and these cells were used for the UMAP clustering along with colorectal-like cells.

Quantifying the similarity of gene expression distributions of cell lineages and clusters within and across patients.—The Bhattacharyya distance metric (a distance metric that is effective at comparing pairwise probability distributions) was used to measure the similarity of gene expression distributions for all pairs of cell clusters between

the long- and short-term survivors. We embedded cell clusters in two-dimensional space with PCA using the highly variable genes and retained the top 50 principal components for subsequent analysis. We randomly sampled 500 cells from each tumor cell cluster of short-term survivors and long-term survivors, repeated 100 times and computed the Bhattacharyya pairwise distance between clusters as follows (a similar approach as described previously[57,58]):

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln\left(\frac{\det\boldsymbol{\Sigma}}{\sqrt{\det\boldsymbol{\Sigma}_1 \det\boldsymbol{\Sigma}_2}}\right)$$

where μ_1 and μ_2 are the mean vectors of each distribution, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2$.

For comparison, we also generated background distributions of clusters of short-term survivors (Short) by randomly sampling 100 pairs of cells and computing the Bhattacharyya distance between each pair of cells. In addition, we also generated the Bhattacharyya distance between cells randomly sampled independent of survival status (Random). We only evaluated clusters that had 500 or more cells. Similarly, we computed the Bhattacharyya pairwise distance between different cell lineages (inferred by HCL) within and across patients. For the patient-level analysis, we randomly sampled 100 cells from each inferred cell lineage (for example, stomach pit cells, stomach mucosal cells, colorectal-like cells; Fig. 1h) and repeated 100 times, and computed the Bhattacharyya pairwise distance between different lineages as described above. For comparison, we also generated background distributions of each lineage by randomly sampling 100 cells twice of the same lineage and computing the Bhattacharyya distance between each pair of cells, and also generated the Bhattacharyya distance between cells randomly sampled independent of lineage annotation (Random). We only evaluated the major lineages (for example, stomach pit cells, stomach mucosal cells, colorectal-like cells) that had 500 or more cells.

scRNA-seq imputation.—Markov affinity-based graph imputation of cells (MAGIC)[59] is a commonly used algorithm for denoising scRNA-seq data. It learns the manifold data and imputes likely gene expression in each cell by sharing information across similar cells. Here, we followed the concept of MAGIC and simplified the imputation process as follows: (1) Computation of affinity matrix via the FindNeighbors function from Seurat. This step constructs a *k*-nearest neighbor graph based on the Euclidean distance in PCA space and refines the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard similarity). It takes default *k* and the same number of principal components as used for unsupervised clustering. (2) Symmetrization of the affinity matrix using an additive approach. (3) Row-stochastic Markov-normalization of symmetric affinity matrix (so each row sums to 1) into Markov matrix, representing the probability distribution of transitioning from each cell to every other cell. (4) Imputation expression matrix by multiplying the Markov matrix by the original expression matrix.

Quantifying the contribution of cell lineage diversity to transcriptomic variation within tumors.—Given that detection rate is a major source of cell-to-cell variation for scRNA-seq datasets[60], we applied MAGIC[59] to impute likely gene expression

to reduce this unwanted source of variation. After MAGIC ($t = 3$), the top 15 principal components can explain 81% ~ 93% of the total variance in each patient. To quantify the extent to which lineage diversity explains transcriptomic variation within a tumor, we then performed one-way ANOVA tests on each of the top 15 principal components in each tumor, using a similar approach as previously described[58]. The one-way ANOVA tests were conducted using the *aov* function in the R stats package. ANOVA partitions the total variation into within-lineages variation and between-lineages variation. The percentages of variation that can be explained by lineage diversity in each PCA space were then calculated for each tumor.

Comparison of between- and within-patient variations.—To compare the magnitudes of between- and within-patient variations in the transcriptomic profiles, we performed the one-way ANOVA tests on the top 15 principal components for tumor cells from patients that had 1,000 or more cells. The one-way ANOVA tests were conducted using the oneway.test() function in the R stats package. ANOVA partitions the total variation into between- and within-patient variation. In addition, we also performed the one-way ANOVA tests in the CNV profiles for each inferred lineage (for example, stomach pit cells, stomach mucosal cells, colorectal-like cells). The F statistic is the ratio of between-patient variation to within-patient variation, with F statistic >1 indicating that the between-patient variation is greater than the within-patient variation.

Inferring cell-cycle stage, hierarchical clustering, DEGs and pathway enrichment analysis.—The cell-cycle stage was computationally assigned for each individual cell by the function CellCycleScoring that is implemented in Seurat[15]. Cell-cycle stage was inferred based on the expression profile of the cell-cycle-related signature genes, as previously described[20]. Hierarchical clustering was performed at multiple levels (all tumor cells together, by cell lineage and by patient) using the Ward minimum variance method. DEGs were identified for each cluster using the *FindMarkers* function in Seurat R package[15] and DEG list was filtered with the following criteria: the gene should be expressed in 20% or more cells in the more abundant group; expression fold change > 1.5; and false discovery rate (FDR) Q value < 0.05. The heat map was then generated using the pheatmap function in pheatmap R package for filtered DEGs. For pathway analysis, the curated gene sets (including Hallmark, KEGG and Reactome gene sets, $n = 910$) were downloaded from the Molecular Signature Database (MSigDB, http://software.broadinstitute.org/gsea/msigdb/index.jsp), single-sample GSVA (ssGSVA) was applied to the scRNA-seq data and pathway scores were calculated for each cell using the *gsva* function in the GSVA software package[61]. Pathway enrichment analysis was done with the limma R software package. Significantly enriched signaling pathways were identified with an FDR Q value < 0.01.

To profile the interactions between biological pathways, the differentially expressed pathways were assembled into a network where nodes represent pathways and edges represent their interactions (if they have common genes). The weight of an interaction corresponds to its Jaccard index between each pathway pair. The Jaccard index is a measure of set (here it refers to pathway) similarity, and defines two sets (pathway A and pathway B)

as the ratio of the size of their intersection over the size of their union (see the equation below):

$$J(\mathrm{A,B}) = \frac{|\mathrm{A}\cap\mathrm{B}|}{|\mathrm{A}\cup\mathrm{B}|}$$

For each pathway, its entity membership was represented as the set $_{\mathrm{Pi}} = \{e_1, e_2, \ldots e_n\}$. The Jaccard index was computed between a pair of pathways A and B as J(A, B). Cytoscape[62] was then used to visualize the network.

Generation of the 12-gene prognostic signature.—To generate a gene expression signature that is clinically applicable, we performed multistep analysis (Fig. 5a). First, we compared the gene expression profiles of tumor cells between the gastric-dominant and GI-mixed groups and identified DEGs. The DEG list was then filtered based on the following criteria to select the most significant DEGs. Only the genes with an expression fold change >1.5 or <−1.5 and an FDR < 0.01 and highly expressed (normalized UMI count > 1, expressed in at least 50% of cells from one of the two groups) were selected and taken into subsequent analysis. We next screened each of the DEGs based on their statistical correlation with patient survival using a univariate Cox proportional-hazards regression model, and only the DEGs that showed consistency with patient survival were selected (for example, DEGs highly expressed in the gastric-dominant group with a hazard ratio > 1 or DEGs highly expressed in the GI-mixed group with a hazard ratio < 1). We identified 149 significant DEGs in total (that met the above criteria). After that, the Harrell concordance index (C-index)[63] was applied to quantify the predictive accuracy of the prognosis-related DEGs. C-index was calculated using a univariate Cox proportional-hazards regression model and the R package *survcomp*. A C-index value of 0.5 indicates no predictive ability, whereas a value of 1 represents perfect predictive ability, similar to that previously described[64,65]. Finally, we performed a forward selection process to search for a set of DEGs that can achieve the largest C-index value based on the following procedures. We chose the gene that had a C-index > 0.7 and expression fold change >2.5 or <−2.5 as a seed (refers *TM4SF1*). The rest of the DEGs (n = 148) were then added to the signature, one at a time. Each time, we screened all of the rest of the genes one by one, then evaluated the C-index of the potential signature after adding a specific gene, and finally picked the gene that reached the highest C-index to the signature. We repeated this process until the C-index had reached a plateau and did not increase any more. For each DEG added to the signature, we labeled each sample based on the expression level of the corresponding gene using the following equation:

$$\text{For gene g, } L_{\mathrm{g,j}} = \begin{cases} 1 \times I(\mathrm{g}), V_{\mathrm{g},\,j} \le \mathrm{median}\left(V_{\mathrm{g}}\right) \\ -1 \times I(\mathrm{g}), V_{\mathrm{g},\,j} > \mathrm{median}\left(V_{\mathrm{g}}\right) \end{cases}$$

where $I(g)$ = 1 if gene g is highly expressed in samples of the gastric-dominant group; otherwise, $I(g)$ = −1.

L, g, j and V denote label, gene, jth sample and gene expression, respectively.

We then summed the scores for each sample and quantified the predictive performance of each derived signature based on its corresponding C-index value. Among all of the signatures derived from the seed, a 12-gene signature showed the highest C-index and was chosen for subsequent validation analyses. As we started from a relatively small number of carefully filtered DEGs (n = 149), no further statistic was applied to limit the size of the gene set.

Validation of the 12-gene prognostic signature.—The signature was then subject to validation with both internally generated and publicly available datasets. Briefly, the signature score was calculated for each sample using a similar approach as that used by Kang et al.[66]. The workflow is illustrated in Supplementary Fig. 15. First, a sample–gene expression matrix (for 12 signature genes) was extracted from each normalized bulk RNA-seq or expression microarray dataset. Second, for each sample, a score of 1 or −1 was assigned for each of the 12 signature genes based on its relative expression (> or ≤ median value) and whether the signature gene was associated with gastric-dominant or GI-mixed features. Briefly, if the gene was among one of the seven genes that are associated with the gastric-dominant feature and its expression in a sample was less than or equal to the median value, we assigned a score of 1 for this gene for this specific sample, and we assigned a score of −1 if its expression was greater than the median value. If the gene was among one of the five genes that are associated with the GI-mixed feature and its expression in a sample was greater than the median value, we assigned a score of 1 for this gene for this specific sample, and we assigned a score of −1 if its expression was less than or equal to the median value (Supplementary Fig. 15). After that, the scores of each sample were summed, which constituted the signature score. Finally, the samples were categorized into gastric-dominant or GI-mixed groups based on their corresponding signature scores: ≤median or >median, respectively. For the bulk RNA-seq datasets, the signature scores were calculated using the log-transformed FPKM (fragments per kilobase of transcript per million mapped reads) values. For the bulk expression microarray datasets, the signature scores were calculated using the normalized gene expression values.

Immune cell deconvolution.—CIBERSORT[67] was applied to the normalized bulk RNA-seq and microarray gene expression datasets with the LM22 gene signature to estimate the relative fractions of 22 immune cell types. In addition, the R package MCP-counter[68] was applied to infer the abundance of eight immune cell subpopulations including T cells, CD8 T cells, cytotoxic lymphocytes, NK cells, B lineage cells, monocytic lineage cells, myeloid dendritic cells and neutrophils, as well as endothelial cells and fibroblasts.

