Supplementary Fig. S1

a

Gene numbers in the STRT data

Gene numbers in the 10X data

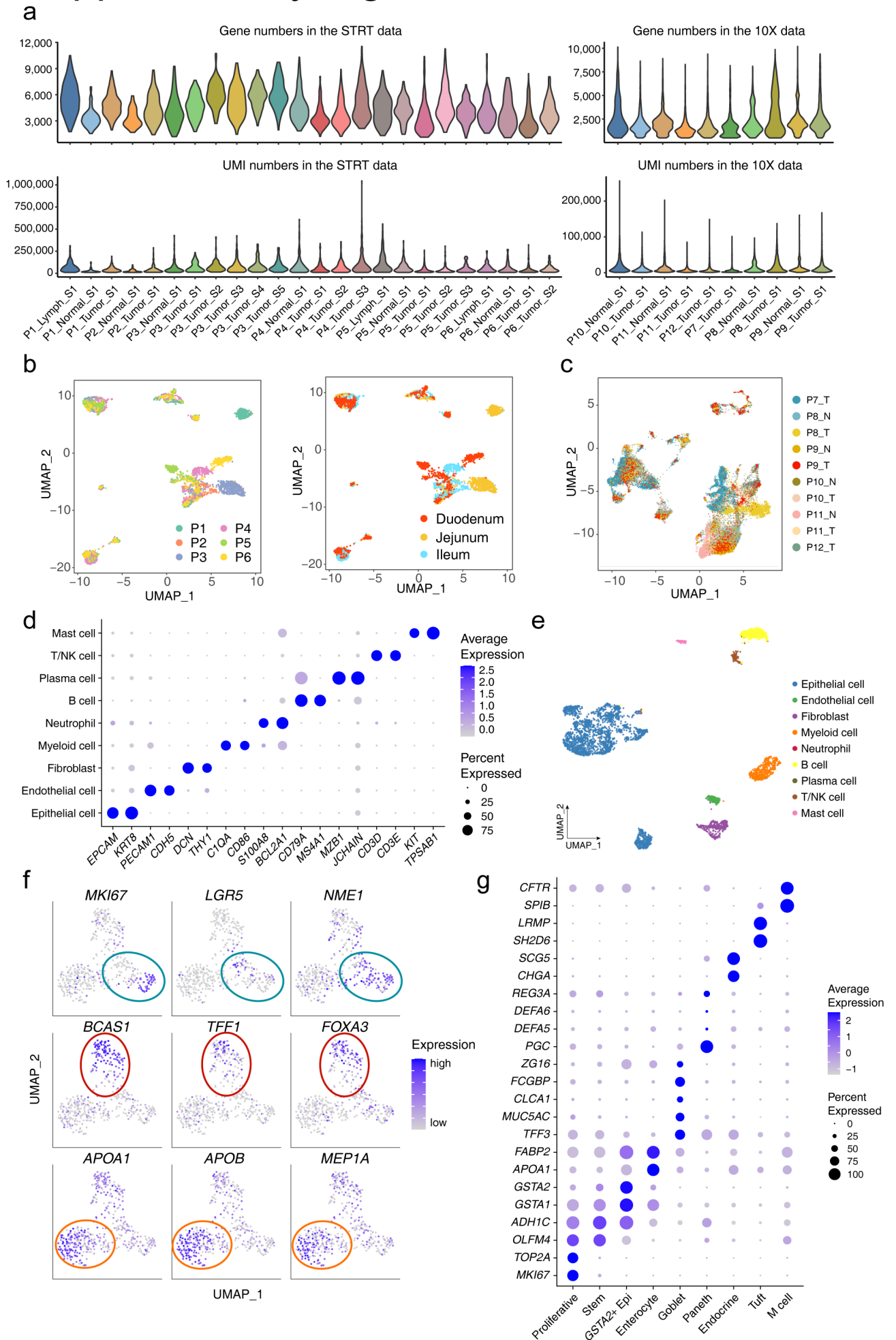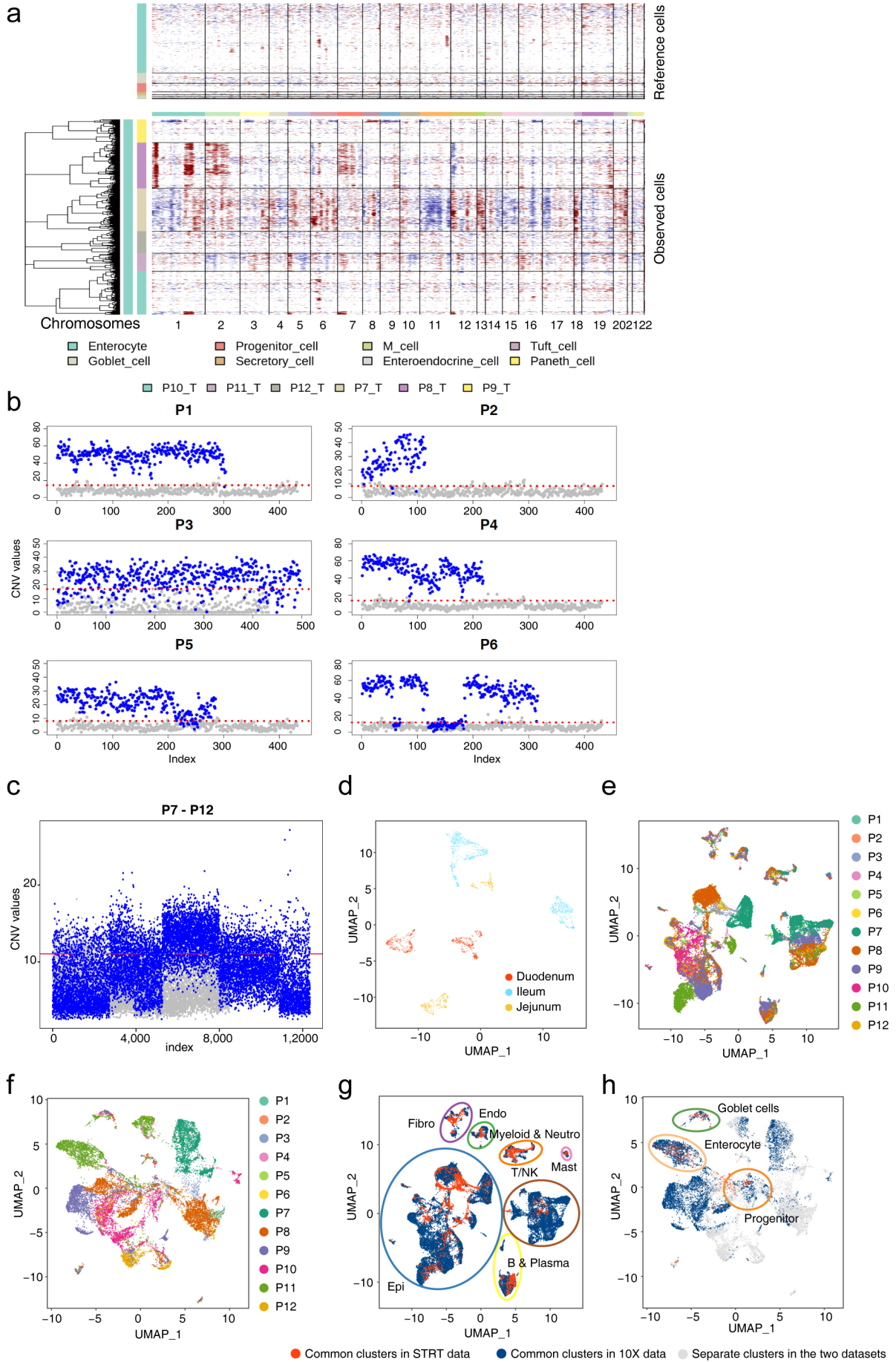UMI numbers in the STRT data

