Cancer Cell, Volume *33*

Supplemental Information

Genomic and Epigenomic Profiling of High-Risk Intestinal Metaplasia Reveals Molecular Determinants of Progression to Gastric Cancer

Kie Kyon Huang, Kalpana Ramnarayanan, Feng Zhu, Supriya Srivastava, Chang Xu, Angie Lay Keng Tan, Minghui Lee, Suting Tay, Kakoli Das, Manjie Xing, Aliya Fatehullah, Syed Muhammad Fahmy Alkaff, Tony Kiat Hon Lim, Jonathan Lee, Khek Yu Ho, Steven George Rozen, Bin Tean Teh, Nick Barker, Chung King Chia, Christopher Khor, Choon Jin Ooi, Kwong Ming Fock, Jimmy So, Wee Chian Lim, Khoon Lin Ling, Tiing Leong Ang, Andrew Wong, Jaideepraj Rao, Andrea Rajnakova, Lee Guan Lim, Wai Ming Yap, Ming Teh, Khay Guan Yeoh, and Patrick Tan

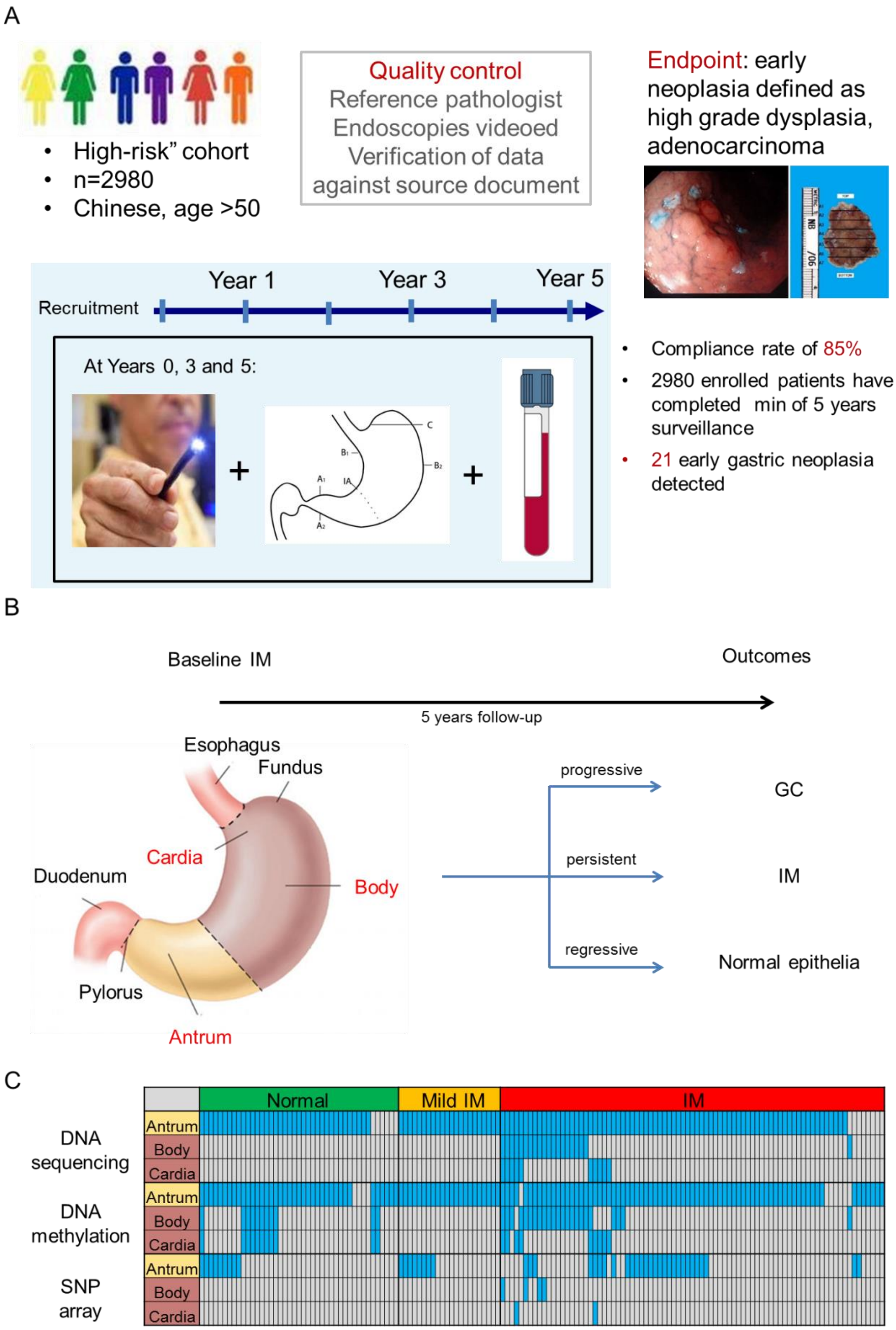
A
High-risk" cohort
n=2980
Chinese, age >50
Quality control
Reference pathologist
Endoscopies videoed
Verification of data against source document
Endpoint: early neoplasia defined as high grade dysplasia, adenocarcinoma
Recruitment
Year 1
Year 3
Year 5
At Years 0, 3 and 5:
+
+
C
B₁
B₂
A₁
IA
A₂
Compliance rate of 85%
2980 enrolled patients have completed min of 5 years surveillance
21 early gastric neoplasia detected
B
Baseline IM
Outcomes
5 years follow-up
Esophagus
Fundus
Cardia
Duodenum
Body
Pylorus
Antrum
progressive
persistent
regressive
GC
IM
Normal epithelia
C
Normal
Mild IM
IM
DNA sequencing
Antrum
Body
Cardia
DNA methylation
Antrum
Body
Cardia
SNP array
Antrum
Body
Cardia