Single-cell somatic variant analysis.—For each sample, the reads were extracted from the original BAM file using the cell-specific barcodes and were aggregated to generate a sub-BAM file for each Monocle-defined cell cluster. Mutect2 (v.4.1.0.0)[69] was then applied to the sub-BAM files to identify somatic point variants. The Mutect2 outputs were run through our pipeline for filtering and annotation. Briefly, only Mutect2 calls located at the 3′ UTR and marked as ‘PASS’ were selected and taken into the next step. Variants with total read coverage < 30, variant read coverage < 6 or variant allelic fraction < 0.1 were removed. After that, common variants reported by the Phase-3 1000 Genomes Project or ExAC (the

Exome Aggregation Consortium) with minor allele frequency greater than 0.5% were further removed. Additionally, we included a virtual normal panel of 33 germline samples from GAP-PC patients to help remove artifacts related to sequencing and mapping errors as well as common single nucleotide polymorphisms. The events that overlapped with variants called from this virtual normal panel were further excluded. Finally, the remaining somatic variants were carefully reviewed on the Integrative Genomics Viewer and variants with noisy background were further discarded. For variant-overlapping analysis at cluster level, we first made a unique list of variants by aggregating all quality-control-passed variants from the Monocle-defined tumor cell clusters of a sample. Then we queried their corresponding sub-BAM files for each unique variant site by chromosome and coordinates and the numbers of reference and variant alleles were counted, which were subsequently used to identify shared and unique variants among tumor cell clusters.

Public datasets.

In addition to the scRNA-seq dataset generated internally for the discovery GAC-PC cohort, we included the bulk transcriptome sequencing (RNA-seq) data generated on an independent GAC-PC cohort from our recent study[9] to validate the 12-gene prognostic signature. Moreover, we downloaded the normalized bulk RNA-seq data generated by TCGA on primary stomach adenocarcinoma (STAD) from the NCI Cancer Genomic Data Commons (NCI-GDC: https://gdc.cancer.gov). The RNA-seq data were processed and normalized by the NCI-GDC bioinformatics team using their transcriptome analysis pipeline. The clinical annotation of the TCGA STAD cohort was downloaded from a recent PanCanAtlas study[70]. The TCGA STAD cohort (n = 411) included both intestinal (n = 176) and diffuse type (n = 69) tumors.

Furthermore, we downloaded three other large-scale primary GAC datasets (GSE62254 (ref. [28]), GSE15459 (refs. [27,71]), GSE84437 (ref. [72])) from the Gene Expression Omnibus database (GEO, https://www.ncbi.nlm.nih.gov/geo/) to further evaluate the prognostic power of the 12-gene signature. The raw gene expression values from microarray experiments were preprocessed (background-corrected and log2-transformed) and quantile-normalized using the robust multi-array average algorithm[73]. For each sample, the expression measurements of all probes corresponding to the same gene ID were averaged to obtain a single measurement. For datasets GSE62254 and GSE15459, the clinical, histopathological and survival data as well as molecular signatures defined by each study[27,28,71] were downloaded and used for the multivariate Cox regression analysis. These primary GAC datasets included both intestinal and diffuse types of GACs. The dataset GSE62254 (n = 300) included 134 diffuse types, 146 intestinal types and 20 mixed types of GACs. The dataset GSE15459 (n = 192) included 75 diffuse types, 99 intestinal types and 18 mixed types of GACs.

Additionally, to test the reliability of the HCL resource as a reference dataset and to evaluate the performance of our approach in cell lineage inference, we downloaded from the Data Portal of Human Cell Atlas (SCP259, https://data.humancellatlas.org) the scRNA-seq dataset generated on normal human colon tissues using a SMART-Seq2 protocol by a recent study[14]. The same approach as outlined in the section ‘Cell-of-origin inference’ was applied to the SCP259 dataset for cell lineage inference.

Statistical analysis.

In addition to the bioinformatics approaches described above for scRNA-seq data analysis, all other statistical analyses were performed using statistical software R v.3.5.2. Analysis of differences on a continuous variable (for example, gene expression, pathway score) across two groups (a categorical independent variable, such as gastric-dominant versus GI-mixed) was performed by the nonparametric Mann–Whitney *U* test. The nonparametric Kruskal–Wallis test was applied to assess the significant difference on a continuous variable by a categorical independent variable with multiple groups (for example, across different tumor cell lineages/types). For survival analyses, including overall survival (OS), progression-free interval, disease-free survival (DFS), disease-specific survival (DSS), disease-free interval (DFI) and survival time from peritoneal metastasis, we used the log-rank test to calculate *P* values between groups, and the Kaplan–Meier method to plot survival curves. For the TCGA dataset, the clinical annotation and the times calculated for OS, DFS, DSS and DFI were downloaded from the PanCanAtlas study[70]. For other large-scale primary GAC datasets downloaded from GEO, the OS times were downloaded from their corresponding published studies[27,28,71,72]. The hazard ratios were calculated using the multivariate Cox proportional-hazards model. All statistical significance testing in this study was two-sided. To control for multiple hypothesis testing, we applied the Benjamini–Hochberg method to correct *P* values and the FDR *Q* values were calculated. Results were considered statistically significant at *P* value or FDR *Q* value of <0.05.

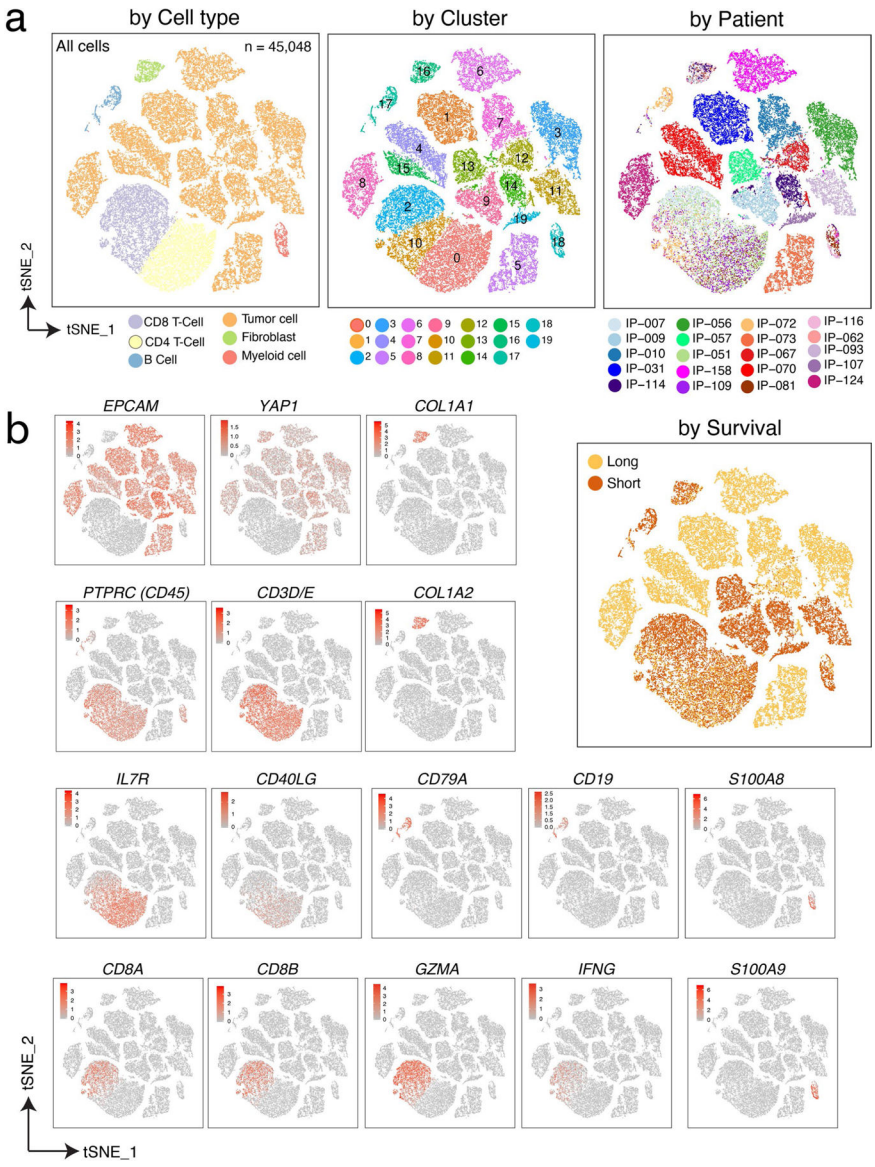
Statistics and reproducibility.

Supplementary Fig. 2b was generated from $n = 100$ repeated k-BET runs. A random sampling of cells ($n = 100$ times) was performed to generate the simulated doublets in Supplementary Fig. 6b, to calculate the Bhattacharyya pairwise distance between tumor cell clusters from samples of long- and short-term survivors in Extended Data Fig. 2c and to calculate the Bhattacharyya distance between and within inferred cell lineages in Extended Data Fig. 6. The statistical methods used for each analysis are described within the figure legends. The key findings of this study were validated by analyzing large-scale public datasets as described above in the section ‘Public datasets’. For the histology image shown in Fig. 1f: because of the nature of clinical care in this disease, only one diagnostic biopsy specimen per patient was available. Obviously, this is representative of the tumor and sufficient for making clinical decisions, but is also used here for analysis that parallels how we practice in the clinic. Keeping with the theme of our report on ITH, it is likely that metaplasia has been overgrown by tumor cells; however, the phenotypic appearance of metaplasia and the classic appearance of goblet cells are highly reliable patterns and true representations of the presence of a premalignant element in a particular GAC. Lack of metaplasia in the primary specimen, however, does not exclude its previous existence (sampling errors or abolition by eventual GAC).

Reporting Summary.

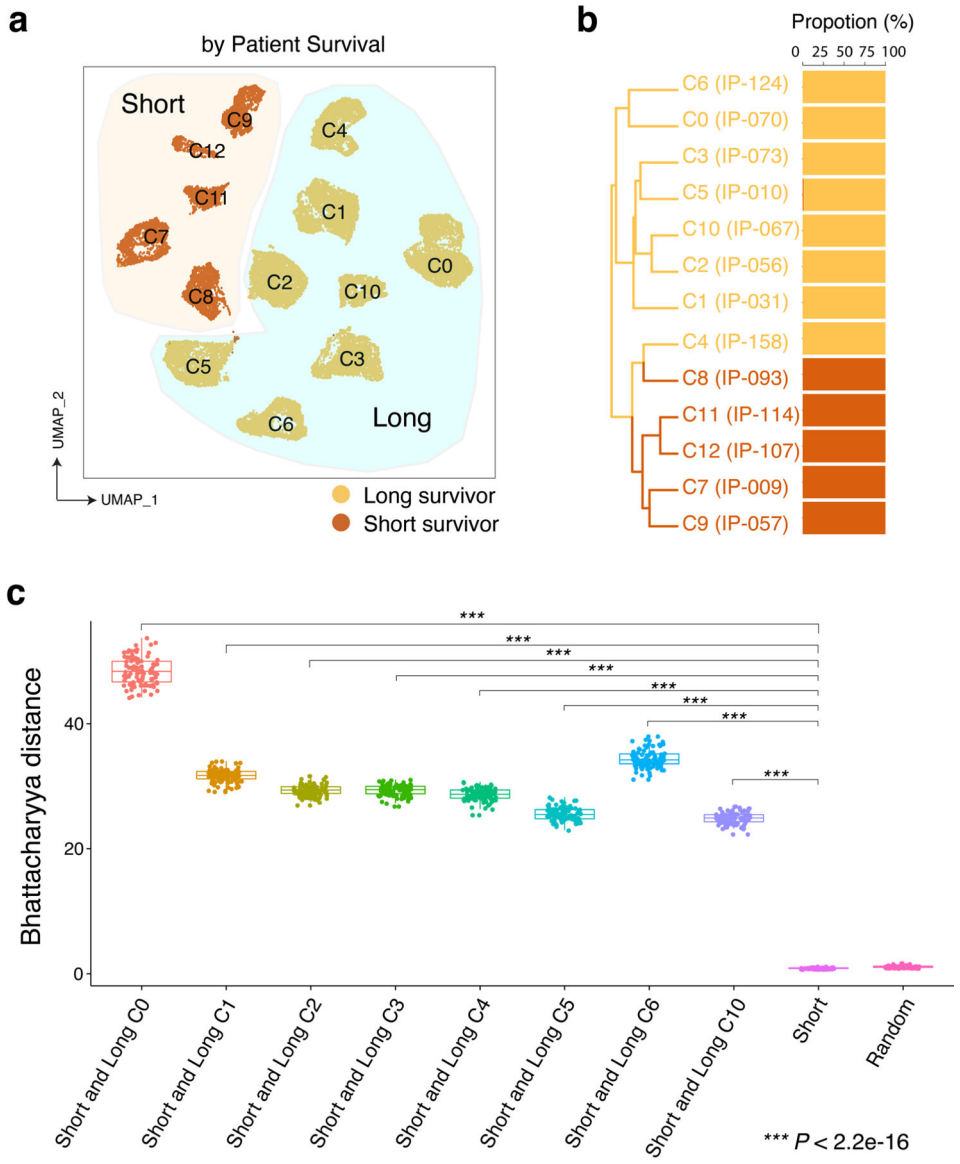
Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Extended Data