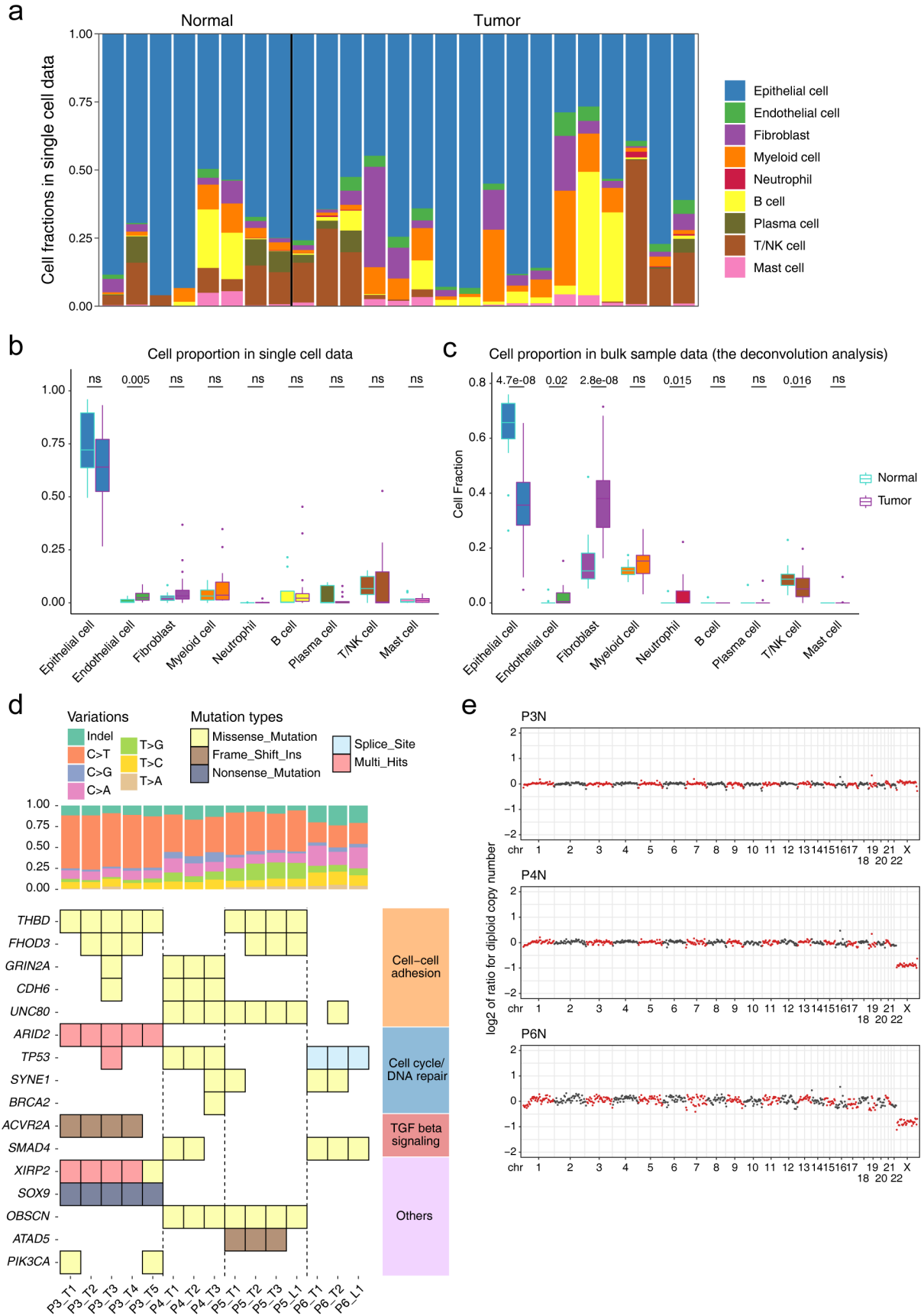UMI numbers in the 10X data

b

c

d

e

f

g

Supplementary Fig. S1 Global gene expression patterns of the STRT and 10×

datasets in SBA. a Gene number and UMI number of the STRT and 10× datasets. **b, c**

UMAP plot exhibiting the patient and intestinal region sources of the STRT and 10×

datasets. **d** Dot plot presenting the normalized expression levels of corresponding

markers of each cell cluster in the 10× dataset. **e** UMAP plot exhibiting the identified

clusters based on TF scores of the STRT dataset. **f, g** UMAP feature plot and dot plot

presenting the normalized expression levels of corresponding markers of normal

epithelial cell clusters in the STRT and 10× datasets.

Supplementary Fig. S2

a

Reference cells

Observed cells

Chromosomes 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Enterocyte   Progenitor_cell   M_cell   Tuft_cell
Goblet_cell   Secretory_cell   Enteroendocrine_cell Paneth_cell

P10_T   P11_T   P12_T   P7_T   P8_T   P9_T

b

P1   P2   P3   P4   P5   P6

CNV values

Index   Index

c

P7 - P12

CNV values

index

d

UMAP_2   UMAP_1

Duodenum   Ileum   Jejunum

e

UMAP_2   UMAP_1

P1   P2   P3   P4   P5   P6   P7   P8   P9   P10   P11   P12

f

UMAP_2   UMAP_1

P1   P2   P3   P4   P5   P6   P7   P8   P9   P10   P11   P12

g

Fibro   Endo   Myeloid & Neutro   T/NK   Mast   B & Plasma   Epi

UMAP_2   UMAP_1

h

Goblet cells   Enterocyte   Progenitor

UMAP_2   UMAP_1

Common clusters in STRT data   Common clusters in 10X data   Separate clusters in the two datasets