Figure S1. Overview of Gastric Cancer Epidemiology Program (GCEP), Related to Figure 1
(A) Patients presenting at gastrointestinal endoscopy clinics were recruited, and quality control metrics (shown) implemented to ensure accurate diagnoses. Samples were collected at years 0, 3, and 5 (multiple samples per patient). Development of early gastric neoplasia was defined as an end-point.
(B) Samples were collected from gastric regions in red (cardia, body, antrum). Patients were longitudinally tracked to monitor disease outcomes, including progression to GC (progressive), no major changes in IM status (persistent), or histologic regression (regressive).
(C) Samples (blue boxes) from the antrum (yellow) or body/cardia (brown) were subjected to targeted DNA sequencing (776 genes), DNA methylation arrays (epigenetics), and SNP arrays (copy number alterations). For targeted sequencing and SNP arrays, GCEP samples (gastric normal, mild IM, IMs) were compared against patient matched germline blood DNAs.

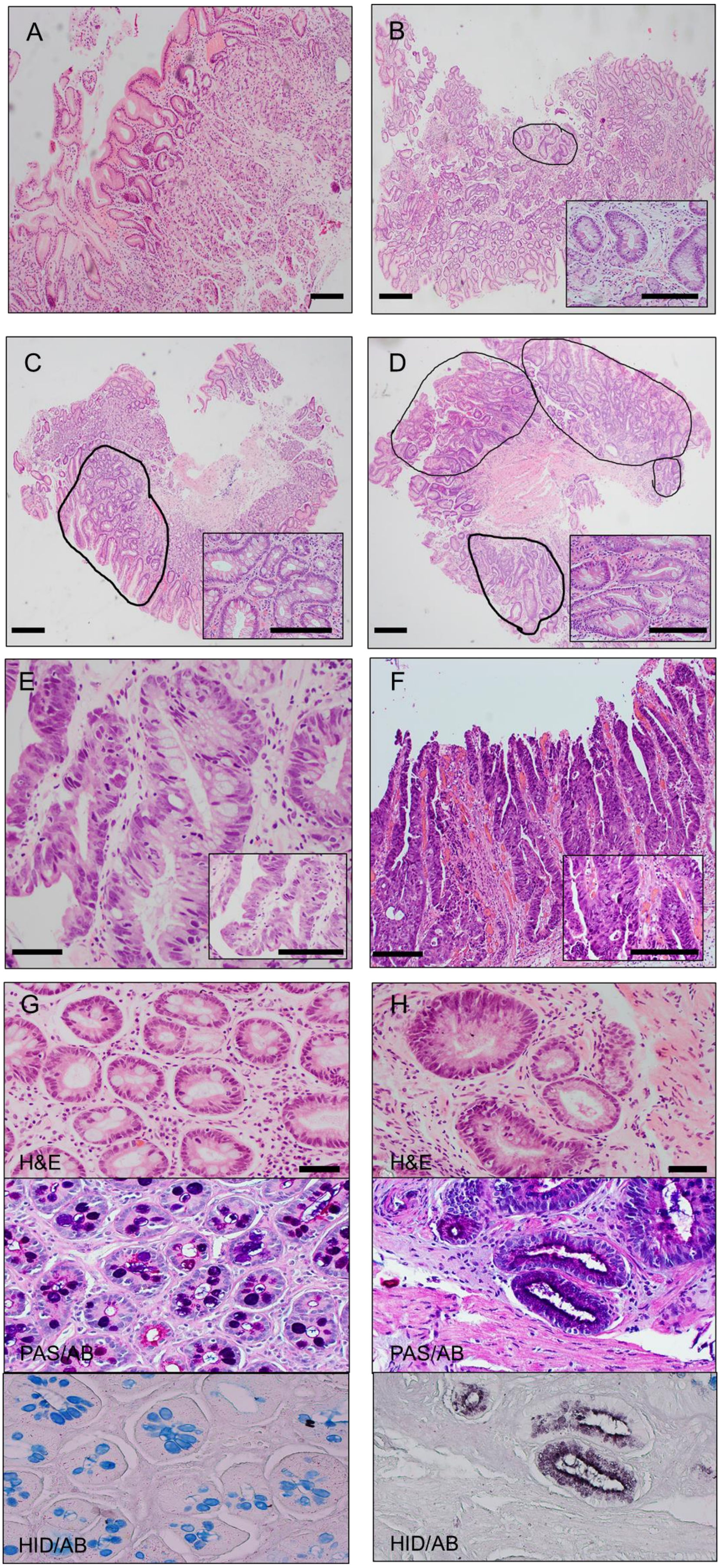
A
B
C
D
E
F
G
H
H&E
H&E
PAS/AB
PAS/AB
HID/AB
HID/AB

Figure S2. Histological Categories of Intestinal Metaplasia, Related to Figure 1

(A-D) Images of normal gastric mucosa (A) and mild (B), moderate (C) and marked (D) IM. Scale bars, 250 µm.

(E) Image showing gastric high-grade dysplasia (HGD). The inset shows dysplastic and distorted gastric glands with vesicular nuclei, prominent nucleoli and dysmorphic goblet cells. In GCEP, HGD cases were treated as early gastric neoplasias (EGNs). Scale bars, 250 µm.

(F) Image showing gastric adenocarcinoma (intestinal type). The inset shows malignant glands with vesicular nuclei, prominent nucleoli, high nucleo-cytoplasmic ratio cells and multiple mitotic figures. Scale bars, 250 µm.