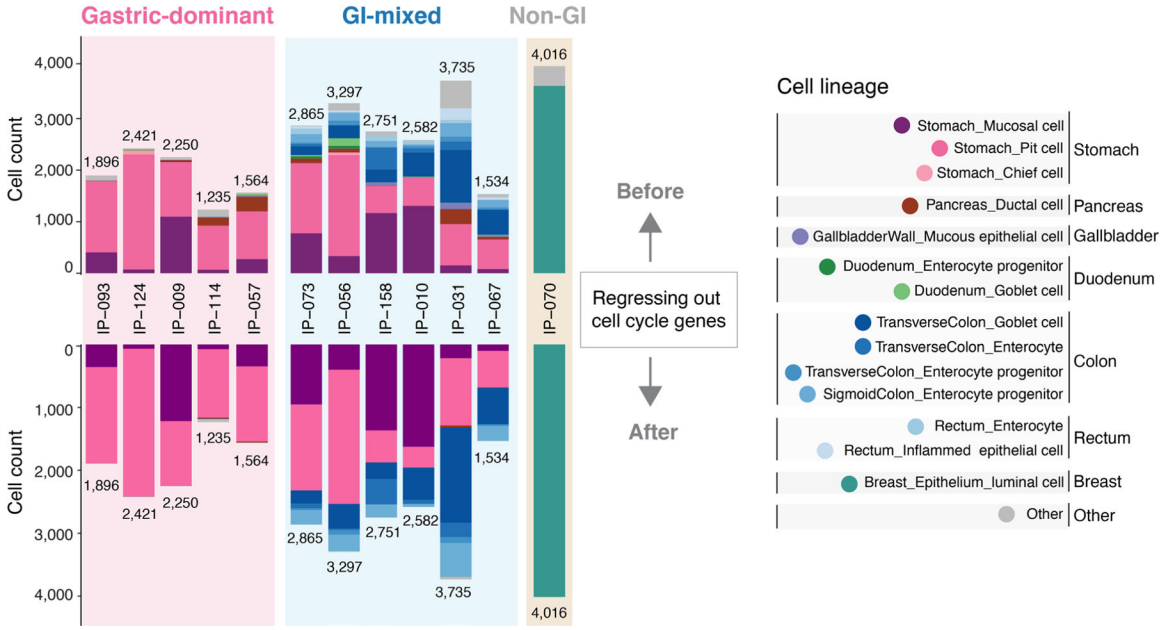
Extended Data Fig. 1 |. A single cell transcriptome map of PC.

a, t-SNE (t-distributed stochastic neighbor embedding) plots showing unbiased clustering analysis of 45,048 single cells that passed quality control in this study. Each dot represents a single cell. Cells are color coded for (left to right): the associated cell types, cell clusters, the corresponding patient origins, and survival status. **b**, t-SNE as in a, showing expression of canonical marker genes used for cell types assignment.



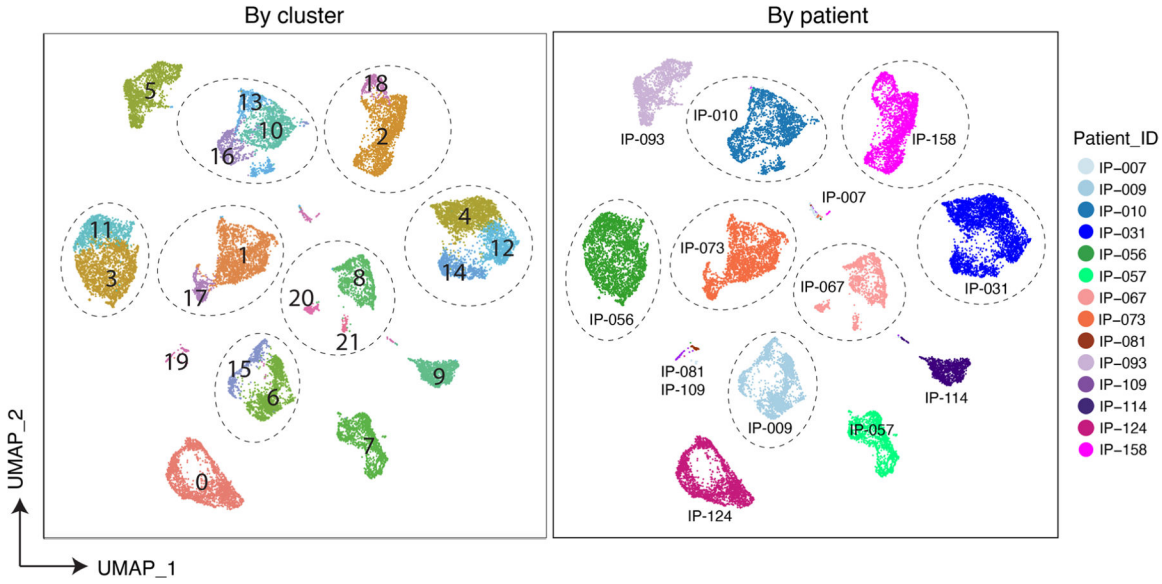
Extended Data Fig. 2 |. Relationships between tumor cell clusters and correlation with patient survival.

a, the UMAP (uniform manifold approximation and projection) plot of PC tumor cells, showing the global data structure. Tumor cell clusters from short-term survivors appeared closer to each other on the UMAP plot than to cell clusters from long-term survivors. **b**, the dendrogram showing relationships between tumor cell clusters. **c**, the Bhattacharyya pairwise distance between tumor cell clusters from samples of long and short-term survivors. Overall, the pairwise distance between clusters of long and short survivors was significantly larger than that within the clusters of Short or Random, indicating distinct transcriptomic profiles associated with survival. Each dot represents one sampling, in totally 100 times. Box, median ± interquartile range. Whiskers, the minimum and maximum values. *P* values were calculated by a two-sided Wilcoxon rank sum test with Benjamini-Hochberg correction. $P < 2.2e\text{-}16$ represents a P value approaching 0.



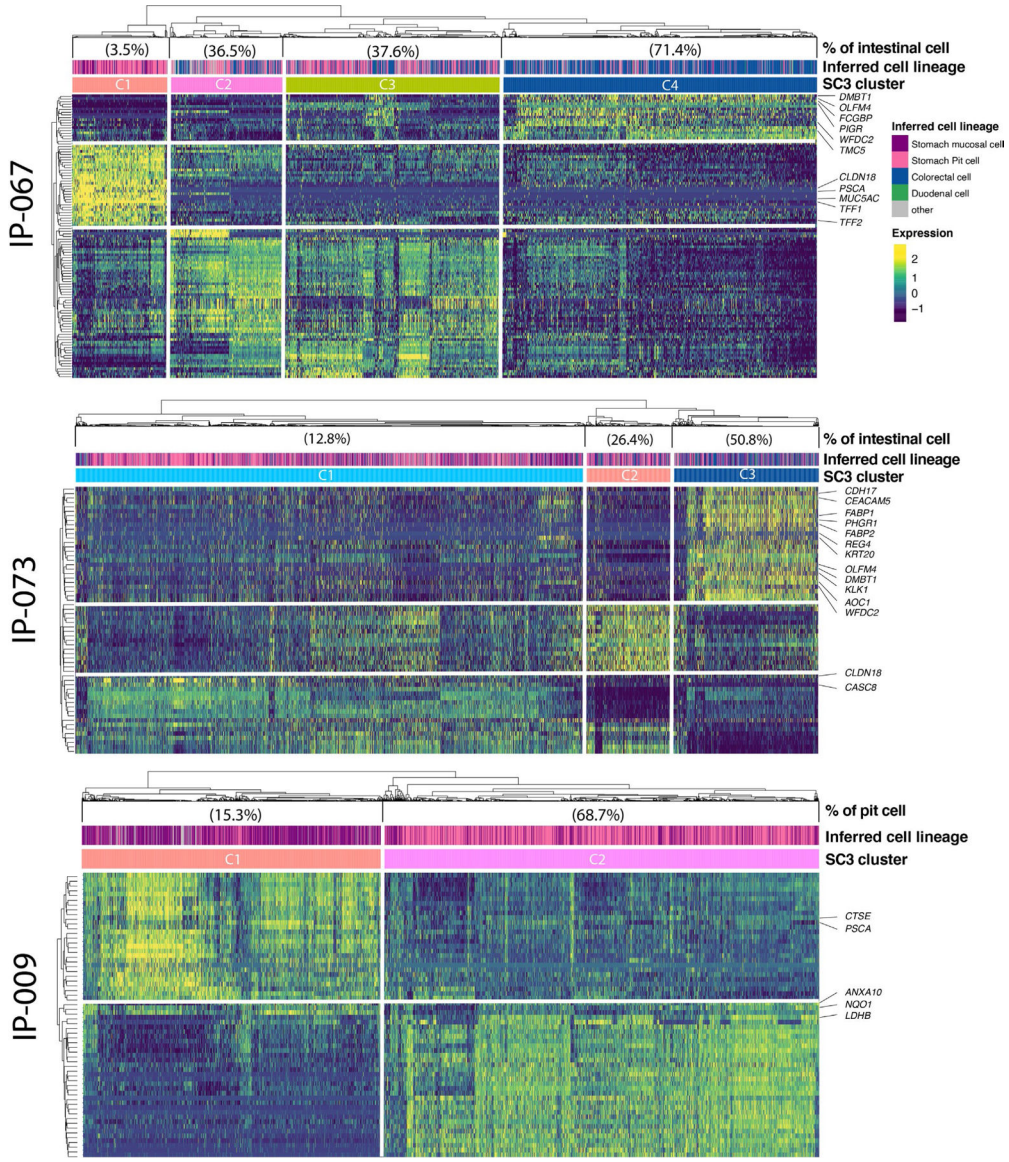
Extended Data Fig. 3 |. Cell lineage assignment was not confounded by differences in cell cycle states.

The histograms showing tumor cell lineage compositions before (top) and after (bottom) regressing out cell cycle-related genes, respectively.