Supplementary Fig. S2 Inferred CNV features and integration of normal epithelial cells from the STRT and 10× datasets. a The upper panel exhibiting large-scale CNVs of single cells inferred based on normal epithelial cells in the 10× dataset. The lower panel exhibiting large-scale CNVs of single cells inferred based on epithelial cells from tumor tissues in the 10× dataset. **b**, **c** Scatterplot exhibiting the partition of the malignant cells with normal epithelial cells in the STRT and 10× datasets. **d** UMAP plot exhibiting cell clusters of the malignant cells from patients in the STRT dataset. **e**, **f** UMAP plot showing the integration of all the acquired cells and epithelial cells from the STRT and 10× datasets. **g**, **h** UMAP plot showing the integration of all the acquired cells and epithelial cells from the STRT and 10× datasets, with colors presenting the STRT or 10× datasets.
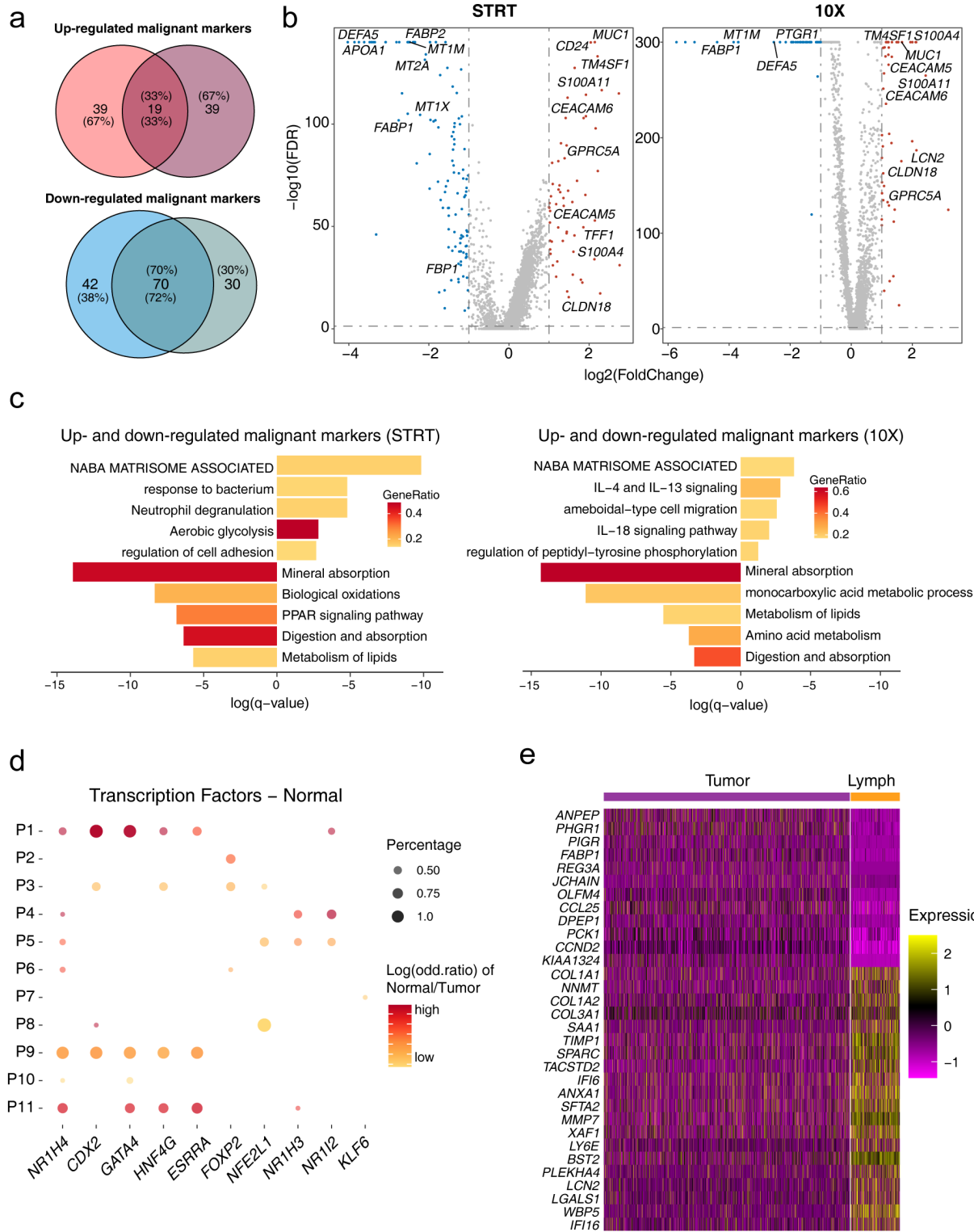
Supplementary Fig. S3

a

Normal   Tumor

Cell fractions in single cell data

Epithelial cell
Endothelial cell
Fibroblast
Myeloid cell
Neutrophil
B cell
Plasma cell
T/NK cell
Mast cell

b

Cell proportion in single cell data

c

Cell proportion in bulk sample data (the deconvolution analysis)

Normal
Tumor

d

Variations
Indel
C>T
C>G
C>A
T>G
T>C
T>A

Mutation types
Missense_Mutation
Frame_Shift_Ins
Nonsense_Mutation
Splice_Site
Multi_Hits