(G) Image shows a representative case of complete IM (type I) by H&E, PAS/AB and HID/AB staining. Histologically the metaplastic epithelium resembles small intestine on H&E. PAS/AB stains the acidic mucins in the goblet cells as purple while HID/AB stains the acidic mucins in goblet cells as blue. Scale bar, 100 µm.

(H) Image shows a representative case of incomplete IM (type III) by H&E, PAS/AB and HID/AB staining. Histologically the metaplastic epithelium resembles large intestine and appears irregular on H&E. PAS/AB stains the mucins in the goblet cells and columnar mucosa as purple while HID/AB stains the acidic (sulfomucins) mucins in goblet cells and columnar mucosa as brownish black in color. Scale bar, 100 µm.

Table S1. Clinical parameters of GCEP cohort used in this study, Related to Figure 1

	Normal Mucosa	**Mild IM (<30%)**	**IM (≥30% cellularity)**	**p value (IM vs Normal)**
	n=43	**n=22**	**n=83**	
Age (year), mean ±SD	62±7	60±7	62±7	0.17
Race				
Chinese	43 (100)	22 (100)	83 (100)	-
Gender (%)				
Male	22 (51)	12 (55)	42 (51)	1
Female	21 (49)	10 (45)	41 (49)	
Smoking (%)				0.048
Current/ Ex-Smoker	9 (21)	4 (18)	32 (39)	
Non-smoker	34 (79)	18 (82)	51 (61)	
Alcohol consumption (%)	5 (12)	5 (23)	20 (24)	0.1
Family history of GC in first-degree relative (%)	7 (16)	3 (14)	13 (16)	1
Hp serology positivity (%)	43 (100)	22 (100)	83 (100)	-
Chronic gastritis (%)	38 (88)	22 (100)	83 (100)	0.004
Atrophic gastritis (%)	0 (0)	0 (0)	67 (81)	-
EGN (%)	0 (0)	0 (0)	4 (5)	-
Endoscopy surveillance (months), mean ±SD	56±12	58±8	49±18	0.04

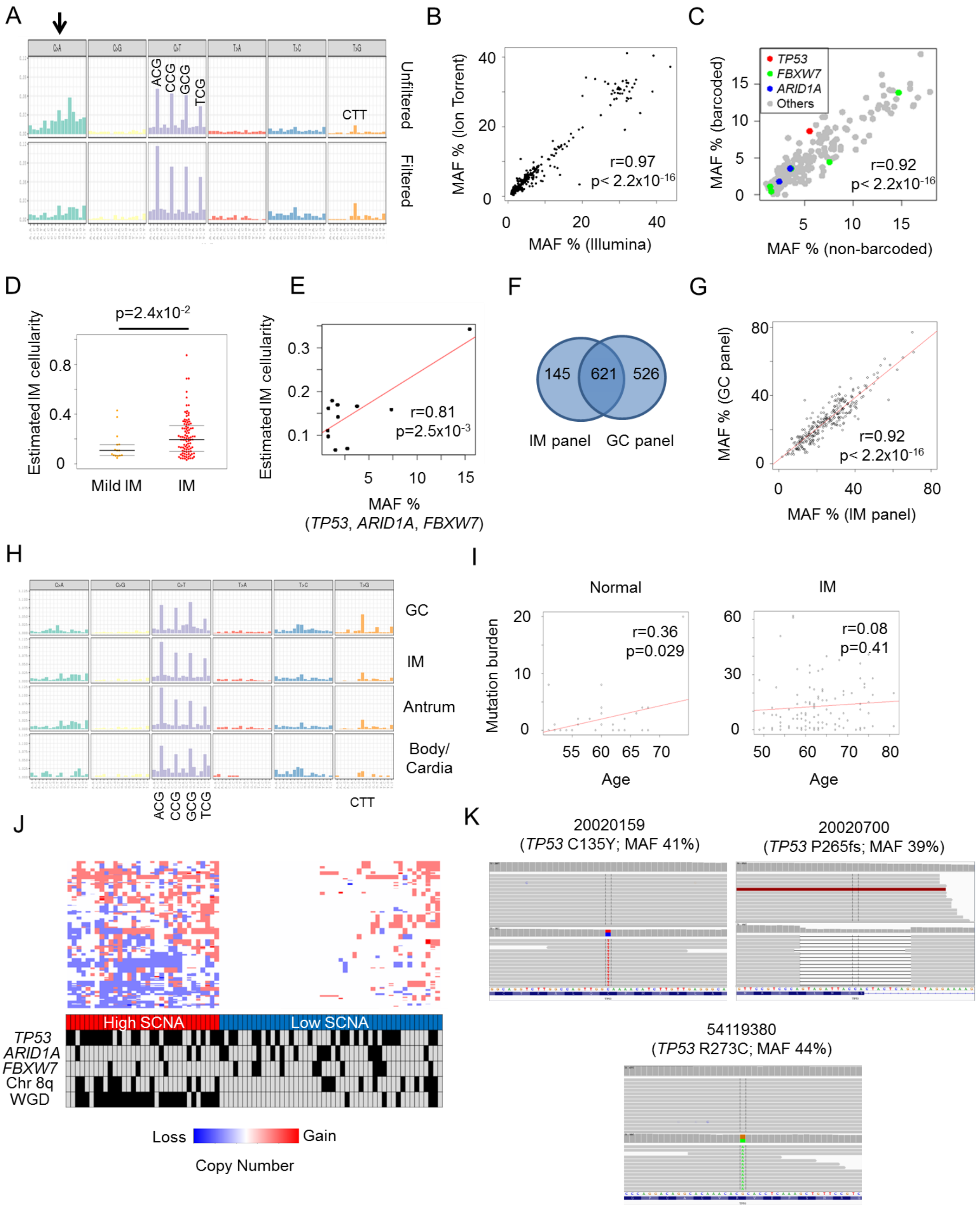
A
C>A
C>G
C>T
T>A
T>C
T>G
ACG
CCG
GCG
TCG
CTT
Unfiltered
Filtered
B
MAF % (Ion Torrent)
MAF % (Illumina)
r=0.97
p< 2.2x10-16
C
TP53
FBXW7
ARID1A
Others
MAF % (barcoded)
MAF % (non-barcoded)
r=0.92
p< 2.2x10-16
D
p=2.4x10-2
Estimated IM cellularity
Mild IM
IM
E
Estimated IM cellularity
r=0.81
p=2.5x10-3
MAF %
(TP53, ARID1A, FBXW7)
F
145
621
526
IM panel
GC panel
G
MAF % (GC panel)
MAF % (IM panel)
r=0.92
p< 2.2x10-16
H
GC
IM
Antrum
Body/
Cardia
ACG
CCG
GCG
TCG
CTT
I
Normal
IM
Mutation burden
r=0.36
p=0.029
r=0.08
p=0.41
Age
Age
J
High SCNA
Low SCNA
TP53
ARID1A
FBXW7
Chr 8q
WGD
Loss
Gain
Copy Number
K
20020159
(TP53 C135Y; MAF 41%)
20020700
(TP53 P265fs; MAF 39%)
54119380
(TP53 R273C; MAF 44%)