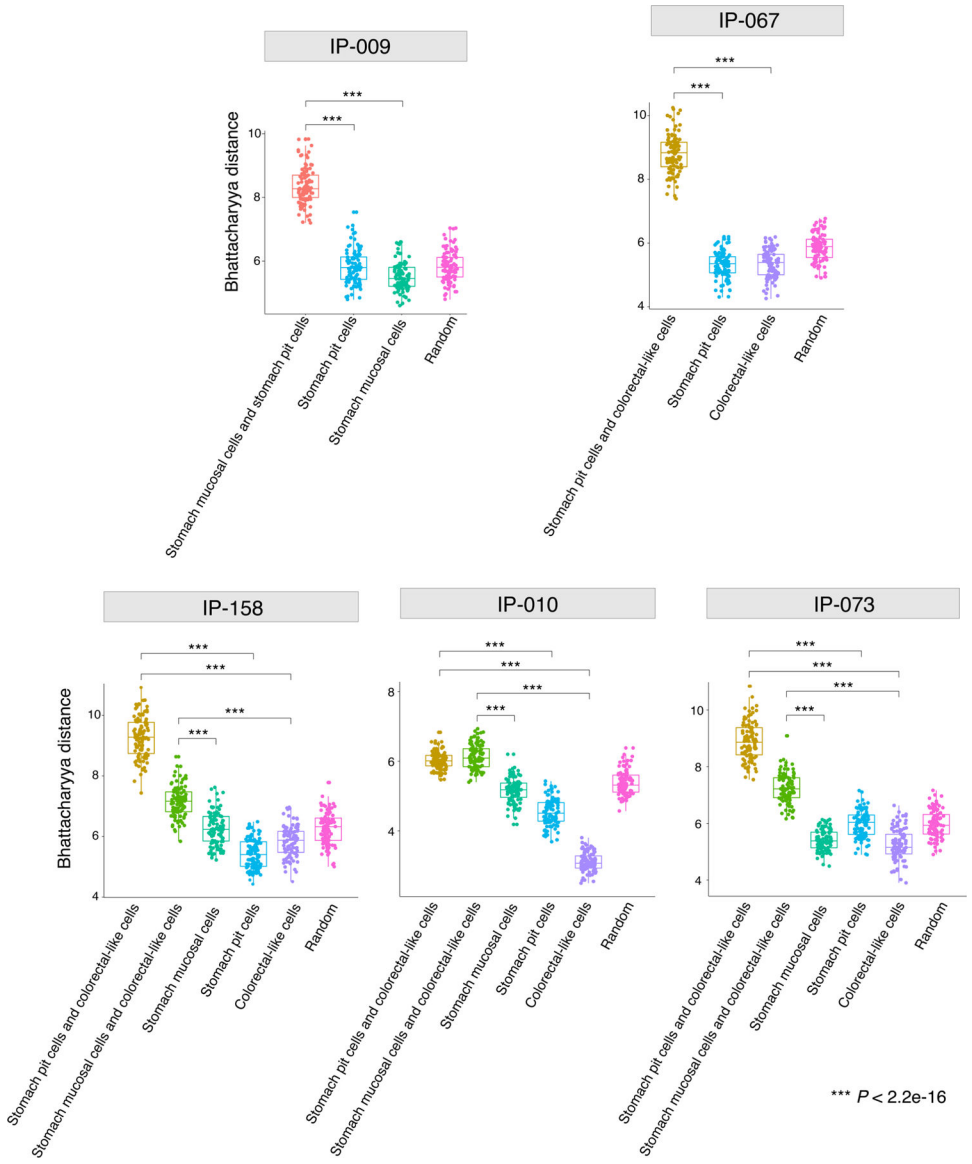


Extended Data Fig. 4 |. Unsupervised clustering analysis revealed inter-patient and intra-tumoral transcriptome heterogeneity in PC tumor cells.

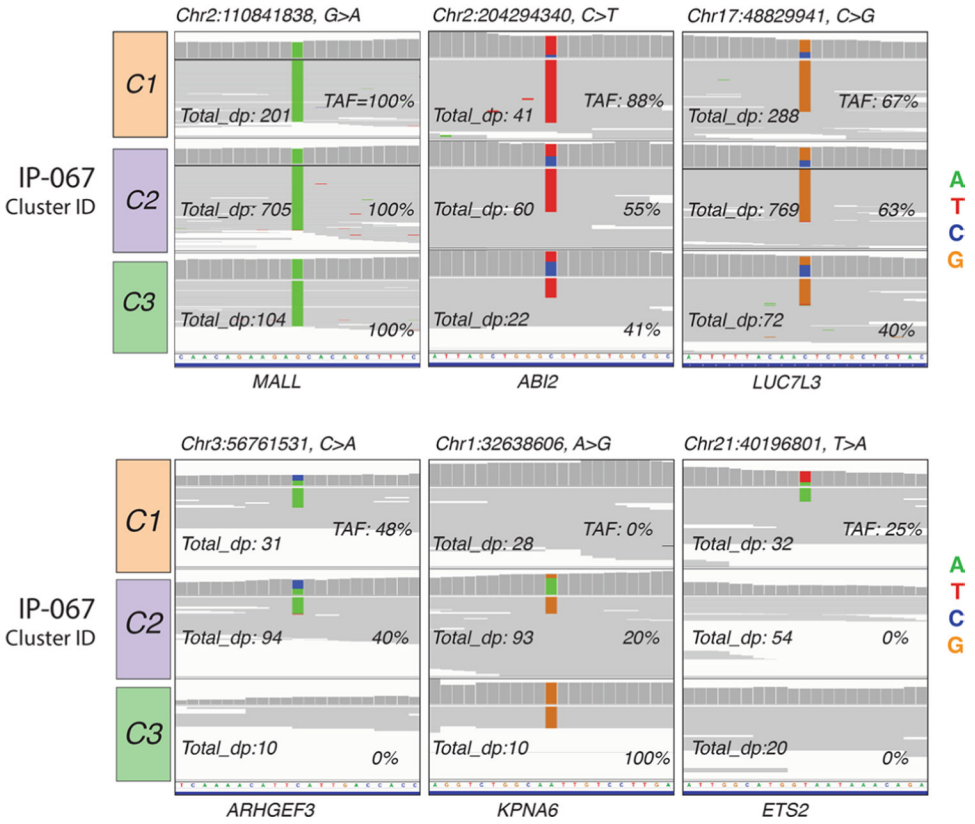
The UMAP plots showing unsupervised clustering analysis of tumor cells (using Seurat) from 14 samples underwent HCL mapping and cell lineage inference as in Fig. 1g. Cells are colored by their corresponding cluster IDs (left) and sample origins (right). Dashed circles highlight samples that formed two or more tumor cell clusters (related to Fig. 1g).



Extended Data Fig. 5 |. SC3 unsupervised clustering analysis of PC tumor cells by patient.
SC3 results of 3 representative patients are shown. Each column represents a cell. The lineage annotation is shown in the top annotation track. The fractions of intestinal cells (IP-067, IP-073) or stomach pit cells (IP-009) in each SC3 defined cell clusters are labelled at the top. Some of the representative marker genes of intestine and stomach origins are labelled on the right. Two-sided proportion tests were performed between C1 and C4 (IP-067), C1 and C3 (IP-073), and C1 and C2 (IP-009), and all are significant ($P < 2.2e{-}16$).

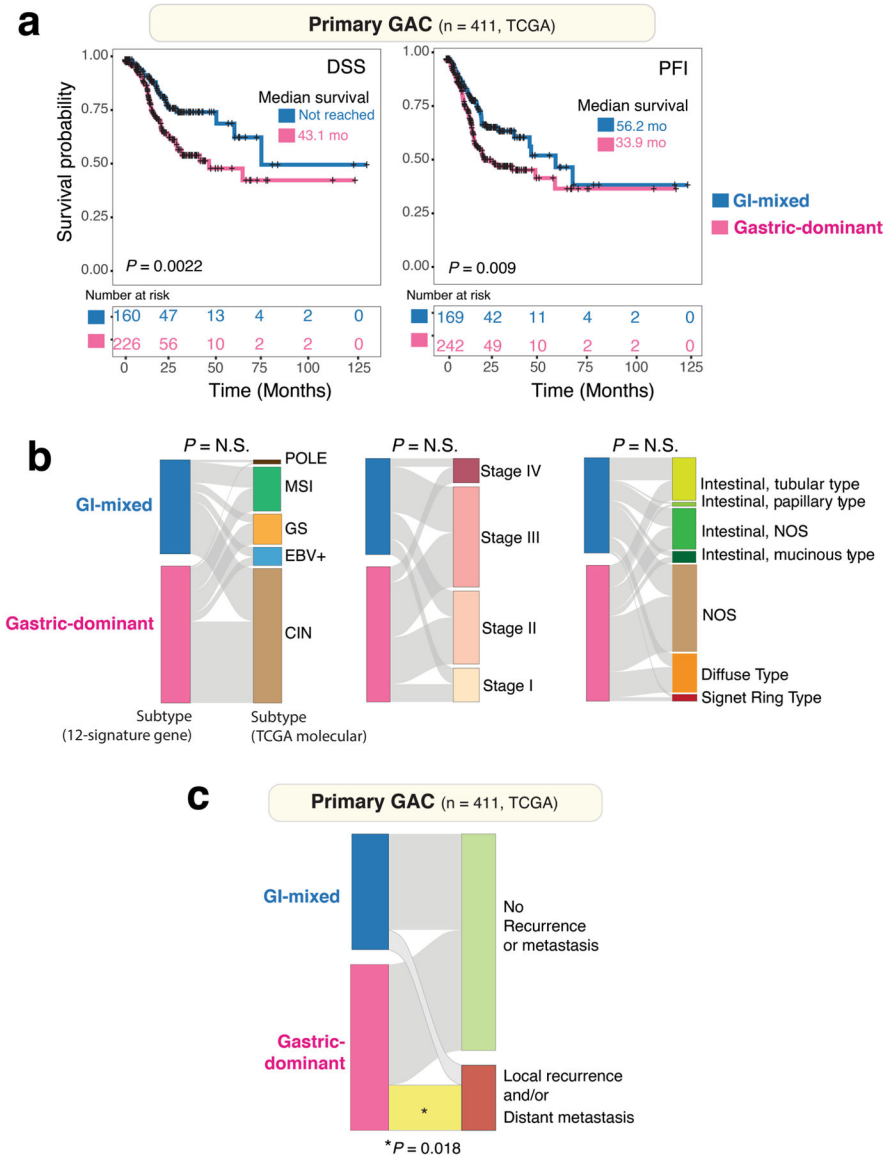


Extended Data Fig. 6 |. The Bhattacharyya distance between and within inferred cell lineages. The Bhattacharyya pairwise distance between different tumor cell lineages was computed as previously described (see Methods). Only the major lineages that had 500 or more cells were included in the analysis. The Bhattacharyya distance between cells of the same lineage and the Bhattacharyya distance between cells randomly sampled independent of lineage annotation (Random) was also computed to provide background distributions for statistical comparison. Each dot represents one sampling, in total 100 times. Box, median ± interquartile range. Whiskers, the minimum and maximum values. P values were calculated by a two-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction. $P < 2.2e{-}16$ represents a P value approaching 0.