THBD
FHOD3
GRIN2A
CDH6
UNC80
ARID2
TP53
SYNE1
BRCA2
ACVR2A
SMAD4
XIRP2
SOX9
OBSCN
ATAD5
PIK3CA

Cell-cell adhesion

Cell cycle/ DNA repair

TGF beta signaling

Others

e

P3N

P4N

P6N

log2 of ratio for diploid copy number

Supplementary Fig. S3 Proportions of different cell types and somatic mutations landscape of SBA. a Histogram showing cell type proportion of different cell clusters in the single-cell dataset. **b** Boxplot showing the cell composition percentage in the single-cell data, with "ns" representing no significance and *P* values calculated by *t*-test. **c** Boxplot showing the cell composition percentage inferred by the deconvolution analysis, with "ns" representing no significance and *P* values calculated by *t*-test. **d** The landscape of somatic mutations of samples from P3–P6. **e** The CNV plot representing the CNV values of normal epithelial samples from P3, P4 and P6.
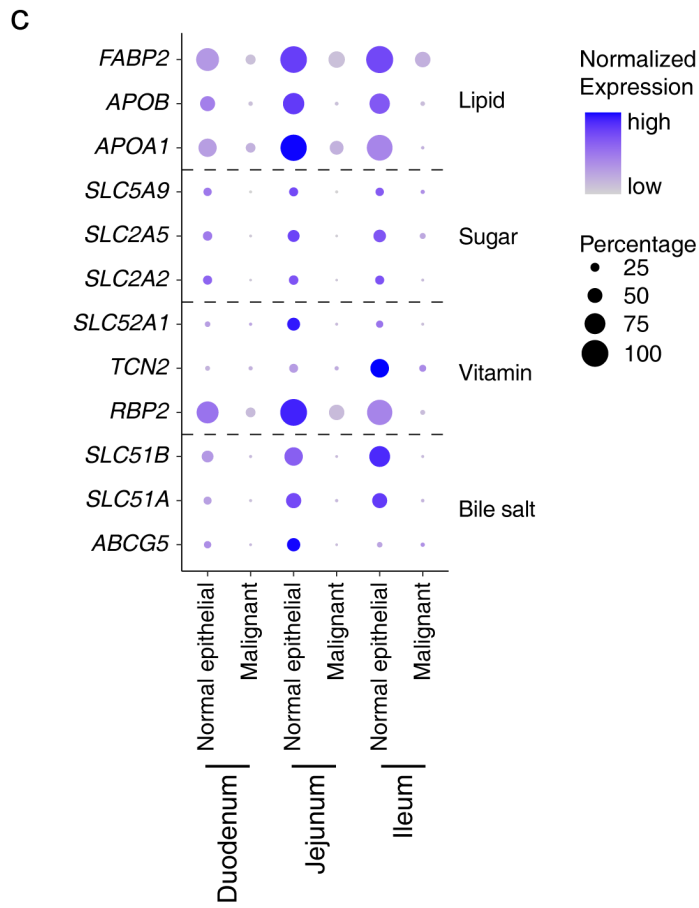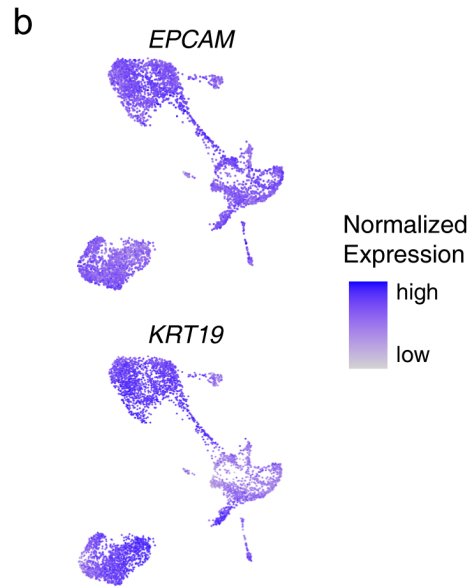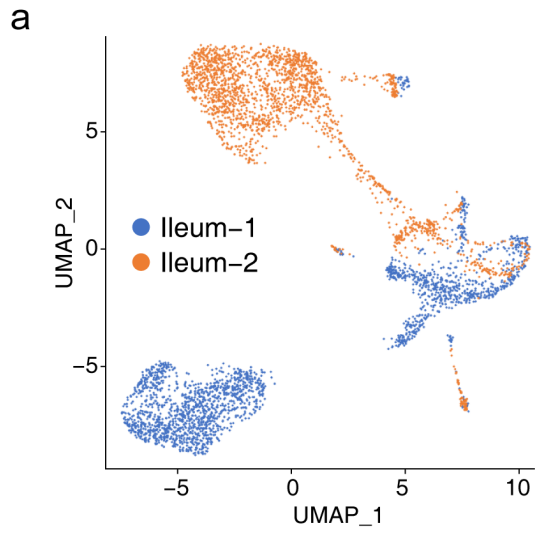
Supplementary Fig. S4

Supplementary Fig. S4 DEGs and Transcriptomic signatures of malignant cells of SBA. a Venn plot showing interactions of up- and downregulated malignant markers between STRT and 10× datasets. **b** Volcano plot exhibiting DEGs of the malignant cells compared with normal epithelial cells in STRT and 10× datasets. **c** Enrichment terms for up- and downregulated marker genes of the malignant cells in the STRT and 10× datasets. **d** Dot plot showing the log odd ratio values of normal/tumor of corresponding TFs, with *P* values calculated by χ^2 test. **e** Heatmap exhibiting DEGs of malignant cells from lymph node and primary tumor sites.

Supplementary Fig. S5

a

b

EPCAM

KRT19

Normalized
Expression

high

low

c

Supplementary Fig. S5 Analysis of the published ileal samples and the absorption and transportation related markers. a, **b** UMAP plot exhibiting the integrated datasets of two ileal normal samples with epithelial cell markers. The dataset is a subset of the GSE125970 dataset in the GEO database. **c** Dot plots presenting the normalized expression levels of absorption and transportation related markers in integrated datasets.
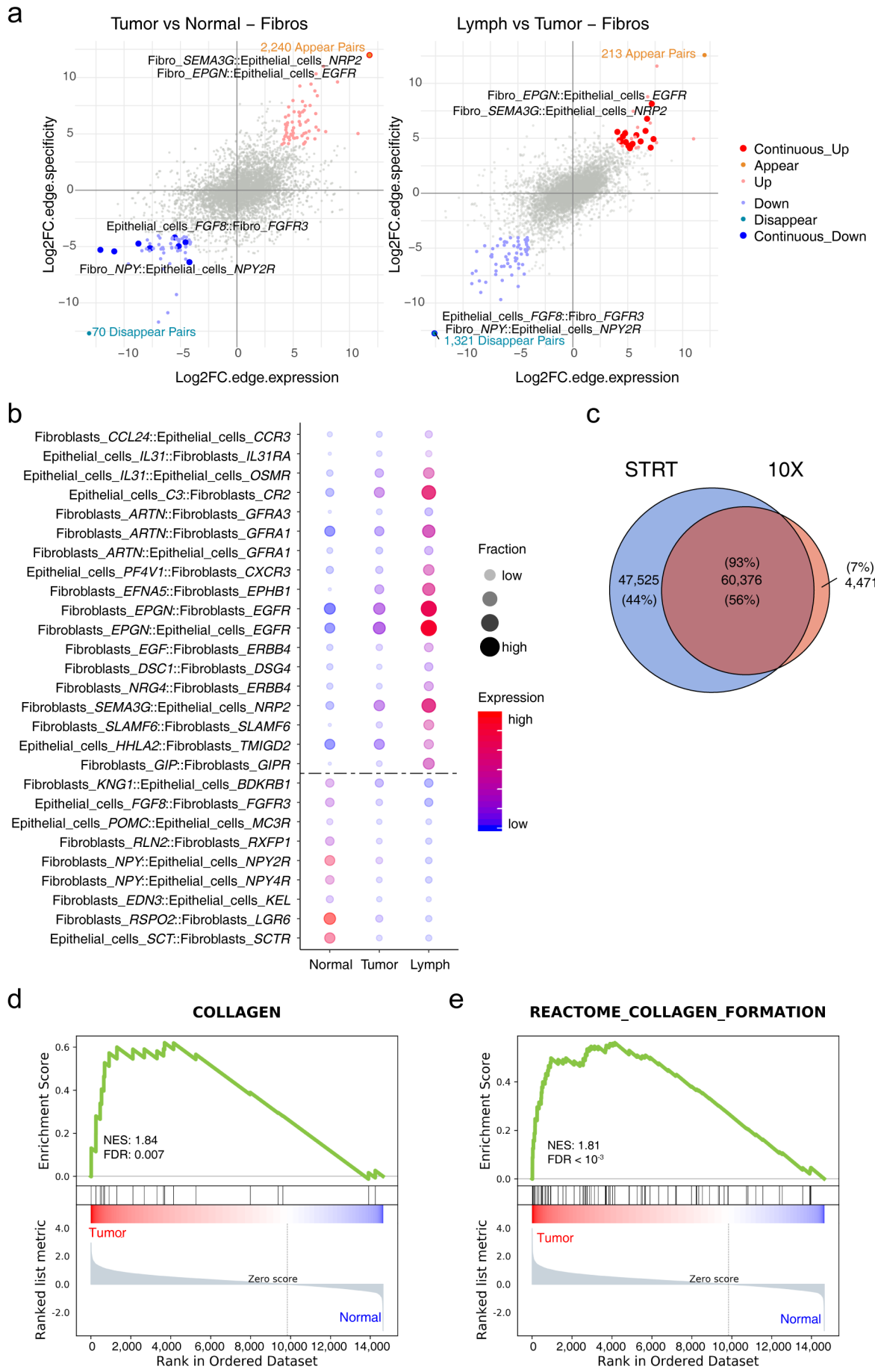
Supplementary Fig. S6

a

-log10(P)