Figure S3. Mutation Signatures, IM cellularity, and GC Profiles, Related to Figure 1

(A) Mutational spectra of unfiltered (top) and filtered (bottom) datasets. Common patterns of base changes in GC (CpG >TpG and CTT > CGT) are indicated. Mutational spectra (C/G>A/T) associated with OxoG sequencing artifacts are indicated with the arrow.

(B) Correlation between MAFs predicted with Illumina and Ion Torrent sequencing platforms. Pearson correlation coefficients (r) and statistical significance (p) are shown.

(C) Correlation between MAFs predicted with and without NGS molecular barcodes. Pearson correlation coefficients (r) and statistical significance (p) are shown. Correlations remain significant when considering only mutations in highlighted genes (*TP53*, *ARID1A* and *FBXW7*; r=0.93, p value=1.0×10^{-4}) or *FBXW7* alone (r=0.98, p value=4.4×10^{-3}). All *FBXW7*-mutated samples for which clinical material was available were resequenced using molecular barcodes (5 cases).

(D) Beeswarm plots of estimated IM cellularity (from MAF) in mild IM (left) and marked/moderate IM (right) tissues. Thick bars indicate medians and thin bars indicate first and third quartiles. P values were calculated using Welch t-test.

(E) Correlation between IM cellularity estimated from DNA mutations and MAFs of *TP53*, *ARID1A* or *FBXW7*. The Pearson correlation coefficient (r) and the significance of the Pearson correlation test (p) is indicated.

(F) Overlap between targeted genes in IM (n=766 genes) and GC panels (n=1147 genes).

(G) Correlation of MAFs for 6 GC-normal pairs sequenced on both IM and GC panels. Pearson correlation coefficients (r) and statistical significance (p) are shown.

(H) Common patterns of base changes in GC (CpG >TpG and CTT > CGT). Mutational patterns in antrum and body/cardia IM are also plotted.

(I) Correlation between mutation burden and age in normal gastric samples and IM samples.

(J) Copy number and mutation landscape in GCs. Sequencing of 80 GC/normal pairs was performed using the GC panel (see F and G). *TP53* and *ARID1A* are frequently mutated and often in an exclusive fashion. sCNAs were estimated using the GATK ACNV workflow and WGDs were estimated using ABSOLUTE. Unsupervised hierarchical clustering grouped GC samples into two clusters (high sCNA and low sCNA), distinguished by the levels of copy number alterations.

(K) Integrated Genome Viewer (IGV) view of predicted *TP53* somatic mutations in GC sequenced using the GCEP panel. Upper panel shows the aligned reads from the matched normal and bottom panel shows the aligned reads from tumor samples.