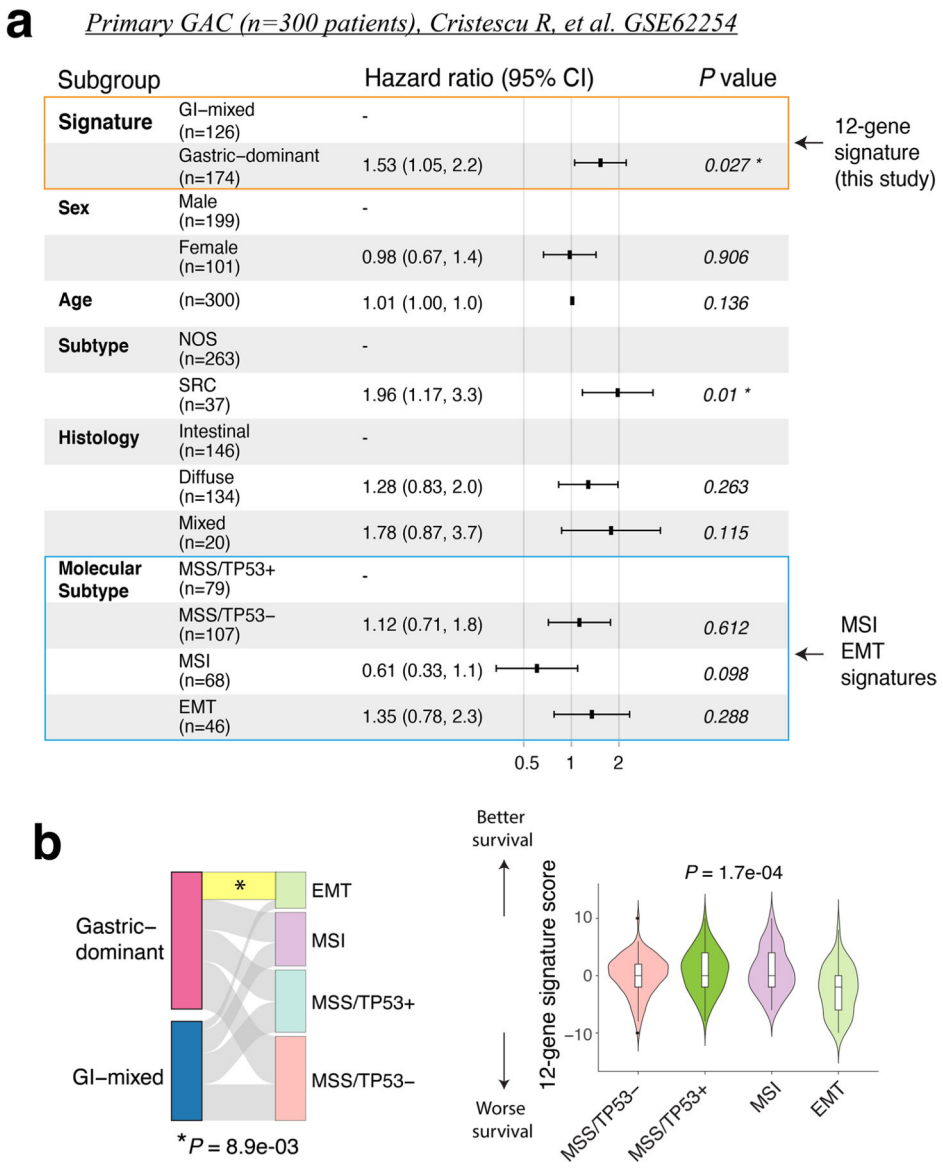
Extended Data Fig. 7 |. Representative examples of somatic variants identified on 3'UTR using scRNA-seq data.

Integrative Genomics Viewer (IGV) was used for visualization of the QC-passed somatic variants. The Bam files of Monocle defined cell clusters C1, C2, C3 of sample IP-067 were loaded to IGV and snapshots of 3'UTR mutations are shown for representative events: somatic mutations shared by PC tumor cells from all three clusters (top); mutations shared by only two of the three clusters (bottom left and middle), and mutations that were unique to one of the three clusters (bottom right) are shown. For each representative mutation across Monocle cell clusters, the gene name, chromosome, start position, base change, total read coverage, and tumor variant allele fraction (TAF) are shown. Total_dp: total read depth.



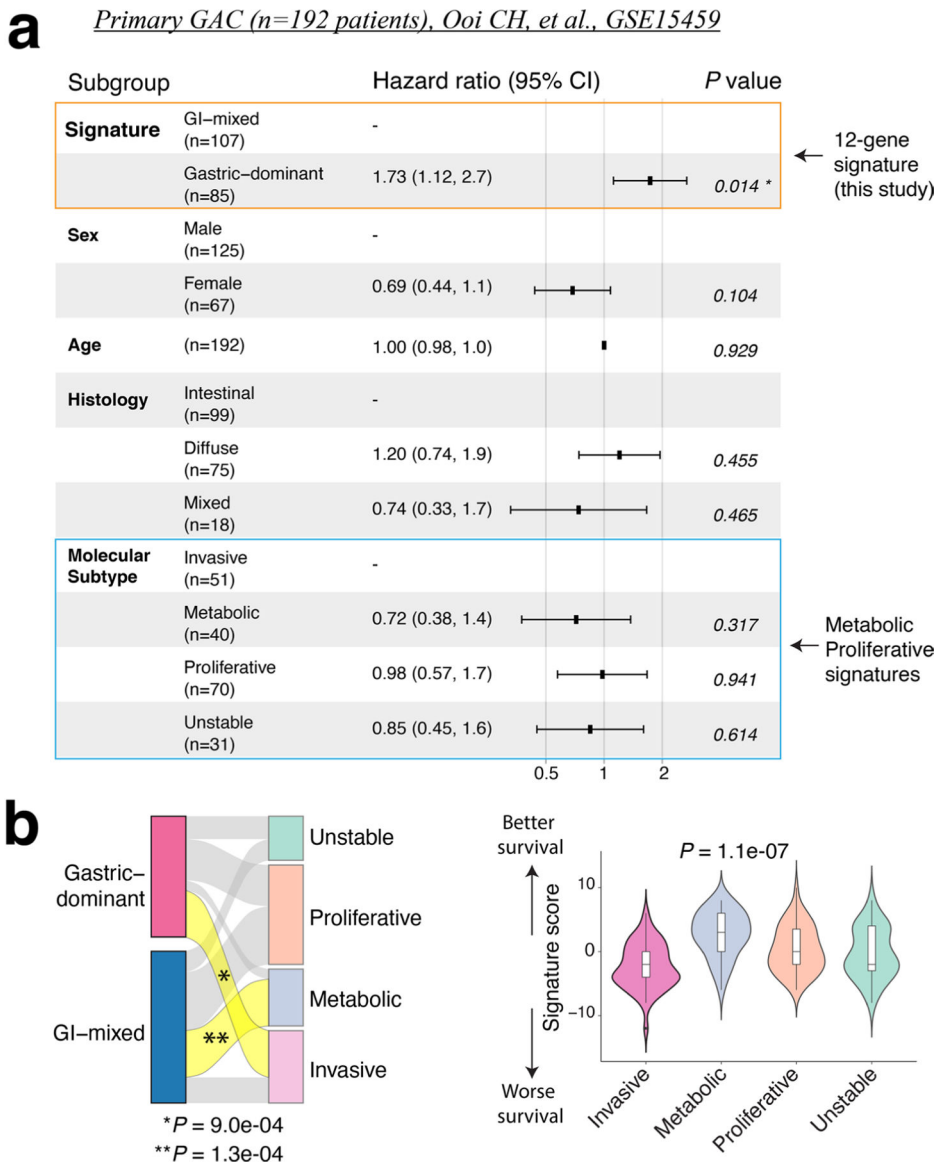
Extended Data Fig. 8 |. Prognostic significance of 12-gene signature in TCGA primary gastric cancer cohort and correlation with molecular subtypes and clinical variables.

a, Disease-specific survival (DSS, left) and progression-free interval (PFI, right) of patients whose PCs were in the GI-mixed and gastric-dominant groups defined by expression of the 12-gene signature. The analyses were performed with the Kaplan–Meier estimates and two-sided log-rank tests. Twenty-five out of 411 patients whose DSS information were not available were excluded from survival analysis. **b**, the alluvial plots display relationships between the PC subtypes defined by the 12-gene signature (left strip) and the molecular subtypes defined by TCGA multi-omic analysis (left), tumor stages (middle), histology types (right), and presence of local recurrence and/or distant metastasis (c). N.S., not statistically significant. *P* value for alluvial plots were calculated by a two-sided Fisher's Exact test.



Extended Data Fig. 9 |. Validation of the 12-gene signature in a large-scale localized GAC cohort from Cristescu R, et al.

a, The multivariate Cox proportional hazard model analysis. The 12-gene signature, clinical and histopathological variables as well as the molecular signatures defined by the original study were included. For each variable, the reference level is the first one. Block in center of error bars represent the weighted mean. Whiskers of error bars represent the 95% confidence interval. **b**, (left) Alluvial plot shows the relationships between the PC subtypes (left strip) and the molecular signatures (right strip). The two-sided Fisher's Exact test was used to calculate the P values and asterisks indicate significant enrichment events. (right) The 12-gene signature scores were calculated and compared across the four molecular groups defined by the original the study. Box, median ± interquartile range. Whiskers, 1.5X interquartile range. *P* value was calculated by one-way Kruskal-Wallis rank-sum test.



Extended Data Fig. 10 |. Validation of the 12-gene signature in a large-scale localized GAC cohort from Ooi CH, et al.

a, The multivariate Cox proportional hazard model analysis. The 12-gene signature, clinical and histopathological variables as well as the molecular signatures defined by the original study were included. For each variable, the reference level is the first one. Block in center of error bars represent the weighted mean. Whiskers of error bars represent the 95% confidence interval. **b**, (left) Alluvial plot shows the relationships between the PC subtypes (left strip) and the molecular signatures (right strip). The two-sided Fisher's Exact test was used to calculate the P values and asterisks indicate significant enrichment events. (right) The 12-gene signature scores were calculated and compared across the four molecular groups defined by the original the study. Box, median ± interquartile range. Whiskers, 1.5X interquartile range. *P* value was calculated by one-way Kruskal-Wallis rank-sum test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported by the start-up research fund provided to L.W. by the UT MD Anderson Cancer Center (MDACC); the Andrew Sabin Family Fellowship Program to L.W. by the Andrew Sabin Family Foundation; the DOD grants no. CA150334 and no. CA160445 to J.A.A.; the DOD grants no. CA160433 and no. CA170906 to S.S.; and the generous support from the Caporella, Dallas, Sultan, Park, Smith, Frazier, Oaks, Vanstekelenberg, Planjery, McNeil, Hyland and Cantu families; as well as from the Schecter Private Foundation, the Rivercreek Foundation, the Kevin Fund, the Myer Fund, the Stupid Strong Foundation, the V. Foundation, the Dio Fund, the Milrod Fund and the MDACC multidisciplinary grant programs. This study was also supported by the NIH grant no. 1S10OD024977-01 Award to the Advanced Technology Genomics Core (ATGC) and the Core grant no. CA016672 (ATGC). We thank E. J. Thompson and D. P. Pollock from the ATGC for their excellent technical assistance. We thank all of the patients who participated in this study.

Data availability

All single-cell RNA-sequencing data generated by this study have been be deposited in the European Genome-Phenome Archive (EGA, https://ega-archive.org/). The data can be accessed under the accession number EGAS00001004443. Bulk mRNA-seq expression data (normalized) generated by The Cancer Genome Atlas (TCGA) on primary stomach adenocarcinoma were downloaded from NCI Cancer Genomic Data Commons (NCI-GDC: https://gdc.cancer.gov). Three large-scale primary GAC datasets (GSE62254 (ref. [28]) and GSE15459 (refs. [27,71]), GSE84437 (ref. [72])) were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/).