0 2 4 6 10 20

Chromosome segregation
DNA conformation change
Phagocytosis, recognition
Metallothioneins bind metals
Cellular responses to stimuli
Positive regulation of protein kinases
Cholesterol biosynthesis
Neutrophil activation
NABA MATRISOME ASSOCIATED
Response to toxic substance
Regulation of peptidase activity
Response to bacterium
Response to wounding
Response to peptide
Response to molecule of bacteria
Multicellular organismal homeostasis
Regulation of viral life cycle
Interleukin-4 and Interleukin-13 signal
NABA CORE MATRISOME
Extracellular matrix organization

EMT
Metallothioneins bind metals
Epithelial 2
Epithelial 1
Immune
Cell cycle
Metabolism

b

P5TumorS2_1 P5TumorS2_1 P6TumorS1_1 P6TumorS1_2

P5TumorS2_3 P5TumorS2_4 P6TumorS1_3 P6TumorS1_4

c

Cell types

- Cell_cycle
- Mitochondria_related
- Metabolism_related
- EMT
- Mesenchymal
- Secretory_cell

Patient

- P5
- P6

Expression

-2 0 2

AURKB
MKI67
CDK1
CDCA8
CCNB1
ATP5B
ATP5A1
C1QBP
UQCRC2
SUCLG1
RPL13
ID2
HES1
EIF5B
CD59
LAMC2
TIMP2
TIMP1
SEMA3D
DCLK1
TCF4
ZEB2
ADAM22
MUC4
DUOX2
FABP1
DEFA5
REG1B

d

Cell cycle program

Cell Cycle
Cell division
Mitotic cell cycle phase transition
Microtubule cytoskeleton organization
DNA replication
DNA conformation change
Regulation of mitotic nuclear division
PID AURORA B PATHWAY
Cell cycle G1/S phase transition
Chromosome Maintenance
PID PLK1 PATHWAY

0 20 40 60 80
-log10(P)

Mitochondria program

Oxidative phosphorylation
GTP hydrolysis of the 60S ribosomal subunit
Humoral immune response
Proton transmembrane transport
Energy coupled proton transmembrane transport
Retina homeostasis
Antimicrobial humoral response
Aging
Response to inorganic substance
PID IL4 2PATHWAY
Negative regulation of glial cell differentiation
Cardiac muscle contraction

0 2 4 6 8 10 12 14 16
-log10(P)

Metabolism program

The citric acid (TCA) cycle and respiratory electron transport
Antibiotic metabolic process
Detoxification
Pyruvate metabolism
Response to inorganic substance
F1F0-ATP synthase, mitochondrial
Regulated exocytosis
CTCF-nucleophosmin-PARP-HIS-KPNA-LMNA-TOP complex
Translation
Ethanol oxidation
Positive regulation of kinase activity

0 2 4 6 8 10 12 14 16
-log10(P)

EMT program

Myeloid leukocyte activation
Regulation of cell adhesion
Negative regulation of cell proliferation
Cell-cell adhesion via adhesion molecules
Response to wounding
Positive regulation of defense response
Positive regulation of locomotion
Viral life cycle
Regulation of inflammatory response
Response to bacterium
Acute inflammatory response

0 2 4 6 8 10 12
-log10(P)

e

Gene module correlation

1.00	0.85	0.73	0.65
0.85	1.00	0.79	0.77
0.73	0.79	1.00	0.63
0.65	0.77	0.63	1.00

Cell cycle Mitochondria Metabolism EMT

Spearman Correlation
high
low

f

Component_2
Component_1

Signaling Entropy Rate
0.93
0.90
0.87
0.84

g

P2 - P4   P7 - P12

AURKB   AURKB
CENPF   CENPF
TOP2A   TOP2A
PRC1   PRC1
MKI67   MKI67
UBE2C   E2F8

MT-ND5   ATP5I
MT-ND2   MT-ND5
MT-ATP6   MT-ND2
             MT-ATP6

ALDH9A1   EIF3E
ALDH1A1   EIF3I
EIF3E

LAMC2   TFF1
CEACAM7   SAA2
DUSP5   LAMC2
             CEACAM7
             DUSP5

Pseudotime
- Pre−branch
- Cell fate 1
- Cell fate 2

Cluster
- 1
- 2
- 3
- 4

Expression
high
low

Supplementary Fig. S6 Validation of gene expression programs of malignant cells of SBA. a Heatmap showing enrichment terms for clustered expression signatures. **b** Heatmap exhibiting PC1–4 from the representative PCA results of P5_Tumor_S2 and P6_Tumor_S1 samples from P5 and P6. **c** Heatmap showing corresponding marker genes of the malignant cells related to gene expression programs in P5 and P6. **d** Enrichment terms for marker genes of identified gene expression programs. **e** Heatmap showing the Spearman correlation of the malignant cells belong to different expression programs. **f** Scatterplot showing the developmental trajectory of malignant cell labeled with the signaling entropy rate. **g** Heatmap showing the identified gene expression programs of the malignant cells along the pseudotime trajectory of P2–P4 in the STRT dataset and P7–P12 in the 10× dataset.

Supplementary Fig. S7

a

WNT5A
BMP2
BMP4
BMP5
PLA2G2A
CXCL12
ADH1B
MFAP5
CTHRC1
INHBA
LMO7
KCNQ1OT1
ADAMDEC1
CCL8
CCL11
CCL13
TOP2A
CCNB1
PRC1
MKI67
ACTA2
TAGLN
BGN
MYH11

C1 C2 C5 C6 Proliferative Myofibroblast

Average Expression
2
1
0
-1

Percent Expressed
0
25
50
75
100

b

ACTA2

Normal   Tumor

c

Normal   Tumor

MGST1
SFRP1
TNXB
CFD
MFAP5
PLA2G2A
FBLN2
DPT
FBN1
GSN
DCN
ADH1B
SFRP2
FBLN1
S100A4
CD55
MGP
NOTCH3
TPM2
RGS5
BGN
TAGLN
COL4A1
COL4A2
COL18A1
ACTA2
MCAM
HIGD1B
COX4I2
PHLDA1
HES4
CSRP2
ADIRF
HSPA6
LYZ