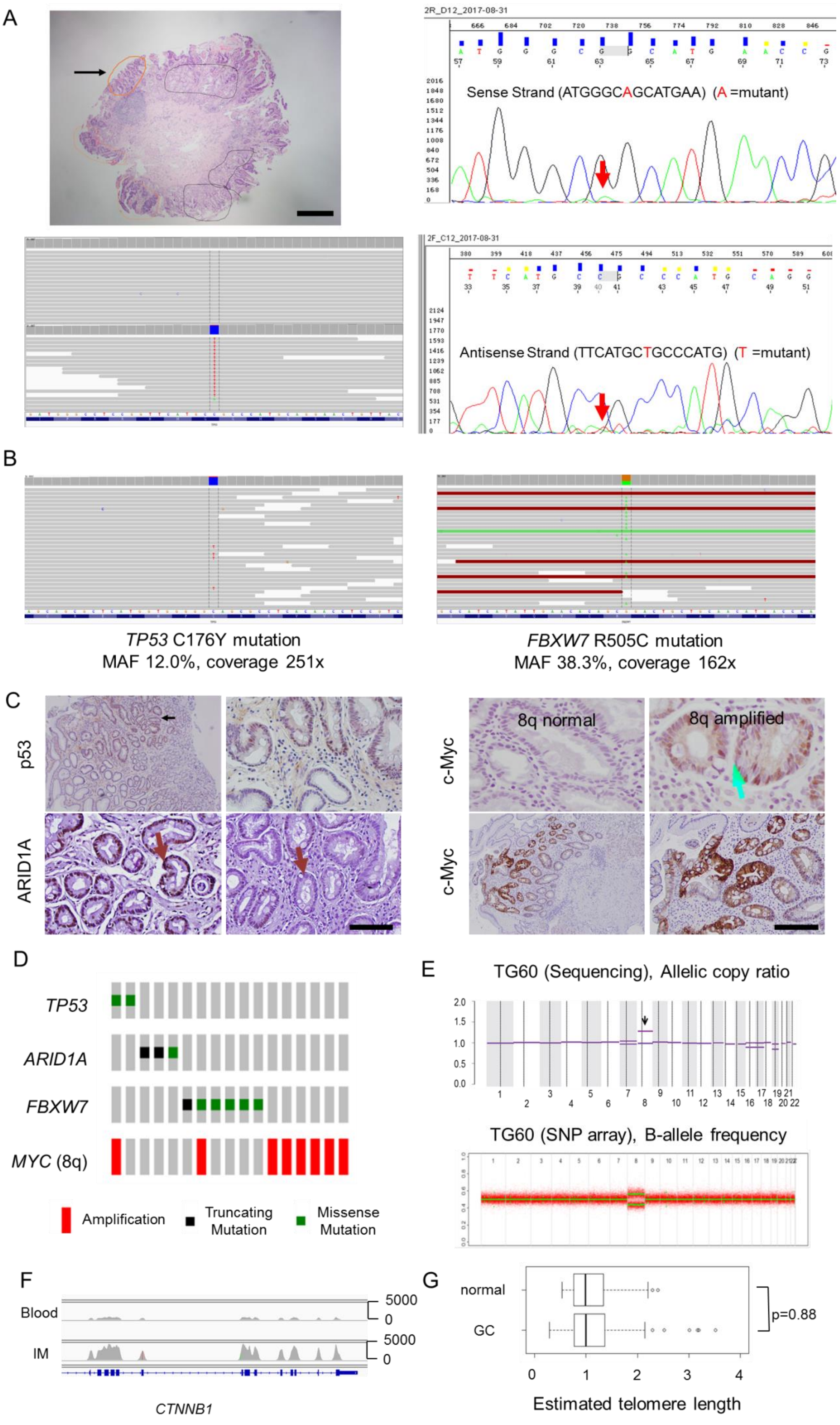
A
2R_D12_2017-08-31
Sense Strand (ATGGGCAGCATGAA) (A =mutant)
2F_C12_2017-08-31
Antisense Strand (TTCATGCTGCCCATG) (T =mutant)
B
TP53 C176Y mutation
MAF 12.0%, coverage 251x
FBXW7 R505C mutation
MAF 38.3%, coverage 162x
C
p53
ARID1A
c-Myc
8q normal
8q amplified
c-Myc
D
TP53
ARID1A
FBXW7
MYC (8q)
Amplification
Truncating Mutation
Missense Mutation
E
TG60 (Sequencing), Allelic copy ratio
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
TG60 (SNP array), B-allele frequency
F
Blood
IM
5000
0
5000
0
CTNNB1
G
normal
GC
p=0.88
0 1 2 3 4
Estimated telomere length

Figure S4. Genomic Alterations in IM, Related to Figure 1 and Figure 2

(A) Hematoxylin and eosin (H&E) stain of an IM section (Sample TG21) at the top left. IM regions are marked with circles and the IM region isolated by LCM is marked with an arrow. IGV plots show a somatic *TP53* mutation found at a MAF of 5.5% found in TG21. Bottom panel of IGV plot shows mapped reads in TG21 while the upper panel shows mapped reads in matched blood samples. Bidirectional Sanger sequencing shows the same *TP53* mutation in both DNA strands of the LCM-captured IM section. Scale bars, 250 μm.

(B) IGV plots show *TP53* (right) and *FBXW7* (left) variants in LCM-purified IM cells. Both variants (*TP53* C176Y and *FBXW7* R505C) are likely somatic mutations as they correspond to known oncogenic mutation hotspots in the COSMIC database.

(C) (Top left) Nuclear expression of p53 in IM glands (black arrow) and no expression in normal foveolar gastric glands (marked by *). Right image is a close-up image showing higher magnification of the IM glands with nuclear expression of p53. (Bottom left) AR1D1A wild-type expression in *ARID1A*-wild-type IM glands (red arrows) on left; while *ARID1A*-mutated IMs show loss of expression of AR1D1A in IM glands on right (red arrows) (Top right) c-MYC protein expression in IM samples by immunohistochemistry. Figure shows no c-MYC expression in an IM sample with normal chromosome 8 (left) and strong nuclear expression of c-MYC in an IM sample with 8q gain (right). (Bottom right) Image showing both strong cytoplasmic and nuclear expression of c-MYC in IM glands. The right image shows a higher magnification. Scale bars, 100 μm.

(D) Mutational co-occurrence of *TP53*, *ARID1A*, *FBXW7* and *MYC* alterations in IM. Each row represents a gene and each column represents an IM sample. Red bars indicate gene amplifications, green squares are missense mutations and black squares are truncating mutations.

(E) Example of a copy number alteration detected by sequencing data and confirmed using SNP arrays. Top figure shows the allelic copy ratio from sample TG60 predicted from sequencing data. Arrow indicates the region of predicted copy number alteration (chromosome 8). Bottom figure shows the confirmation of chromosome 8 amplification in the same sample using SNP arrays.

(F) *CTNNB1* amplification in one IM sample. Depths of coverage (0 to 5000x) at the *CTNNB1* locus were visualized using IGV. Grey curves indicate read coverage for IM and blood samples. Amplification is observed at *CTNNB1* exons as targeted capture was performed.

(G) Estimated telomere lengths in GC and adjacent normal samples (normalized to median telomere length in normal gastric tissues). Each box is the interquartile range (IQR) and the line is the median. Whiskers are extended to within 1.5 IQR of the upper and lower quartiles. Data points that fall outside this range are displayed independently. P values were calculated using t-tests.