References

1. Bray F et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin 68, 394–424 (2018). [PubMed: 30207593]
2. Ikoma N et al. Preoperative chemoradiation therapy induces primary-tumor complete response more frequently than chemotherapy alone in gastric cancer: analyses of the National Cancer Database 2006–2014 using propensity score matching. Gastric Cancer 21, 1004–1013 (2018). [PubMed: 29730720]
3. Mizrak Kaya D et al. Risk of peritoneal metastases in patients who had negative peritoneal staging and received therapy for localized gastric adenocarcinoma. J. Surg. Oncol 117, 678–684 (2018). [PubMed: 29205363]
4. Shiozaki H et al. Prognosis of gastric adenocarcinoma patients with various burdens of peritoneal metastases. J. Surg. Oncol 113, 29–35 (2016). [PubMed: 26603684]
5. Chen C et al. Efficacy and safety of immune checkpoint inhibitors in advanced gastric or gastroesophageal junction cancer: a systematic review and meta-analysis. Oncoimmunology 8, e1581547 (2019). [PubMed: 31069144]
6. Taieb J et al. Evolution of checkpoint inhibitors for the treatment of metastatic gastric cancers: current status and future perspectives. Cancer Treat. Rev 66, 104–113 (2018). [PubMed: 29730461]
7. Bartley AN et al. HER2 testing and clinical decision making in gastroesophageal adenocarcinoma: guideline from the College of American Pathologists, American Society for Clinical Pathology, and the American Society of Clinical Oncology. J. Clin. Oncol 35, 446–464 (2017). [PubMed: 28129524]
8. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513, 202–209 (2014). [PubMed: 25079317]

9. Wang R et al. Multiplex profiling of peritoneal metastases from gastric adenocarcinoma identified novel targets and molecular subtypes that predict treatment response. Gut 69, 18–31 (2020). [PubMed: 31171626]
10. Tirosh I et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189–196 (2016). [PubMed: 27124452]
11. Han X et al. Construction of a human cell landscape at single-cell level. Nature 581, 303–309 (2020). [PubMed: 32214235]
12. Buttner M, Miao Z, Wolf FA, Teichmann SA & Theis FJ A test metric for assessing single-cell RNA-seq batch correction. Nat. Methods 16, 43–49 (2019). [PubMed: 30573817]
13. McInnes L, Healy J, Melville J UMAP: uniform manifold approximation and projection for dimension reduction Preprint at https://arxiv.org/abs/1802.03426 (2018).
14. Smillie CS et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell 178, 714–730.e22 (2019). [PubMed: 31348891]
15. Butler A, Hoffman P, Smibert P, Papalexi E & Satija R Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol 36, 411–420 (2018). [PubMed: 29608179]
16. Cao J et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502 (2019). [PubMed: 30787437]
17. Kiselev VY et al. SC3: consensus clustering of single-cell RNA-seq data. Nat. Methods 14, 483–486 (2017). [PubMed: 28346451]
18. Puram SV et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell 171, 1611–1624.e24 (2017). [PubMed: 29198524]
19. Patel AP et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401 (2014). [PubMed: 24925914]
20. Jerby-Arnon L et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell 175, 984–997.e24 (2018). [PubMed: 30388455]
21. Stevens CE & Leblond CP Renewal of the mucous cells in the gastric mucosa of the rat. Anat. Rec 115, 231–245 (1953). [PubMed: 13031133]
22. Karam SM A focus on parietal cells as a renewing cell population. World J. Gastroenterol 16, 538–546 (2010). [PubMed: 20128020]
23. Merzel J & Leblond CP Origin and renewal of goblet cells in the epithelium of the mouse small intestine. Am. J. Anat 124, 281–305 (1969). [PubMed: 5773907]
24. Blanpain C, Horsley V & Fuchs E Epithelial stem cells: turning over new leaves. Cell 128, 445–458 (2007). [PubMed: 17289566]
25. Coker EA et al. canSAR: update to the cancer translational research and drug discovery knowledgebase. Nucleic Acids Res 47, D917–D922 (2019). [PubMed: 30496479]
26. Kim HK et al. A gene expression signature of acquired chemoresistance to cisplatin and fluorouracil combination chemotherapy in gastric cancer patients. PLoS ONE 6, e16694 (2011). [PubMed: 21364753]
27. Ooi CH et al. Oncogenic pathway combinations predict clinical prognosis in gastric cancer. PLoS Genet 5, e1000676 (2009). [PubMed: 19798449]
28. Cristescu R et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat. Med 21, 449–456 (2015). [PubMed: 25894828]
29. Mizrak Kaya D et al. Advanced gastric adenocarcinoma: optimizing therapy options. Expert Rev. Clin. Pharmacol 10, 263–271 (2017). [PubMed: 28094573]
30. Dagogo-Jack I & Shaw AT Tumour heterogeneity and resistance to cancer therapies. Nat. Rev. Clin. Oncol 15, 81–94 (2018). [PubMed: 29115304]
31. Hudler P Challenges of deciphering gastric cancer heterogeneity. World J. Gastroenterol 21, 10510–10527 (2015). [PubMed: 26457012]
32. Gullo I, Carneiro F, Oliveira C & Almeida GM Heterogeneity in gastric cancer: from pure morphology to molecular classifications. Pathobiology 85, 50–63 (2018). [PubMed: 28618420]
33. Oh SC et al. Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. Nat. Commun 9, 1777 (2018). [PubMed: 29725014]

34. Merlo LM, Pepper JW, Reid BJ & Maley CC Cancer as an evolutionary and ecological process. Nat. Rev. Cancer 6, 924–935 (2006). [PubMed: 17109012]
35. Michor F & Polyak K The origins and implications of intratumor heterogeneity. Cancer Prev. Res. (Phila.) 3, 1361–1364 (2010). [PubMed: 20959519]
36. Barros R, Freund JN, David L & Almeida R Gastric intestinal metaplasia revisited: function and regulation of CDX2. Trends Mol. Med 18, 555–563 (2012). [PubMed: 22871898]
37. Moghimi-Dehkordi B, Safaee A & Zali MR Comparison of colorectal and gastric cancer: survival and prognostic factors. Saudi J. Gastroenterol 15, 18–23 (2009). [PubMed: 19568550]
38. Qiu MZ et al. Clinicopathological characteristics and prognostic analysis of Lauren classification in gastric adenocarcinoma in China. J. Transl. Med 11, 58 (2013). [PubMed: 23497313]
39. Petrelli F et al. Prognostic value of diffuse versus intestinal histotype in patients with gastric cancer: a systematic review and meta-analysis. J. Gastrointest. Oncol 8, 148–163 (2017). [PubMed: 28280619]
40. Petitprez F et al. B cells are associated with survival and immunotherapy response in sarcoma. Nature 577, 556–560 (2020). [PubMed: 31942077]
41. Cabrita R et al. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. Nature 577, 561–565 (2020). [PubMed: 31942071]
42. Helmink BA et al. B cells and tertiary lymphoid structures promote immunotherapy response. Nature 577, 549–555 (2020). [PubMed: 31942075]
43. Najafi M et al. Macrophage polarity in cancer: a review. J. Cell. Biochem 120, 2756–2765 (2019). [PubMed: 30270458]
44. Kaneda MM et al. PI3Kγ is a molecular switch that controls immune suppression. Nature 539, 437–442 (2016). [PubMed: 27642729]
45. Kalluri R The biology and function of fibroblasts in cancer. Nat. Rev. Cancer 16, 582–598 (2016). [PubMed: 27550820]
46. Calon A et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat. Genet 47, 320–329 (2015). [PubMed: 25706628]
47. Fiori ME et al. Cancer-associated fibroblasts as abettors of tumor progression at the crossroads of EMT and therapy resistance. Mol. Cancer 18, 70 (2019). [PubMed: 30927908]

References

48. Amin MB et al. AJCC Cancer Staging Manual 8th edn (Springer, 2017).
49. Amin MB et al. The Eighth Edition AJCC Cancer staging manual: continuing to build a bridge from a population-based to a more ‘personalized’ approach to cancer staging. CA Cancer J. Clin 67, 93–99 (2017). [PubMed: 28094848]
50. Niu B et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 30, 1015–1016 (2014). [PubMed: 24371154]
51. Savas P et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat. Med 24, 986–993 (2018). [PubMed: 29942092]
52. Galili T dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31, 3718–3720 (2015). [PubMed: 26209431]
53. Zhang L et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature 564, 268–272 (2018). [PubMed: 30479382]
54. Lambrechts D et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat. Med 24, 1277–1289 (2018). [PubMed: 29988129]
55. Sade-Feldman M et al. Defining T cell states associated with response to checkpoint immunotherapy in melanoma. Cell 175, 998–1013.e20 (2018). [PubMed: 30388456]
56. Nilsen G et al. Copynumber: efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics 13, 591 (2012). [PubMed: 23442169]
57. Cillo AR et al. Immune landscape of viral- and carcinogen-driven head and neck cancer. Immunity 52, 183–199.e9 (2020). [PubMed: 31924475]

58. Azizi E et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell 174, 1293–1308.e6 (2018). [PubMed: 29961579]
59. van Dijk D et al. Recovering gene interactions from single-cell data using data diffusion. Cell 174, 716–729.e27 (2018). [PubMed: 29961576]
60. Hicks SC, Townes FW, Teng M & Irizarry RA Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics 19, 562–578 (2018). [PubMed: 29121214]
61. Hanzelmann S, Castelo R & Guinney J GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinforma 14, 7 (2013).
62. Cline MS et al. Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc 2, 2366–2382 (2007). [PubMed: 17947979]
63. Harrell FE Jr., Lee KL & Mark DB Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med 15, 361–387 (1996). [PubMed: 8668867]
64. Peng F et al. Differential expression analysis at the individual level reveals a lncRNA prognostic signature for lung adenocarcinoma. Mol. Cancer 16, 98 (2017). [PubMed: 28587642]
65. Lau SK et al. Three-gene prognostic classifier for early-stage non small-cell lung cancer. J. Clin. Oncol 25, 5562–5569 (2007). [PubMed: 18065728]
66. Kang J, D'Andrea AD & Kozono D A DNA repair pathway–focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. J. Natl Cancer Inst 104, 670–681 (2012). [PubMed: 22505474]
67. Newman AM et al. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 12, 453–457 (2015). [PubMed: 25822800]
68. Becht E et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol 17, 218 (2016). [PubMed: 27765066]
69. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol 31, 213–219 (2013). [PubMed: 23396013]
70. Liu J et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 173, 400–416.e11 (2018). [PubMed: 29625055]
71. Lei Z et al. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. Gastroenterology 145, 554–565 (2013). [PubMed: 23684942]
72. Yoon SJ et al. Deconvolution of diffuse gastric cancer and the suppression of CD34 on the BALB/c nude mice model. BMC Cancer 20, 314 (2020). [PubMed: 32293340]
73. Irizarry RA et al. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31, e15 (2003). [PubMed: 12582260]

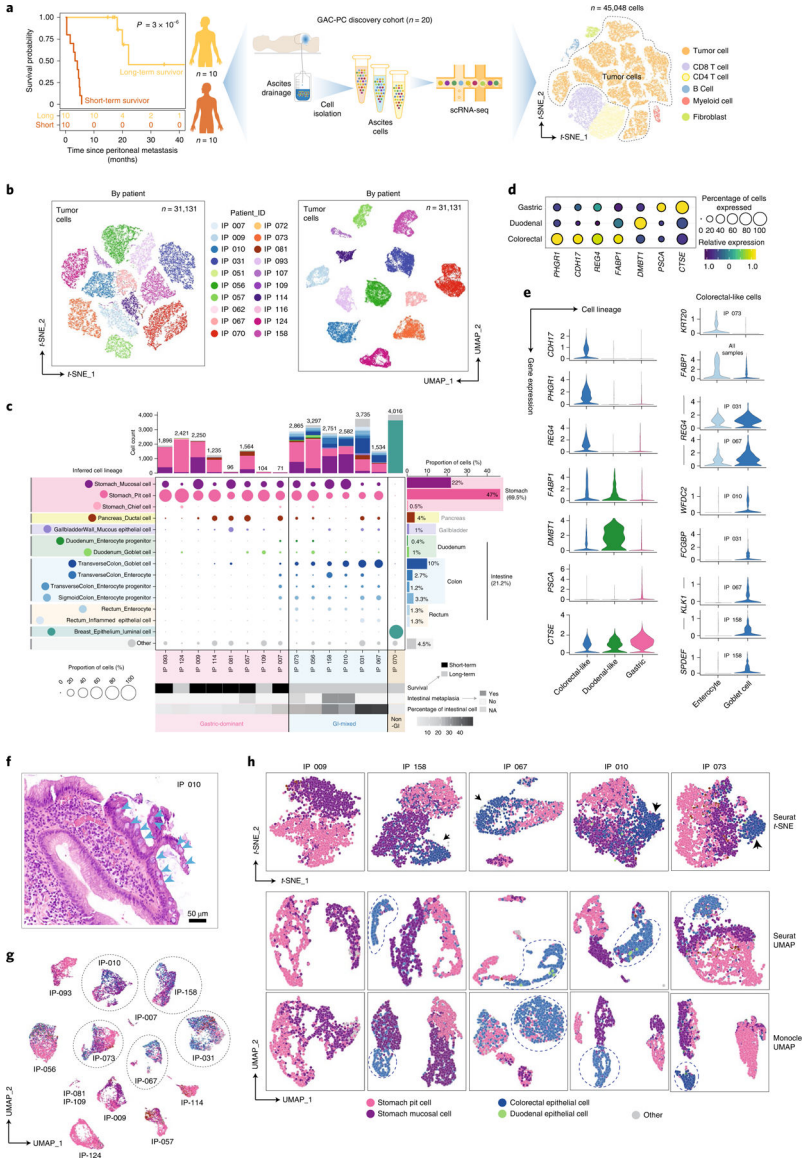


Fig. 1 |. A single-cell transcriptome map of PC and the inferred tumor cell lineages.