Expression
2
1
0
-1
-2

d

Tumor   Lymph

MMP1
CXCL6
TFF3
CXCL5
CXCL14
CXCL1
IER3
WNT5A
MMP3
ADM
IGHG4
IFI6
MMP7
CRABP2
LY6E
IFI27
IGFL2
CALB2
OGN
MGP
IGFBP2
SFRP2
CXCL12
DPT
ADH1B
CCDC80

Expression
2
1
0
-1

e

Mast cell
T cell
B cell
Myeloid cell
Fibro.Tumor
Fibro.Normal
Endothelial cell
Epithelial cell

LY6E MMP7 IFI6 IFI27 CCDC80 DPT SFRP2

Percentage
low
high

Normalized expression
high
low

Supplementary Fig. S7 DEGs and identified markers of fibroblasts during SBA progression. a Dot plots presenting the normalized expression levels of corresponding markers of each fibroblast cluster in the 10× dataset. **b** Representative IHC staining of ACTA2 in adjacent normal and primary tumor tissues (original magnification 100×). The rows of the paired normal and tumor samples are from the same patients, and three individual patients are listed. Scale bar, 100 μm. **c** Heatmap showing DEGs of fibroblasts from tumor tissues compared with normal tissues in the 10× dataset. **d** Heatmap showing DEGs of fibroblasts from primary tumor tissues compared with lymph node tumor tissues in the STRT dataset. **e** Dot plots presenting the normalized expression levels of 8 increasingly expressing markers during tumor progression in the STRT dataset.

Supplementary Fig. S8

Supplementary Fig. S8 Different expression interactions of fibroblasts and epithelial cells during SBA progression. a Scatterplot showing differential expression interaction pairs of epithelial cells and fibroblasts during tumor progression. **b** Dot plot showing continuously differential expression of interaction pairs in different malignant cell and fibroblast pairs. **c** Venn plot showing interaction pairs of different types of the fibroblasts and malignant cells between the STRT and 10× datasets. **d** GSEA result of the collagen and reactome collagen formation gene sets enriched in fibroblasts from tumor tissues compared with normal tissues.

Supplementary Fig. S9

a

b

c
DEGs of Treg cells in tumor and normal tissues

DEGs of cytotoxic T cells in tumor and normal tissues

d

e
DEGs of exhausted-high and exhausted-low T cells

f

g
CXCL13

h
Malignant cells and T cells - GC and CRC
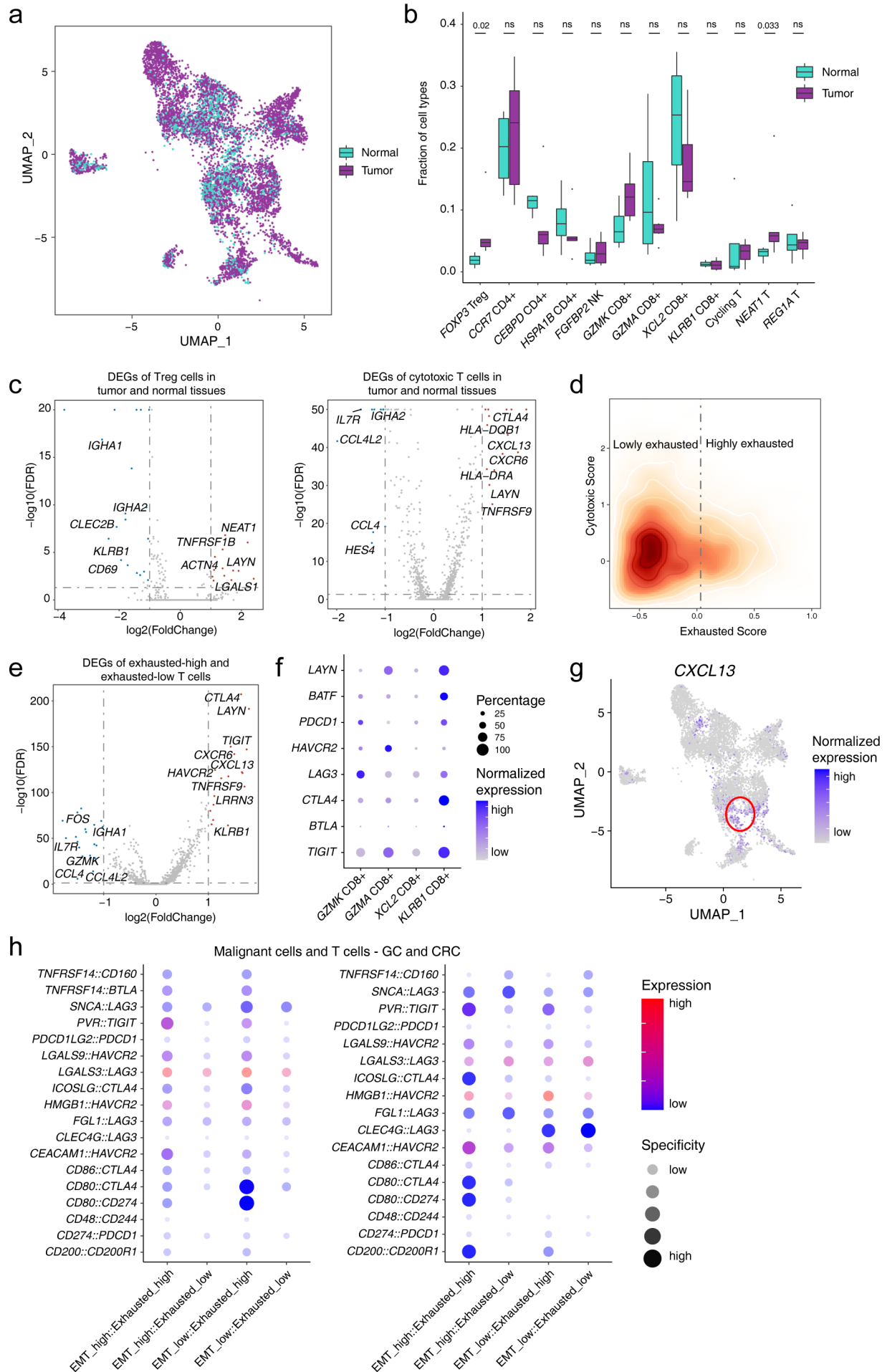
Supplementary Fig. S9 Analysis of T cell exhaustion in SBA. a UMAP plot exhibiting T cells from tumor and normal tissues in the 10× datasets. **b** Boxplot presenting cell percentage of different T cell subclusters between tumor and normal tissues in the 10× dataset, with "ns" representing no significance and P values calculated by t-test. **c** Volcano plot exhibiting DEGs of T cells between tumor and normal tissues in Treg cells and cytotoxic cells. **d** Density plot showing the separation of T cells with low and high T cell exhausted signatures in the 10× dataset. **e** Volcano plot showing DEGs of T cells with high and low exhausted signatures. **f** Dot plots presenting the normalized expression level of exhausted markers of CD8$^+$ T cells in the 10× dataset. **g** UMAP plot showing the expression of *CXCL13* in T cells from the 10× dataset. **h** Dot plot showing expression of interaction pairs in the malignant cells with low and high EMT signatures and T cell with low and high exhausted signatures of GC and CRC in responding 10× datasets. The used datasets are subsets of the GSE134520 and GSE132465 datasets in the GEO database.