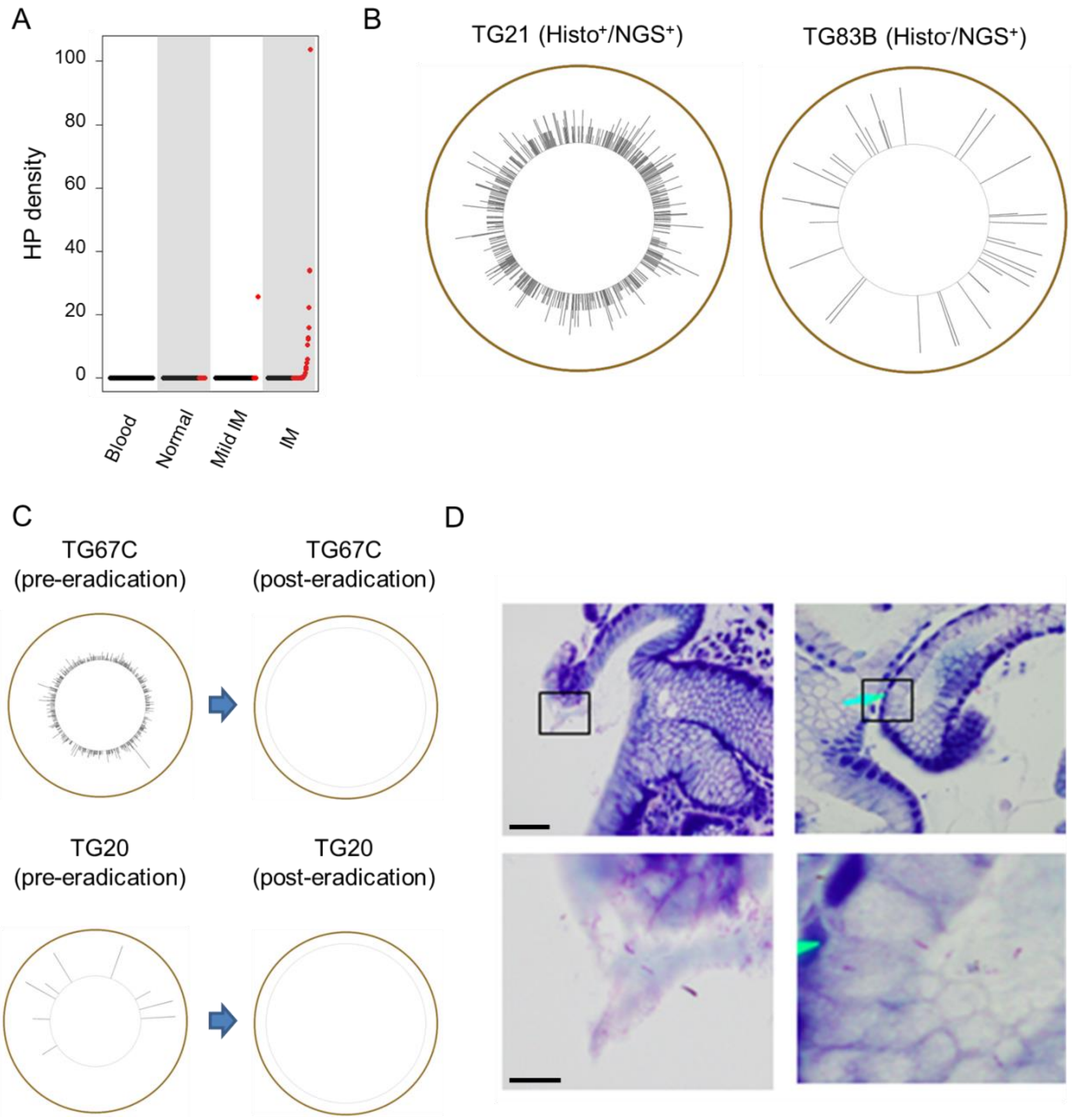

Figure S5. Sequencing of *H. pylori*, Related to Figure 3

(A) Hp sequence read densities (calculated as Hp sequences per million sequencing reads) in blood, normal gastric, mild IM, and IM. Red circles indicate samples with at least one Hp sequence.

(B) Examples of Hp coverage in Histo$^+$/NGS$^+$ IM samples (left; TG21) and Histo$^-$/NGS$^+$ IM sample (right; TG83B). Circos plot of the Hp genome with depth of coverage per 1kbp window. Maximum coverage are 7x for TG21 and 2x for TG83B.

(C) Detection of Hp reads in pre- (left) and post- (right) eradicated IM samples (2 cases are shown). Circos plot of the Hp genome with depth of coverage per 1kbp window. Maximum coverage is 17x (pre-eradication) for TG67C and 2x for TG20 (pre-eradication). No Hp sequences are found in both post-eradicated samples.

(D) Photomicrographs show gastric samples with Hp sequence reads but Hp negative by histology (Histo$^-$/NGS$^+$). Giemsa staining was used to detect low-density Hp infection. The top photographs show the tissue context, while the lower photographs are zoom-ins of the black boxes. Case identities are G80 (left) and G59 (right). Scale bars, 50 µm (top panels) and 10 µm (bottom panels).