This study included ten short-term survivors and ten long-term survivors. **a**, Left, the Kaplan–Meier curve demonstrates a dramatic difference ($P = 3 \times 10^{-06}$ by log-rank test) in the survival time since PC diagnosis between the two groups of patients with GAC; middle, a schema of sample collection for scRNA-seq; right, *t*-SNE plot showing unbiased clustering analysis of 45,048 single cells that passed quality control in this study. Each dot of the *t*-SNE plot represents a single cell. Cells are color coded for their associated cell types. **b**, The *t*-SNE and UMAP plots of the 31,131 PC tumor cells (14 cell clusters) that were selected for subsequent analyses. Cells are color coded by their corresponding patient origins. **c**, The tumor cell lineage compositions inferred by mapping scRNA-seq data to the HCL database. The middle panel shows the HCL-defined cell lineages/types (rows) by patient (columns). The size of the circle represents, for each specific cell lineage/type, the fraction of tumor cells (among the total quality-control-passed tumor cells) in each individual PC. The circles

are color coded by defined cell lineages/types, the same as in the annotation track on the left. The histogram on the top shows, for each individual sample, the number of tumor cells accumulated on listed cell lineages/types (plus other unclassified or rare cell types). The histogram on the right shows, for each specific tumor cell lineage/type, the fraction of tumor cells (among the total quality-control-passed tumor cells) in this cohort. The bottom annotation tracks show (from top to bottom): the corresponding patient IDs, the survival groups to which the patients belong, the presence of intestinal metaplasia in their corresponding primary tumors, fractions of intestinal cells among the total quality-control-passed tumor cells in each individual PC and the PC subtypes. Classification of the PC subtypes was based on tumor cell lineage compositions (gastric-dominant if fraction of intestinal cells <20% and GI-mixed if fraction of intestinal cells ≥20%). **d**, Bubble plot showing expression of lineage-specific marker genes across different cell lineages/types. **e**, Violin plots of representative lineage-specific marker genes. **f**, A representative histology image for IP-010 demonstrating well-formed goblet cells in gastric mucosa (indicated by blue arrow heads). **g**, UMAP plot showing unsupervised clustering of 26,401 PC tumor cells from 14 samples that underwent HCL mapping and cell lineage inference as in **c**. Cells are colored by their inferred cell lineages/types. Dashed circles highlight samples that formed two or more tumor cell clusters (as labeled in the left panel of Extended Data Fig. 4). **h**, *t*-SNE and UMAP plots of PC tumor cells generated from patient-level subclustering analysis, showing that gastric cells (pink, purple) were clustered distinctly from the colorectal-like cells (dark blue).

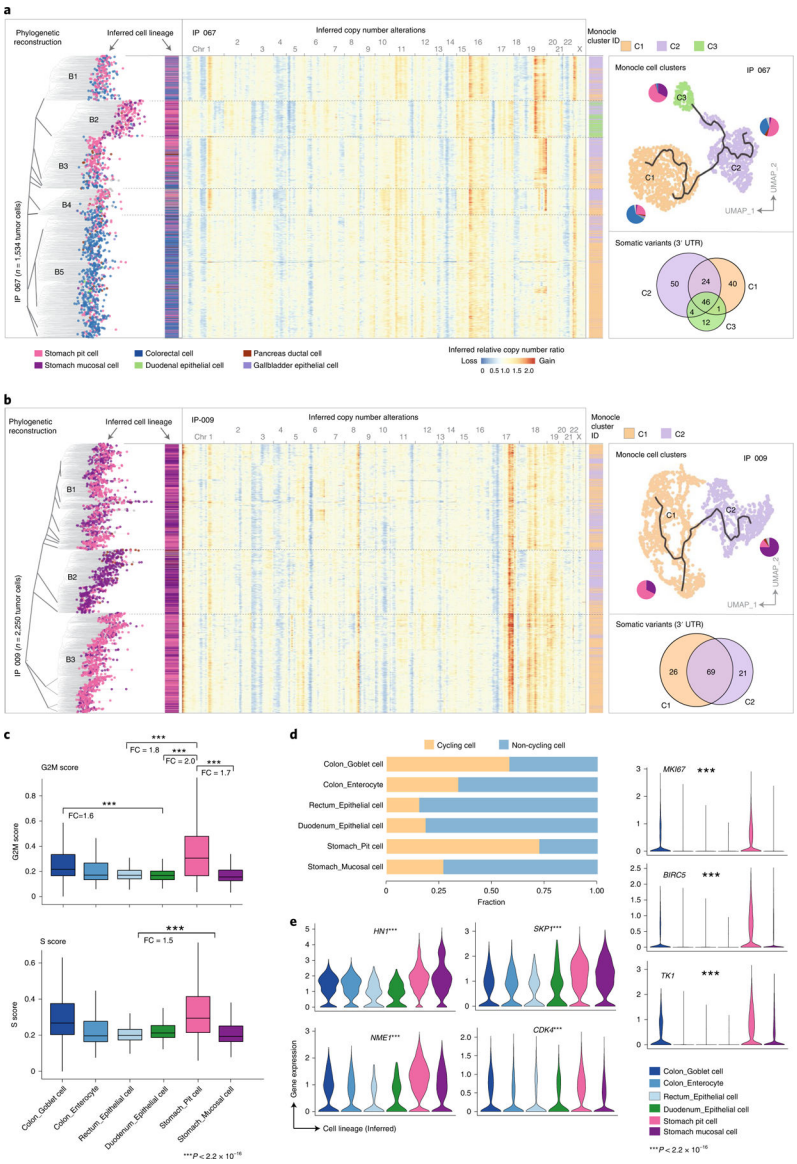


Fig. 2 |. The diversity in tumor cell lineage compositions links to ITH at transcriptomic, genotypic and molecular levels.

a, A representative sample, IP–067. Left, phylogenetic reconstruction analysis of inferred CNVs. B1–5 labels of five tumor cell subpopulations with distinct CNV profiles. Middle, heatmap showing the inferred larger-scale CNVs by chromosome; the annotation track on the left of the heatmap indicates the inferred cell lineages, and the annotation track on the right indicates Monocle-defined cell clusters. Right, top, Monocle-defined cell clusters. For each Monocle-defined cell cluster, its tumor-cell-lineage composition is shown in the small pie chart next to it; right, bottom, the Venn diagram showing shared and unique somatic variants across Monocle-defined cell clusters. Somatic variants were called from scRNA-seq data, and only variants located at the 3′ UTR were counted. **b**, Another representative sample, IP–009. B1–3 labels of three tumor cell subpopulations with distinct CNV profiles. The annotations for the remainder of **b** are in the same format as those of **a**. **c**, Comparison of tumor cell proliferative property across the inferred tumor cell lineages. Box, median ±

interquartile range. Whiskers, the minimum and maximum values. *P* values were calculated by a two-sided Wilcoxon rank-sum test with Benjamini–Hochberg correction. **d**, Proportion of cycling (cells in G2M or S phase) and non-cycling cells across the inferred cell lineages. **e**, The violin plots for representative cell-cycle-related genes that are differentially expressed across tumor cell lineages/types ($P < 2.2 \times 10^{-16}$). *P* values were calculated by one-way Kruskal–Wallis rank-sum test. $P < 2.2 \times 10^{-16}$ represents a *P* value approaching 0. Number of cells for **c** and **e**: colon goblet cells, $n = 2{,}658$; colon enterocyte cells, $n = 1{,}042$; rectum epithelial cells, $n = 1{,}578$; duodenum epithelial cells, $n = 366$; stomach pit cells, $n = 12{,}341$; stomach mucosal cells, $n = 5{,}937$. FC, fold change.

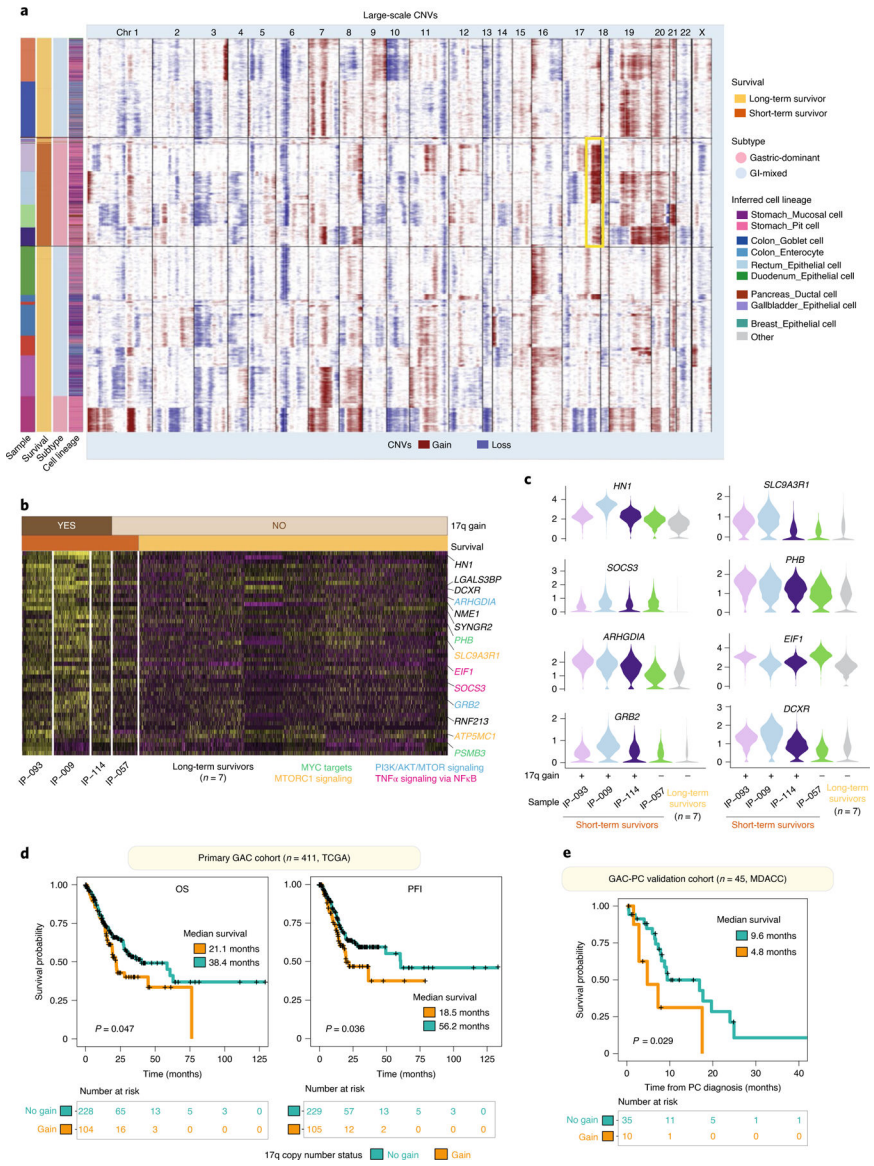


Fig. 3 |. 17q copy number gain is prevalent in cells of stomach origin and significantly associated with inferior survival.

a, The landscape of inferred large-scale CNVs for all of the tumor cells. The annotation tracks on the left indicate (from left to right) the corresponding sample IDs (the same colors as in Fig. 1b), survival groups, PC subtypes and the inferred cell lineages/types. Chromosome numbers are labeled on the top. The yellow rectangle highlights the 17q copy number gain that was nearly exclusively found in cells from the short-term survivors. **b**, The heatmap displays scaled expression values of genes upregulated in three short-term survivors (sample IDs labeled at the bottom) with evident 17q gain (annotated on the top track) and one short-term survivor and seven long-term survivors without detectable 17q changes. Biologically important genes are listed on the right, color coded by their related signaling pathways. **c**, The representative violin plots of eight genes selected from **b**. **d**,**e**, 17q copy number gain was associated with worse patient survival in the TCGA primary GAC cohort (**d**) (n = 411; only cases with survival data available were included) and an independent

GAC-PC cohort (**e**) ($n = 45$). *P* values were calculated by a two-sided log-rank test. Median survival times (in months) are labeled on the plots. MDACC, MD Anderson Cancer Center; OS, overall survival; PFI, progression-free interval.