Supplementary Fig. S10

a GC epithelial cells — Normal epithelial, Malignant; Normalized expression high/low; *EPCAM*; *KRT19*

b CRC epithelial cells — Normal epithelial, Malignant; Normalized expression high/low; *EPCAM*; *KRT19*

c Malignant cells correlation of 10X data

Spearman Correlation high/low

	GC	Duo.Malignant	Jej.Malignant	CRC
GC	1.00	0.69	0.59	0.60
Duo.Malignant	0.69	1.00	0.57	0.62
Jej.Malignant	0.59	0.57	1.00	0.59
CRC	0.60	0.62	0.59	1.00

d Genes: OLFM4, H3F3B, LYZ, ADIRF, HSP90AB1, PHLDA2, DNAJB1, HSPA5, RPLP2, RPL36A, RPL27, UBA52, SLIRP, SNRPF, TFF3, UBE2C, CRIP1 — GC.Malignant, Duo.Malignant, Jej.Malignant, CRC.Malignant; Expression 2, 1, 0, -1, -2

e GC.Malignant, Duo.Malignant, Jej.Malignant, CRC.Malignant

f Up-regulated genes in SBA / Up-regulated genes in CRC: 48 (83%), (27%) 10 (17%), (73%) 27

Down-regulated genes in SBA / Down-regulated genes in CRC: 86 (86%), (26%) 14 (14%), (74%) 40

g CRC, SBA; p.value=2.1e−12; Expression of the hallmark EMT geneset

Supplementary Fig. S10 Analyses of gastrointestinal cancer datasets and comparison of gastrointestinal cancers. a, b Identification of malignant cells and normal epithelial cells and the expression of *EPCAM* and *KRT19* in the GC and CRC 10× datasets. The used datasets are subsets of the GSE134520 and GSE132465 datasets in the GEO database. **c** Heatmap showing Spearman correlation of malignant cells from different intestinal regions of SBA and other gastrointestinal cancers. **d** Heatmap showing the expression of DEGs of GC and CRC in all the malignant cells. **e** UMAP plot representing the integration of gastrointestinal cancers. **f** Venn plot showing interactions of markers of the malignant cells and normal epithelial cells between SBA and CRC in responding 10× datasets. **g** Violin plot showing expression of the hallmark EMT geneset between CRC and SBA in responding 10× datasets. The *P* value was calculated by *t*-test.

Supplementary Fig. S11
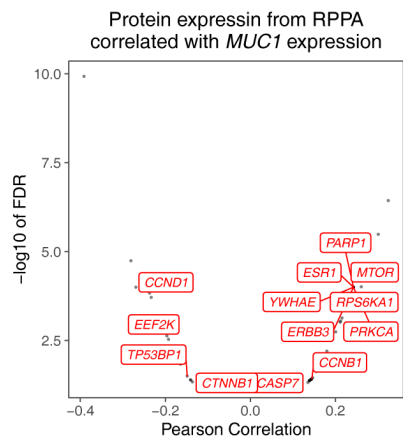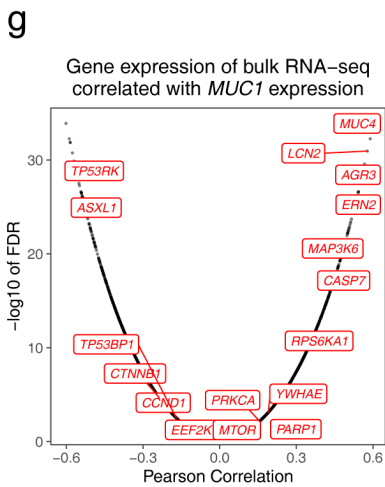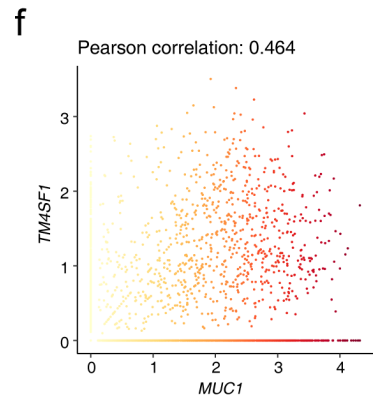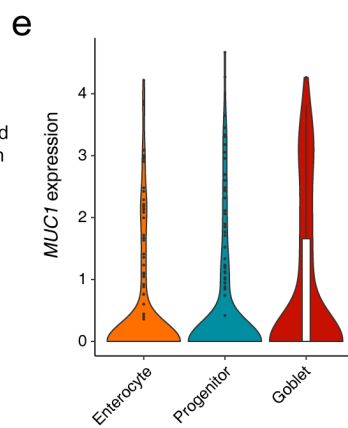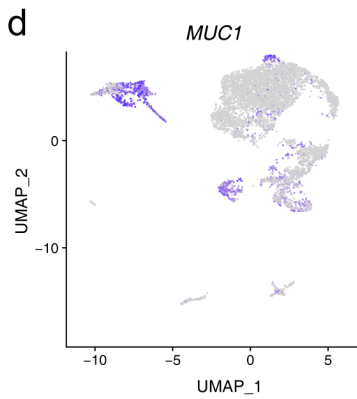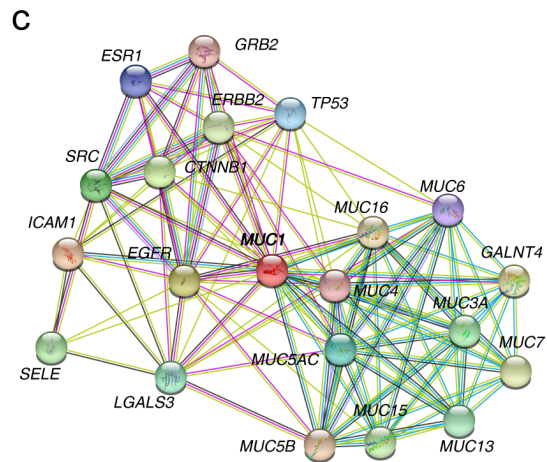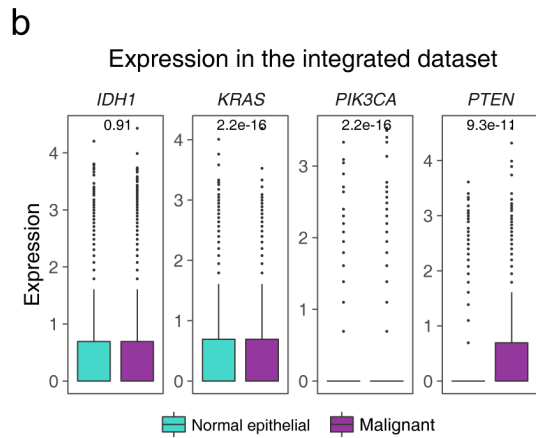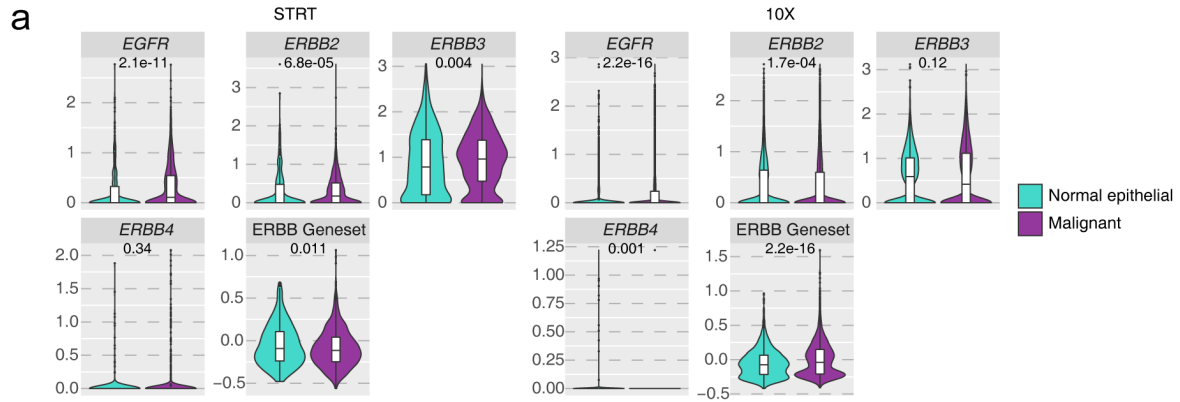
Supplementary Fig. S11 Expression of potential targets and *MUC1* as a malignant marker. a Violin plot exhibiting expression levels of key genes in the ERBB pathway and the ERBB family gene set score in the malignant cells and normal epithelial cells separately in the STRT and 10× datasets, with "ns" representing no significance and *P* values calculated by *t*-test. **b** Boxplot showing the expression of frequently-mutated genes of SBA in the integrated dataset, with "ns" representing no significance and *P* values calculated by *t*-test. **c** The protein-protein interaction network with the hub gene *MUC1* provided by the STRING database (version 11.5). **d** UMAP plot showing the expression of *MUC1* in normal epithelial cells of the 10× dataset. **e** Violin plot showing the expression of *MUC1* in the normal epithelial cells of the STRT dataset. **f** Scatterplot showing the Pearson correlation of *TM4SF1* with *MUC1* in the 10× dataset. **g** The first part of the scatterplot showing the Pearson correlation of the expression of *MUC1* and all the other genes according to the bulk RNA-seq of CRC. The second part of the scatterplot showing the Pearson correlation of the expression of *MUC1* and protein expression levels according to the RPPA data of CRC.
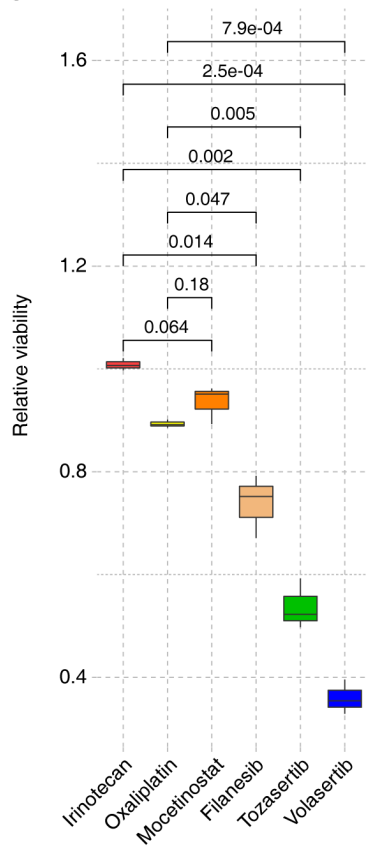
Supplementary Fig. S12