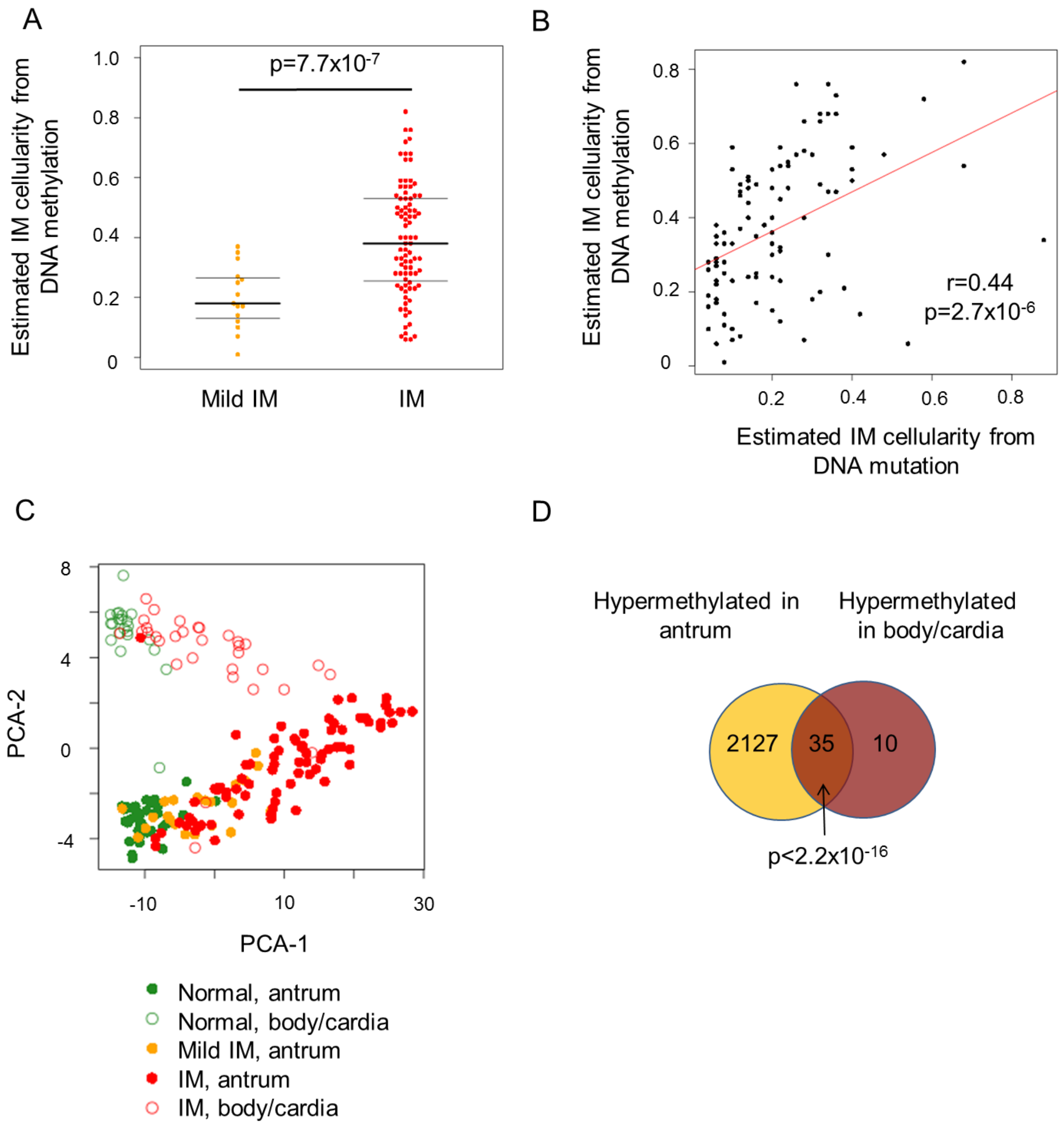

Figure S6. IM cellularity and DNA methylation changes, Related to Figure 4

(A) Beeswarm plots of estimated IM cellularity (from DNA methylation) in mild IM (left) and marked/moderate IM (right) tissues. Thick bars indicate medians and thin bars indicate first and third quartiles. P values were calculated using Welch's t-test.

(B) Correlation between IM cellularity levels, inferred from DNA sequencing or DNA methylation.

(C) Principal component analysis (PCA) of DNA methylation profiles in normal, mild IM and IM gastric biopsies. The 5000 most variable CpG sites were used as input. Normal samples, mild IM and IM are depicted in green, orange and red. Filled circles indicate samples from antral regions and open circles indicate samples from body/cardia regions.

(D) Hypermethylated sites in antral and body/cardia IMs compared against matched normal gastric mucosae from the same region. P values were calculated using the Fisher exact-test. A total of 454342 methylation sites were not hypermethylated in antrum or body/cardia IM compared to normal gastric tissues.