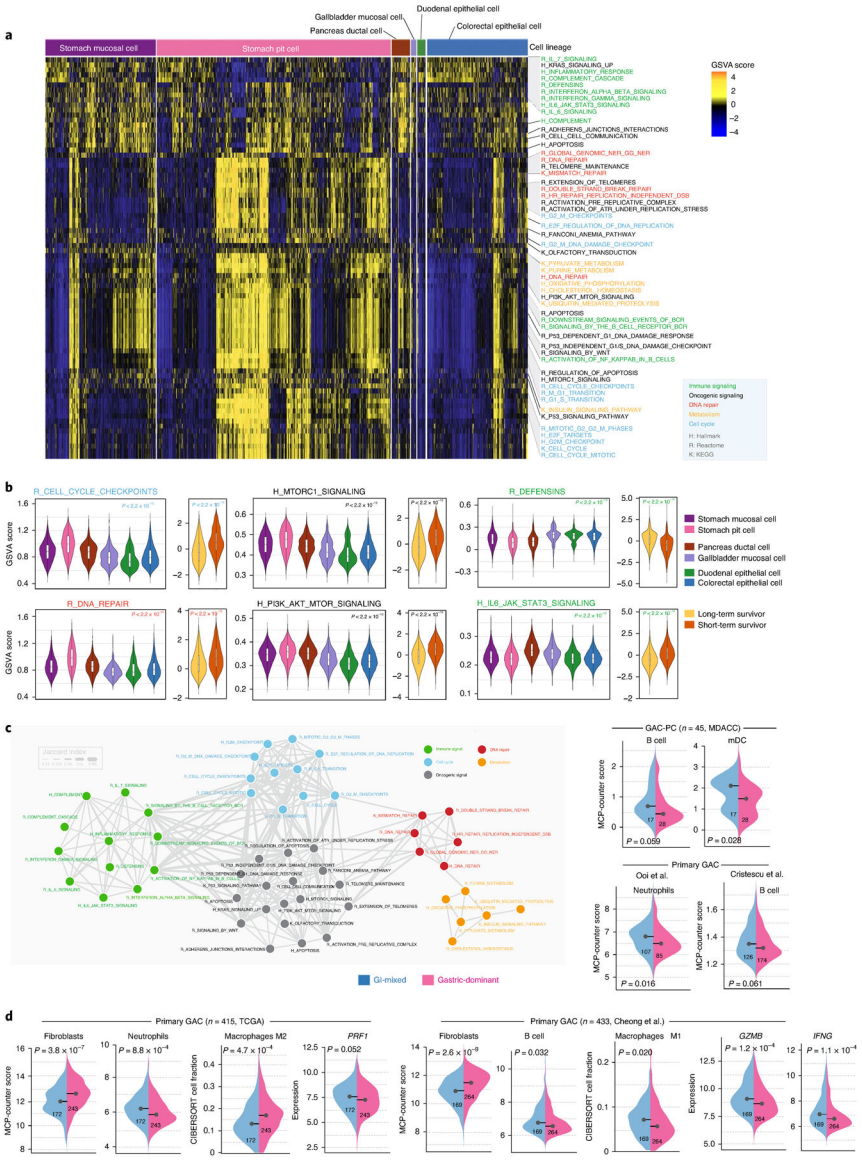


Fig. 4 |. Molecular pathway-based dissection of the transcriptomic iTH and correlation with tumor cell lineage and patient survival.

a, The transcriptomic ITH of curated gene sets, including cancer hallmark gene sets (n = 50) and gene sets from KEGG (n = 186) and reactome (n = 674) pathway databases. Each column represents a single cell. Only the pathways (rows) that were differentially expressed across different tumor cell lineages are shown. The pathway names are labeled on the right and color coded by their biological functions. **b,** representative violin plots of six pathways selected from a and Supplementary Fig. 20 that showed significant correlation with patient survival. Number of cells: stomach mucosal cells, n = 5,937; stomach pit cells, n = 12,341; pancreas ductal cells, n = 1,037; gallbladder mucosal cells, n = 285; duodenal epithelial cells, n = 366; colorectal epithelial cells, n = 5,278; long-term survivors, n = 18,428; short-term survivors, n = 6,816. Box, median ± interquartile range. Whiskers, 1.5 × interquartile range. P values across different tumor cell lineages were calculated by one-way Kruskal–Wallis rank-sum test. P values between two patient groups were calculated by a two-sided

Wilcoxon rank-sum test. $P < 2.2 \times 10^{-16}$ represents a P value approaching 0. GSVA, gene set variation analysis. **c,** The interaction networks of differentially expressed pathways displayed in **a**. The curated gene sets were colored by their biological functions. The weight of a line corresponds to its Jaccard index (a similarity metric) between each pathway pair connected by the line. **d,** Violin plots showing the differences in immune cell composition between the gastric-dominant and GI-mixed groups across multiple validation cohorts. The MCP-counter scores for a specific tumor cell lineage or the CIBErSOrT cell fractions, or normalized gene expression levels, are shown on the y axis. The black, bold, horizontal line with a dot indicates the median value of each group. P values were calculated by a two-sided Wilcoxon rank-sum test. mDC, myeloid dendritic cells.

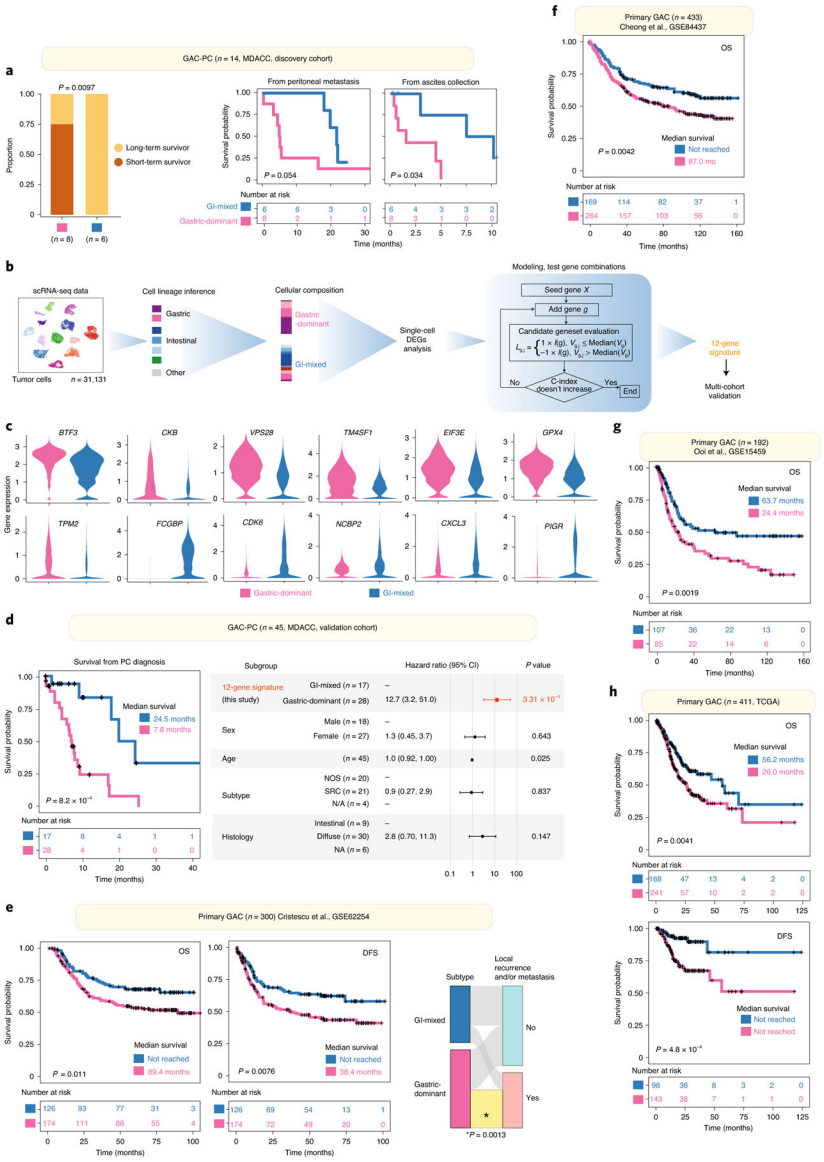


Fig. 5 |. Identification and validation of the 12-gene prognostic signature.

a, Survival analysis of the discovery GAC-PC cohort. Left, histogram showing relative proportions of long- and short-term survivors between the gastric-dominant and GI-mixed groups. *P* value was calculated by the two-tailed Fisher's exact test. Right, Kaplan–Meier plots showing the survival time (in months) since PC diagnosis and survival time since ascites collection, respectively, between patients in gastric-dominant and GI-mixed groups. *P* values were calculated by two-sided log-rank tests. **b**, A schema that illustrates the bioinformatics flow for generation of the 12-gene signature (see details in the Methods). **c**, Differential expression of the 12 signature genes between the gastric-dominant and GI-mixed groups. **d**, Survival analysis of a second independent cohort of GAC-PC patients (n = 45). Left, the Kaplan–Meier curves showing significant differences in patient survival from PC diagnosis between the two PC subtypes (the colors are the same as in panels **a**–**c**). Right, multivariate Cox proportional regression outcomes, with the 12-gene signature included. For

each variable, the reference level is the first one. The block in the center of the error bars represents the weighted mean. Whiskers of error bars represent the 95% confidence intervals. Patients whose PC belongs to the gastric-dominant subtype as defined by the 12-gene signature are significantly associated ($P = 3.31 \times 10{-}4$) with worse survival in this multivariate model. CI, confidence interval. **e**–**h**, Survival analysis of the 12-gene signature across three additional large-scale validation cohorts of localized GACs. For each cohort, the source of the dataset, the sample size, the log-rank *P* value and the median survival time (in months) are labeled on the Kaplan–Meier plot. **e**, The localized GAC cohort from Cristescu and colleagues[28]. The alluvial plots (right) show the relationships between PC subtypes (left strip) and the presence of local recurrence and/or distant metastases (right strip). The yellow band highlights the significant enrichment of local recurrence and/or distant metastases events in patients whose PCs belong to the gastric-dominant subtype. The *P* values for the alluvial plots were calculated by a two-sided Fisher's exact test. **f**, The GAC cohort from Cheong and colleagues. **g**, The GAC cohort from Ooi and colleagues[27]. **h**, The GAC cohort from TCGA.