a

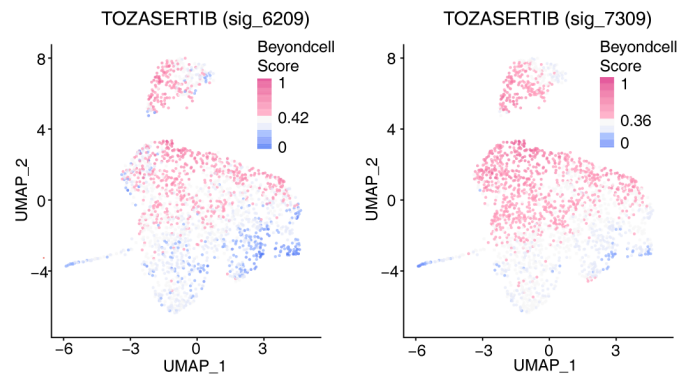| Datasets	Numbers of drugs (remained)	Published date
CTRP	538	2015
PRISM	1,492	2019
GDSC1	318	2016
GDSC2	173	2019

b

Expression in the integrated dataset
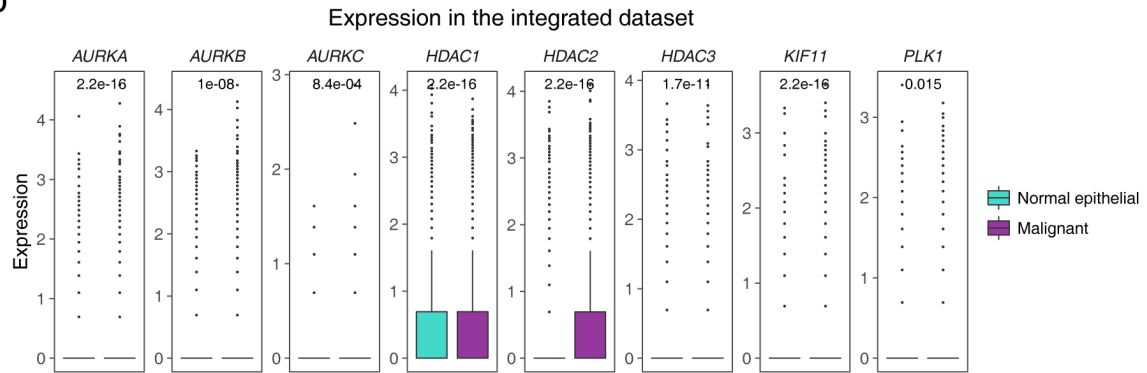
c

d

Supplementary Fig. S12 Analysis of predicted specific drugs and verification in SBA. a Information of used databases for predicting drugs. **b** Boxplot showing the expression of target genes of the four selected drugs in the integrated dataset, with "ns" representing no significance and *P* values calculated by *t*-test. **c** Boxplot showing the relative viability of selected drugs in the HUTU-80 cell line at the concentration of 0.1 μM, with "ns" representing no significance and *P* values calculated by *t*-test. **d** Evaluated drug sensitivity of tozasertib in single cells from the STRT dataset based on Beyondcell.