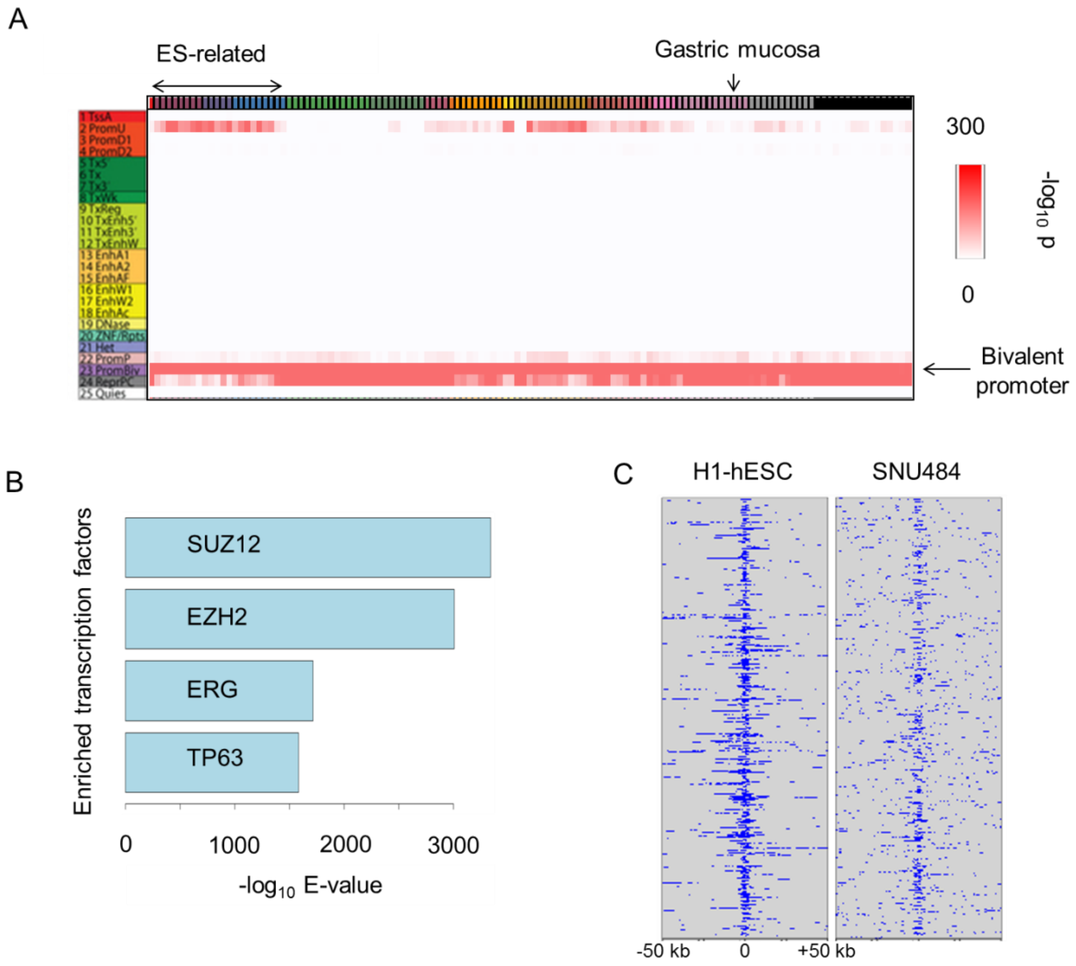

Figure S7. Chromatin Features Related to IM DMRs, Related to Figure 5

(A) Enrichment of IM DMRs at different chromatin states in 127 cell/tissue types. IM DMRs are enriched in bivalent promoter-associated chromatin states.

(B) Enrichment of transcription factor binding sites at IM DMRs, using ReMap

(C) Heatmap of H3K27me3 ChIP-seq enrichment across IM hypermethylated regions in stem cells and GC cells (SNU484). Each row represents a 100 kb window centered on a DMR.

Table S4. Association of IM Histological Subtypes with Molecular Profiles, Related to Figure 5

Complete IM = Type I IM

Incomplete IM = Type II/III IM

All IMs (Body, Cardia, Antrum)

	Complete (n=39)	Incomplete (n=77)	p value
Hp density (Hp reads per million sequence)	1.98	2.88	0.62
Mutation burden	12.62	12.68	0.98
Presence of copy number alteration			1
Yes	4 (13.3%)	9 (12.9%)	
No	30 (86.7%)	61 (87.1%)	
Telomere length	1.2	1.08	0.16
DNA methylation level	0.48	0.48	0.73
Clinical Outcome			0.1
Persistent	33 (84.6%)	60 (77.9%)	
Progression	0 (0%)	8 (10.4%)	
Regression	6 (15.4%)	9 (11.7%)	

Antral IMs only

	Complete (n=25)	Incomplete (n=57)	p value
Hp density (Hp reads per million sequence)	1.9	1.13	0.58
Mutation burden	16	13.83	0.61
Presence of copy number alteration			0.71
Yes	2 (9.1%)	8 (15.4%)	
No	20 (90.9%)	44 (84.6%)	
Telomere length	1.16	1.02	0.17
DNA methylation level	0.49	0.48	0.68
Clinical Outcome			0.22
Persistent	19 (76.0%)	42 (73.7%)	
Progression	0 (0%)	6 (10.5%)	
Regression	6 (24.0%)	9 (15.8%)	

Table S5. Multivariate Analysis of Molecular Factors and IM Outcome, Related to Figure 6

Factors related to IM Regression

	Coefficient	Odds ratio	p value
Mutation burden	-0.04	0.96	0.23
DNA methylation	-67.9	3.3×10^{-30}	0.16
Hp density	-2.2	0.11	0.63

Factors related to IM Progression

	Coefficient	Odds ratio	p value
Telomere length	-5	7.0×10^{-3}	0.147
sCNA (positive)	2.6	14	0.02
Histology (Incomplete)	18.6	1.2×10^{8}	0.995

Table S6. Association of IM Molecular Profiles with Clinical Outcome, Related to Figure 6

Antrum (n=82)

	Regressive	Persistent	Progressive
n	15	61	6
Mutation burden	**8.923**	15.719	17.2
Methylation level	**0.48**	0.485	0.487
Telomere length	0.998	1.099	**0.821**
sCNA (positive)	0	0.123	**0.6**

Body/Cardia (n=34)

	Regressive	Persistent	Progressive
n	0	32	2
Mutation burden	-	8.714	0
Methylation level	-	0.475	0.481
Telomere length	-	1.256	0.916
sCNA (positive)	-	0.103	0

Antrum/Body/Cardia (n=116); 83 subjects

	Regressive	Persistent	Progressive
N	15	93	8
Mutation burden	8.923	13.256	14.333
Methylation level	0.48	0.482	0.486
Telomere length	0.998	1.152	**0.837**
sCNA (positive)	0	0.104	**0.5**

*significantly associated features are indicated with bold text