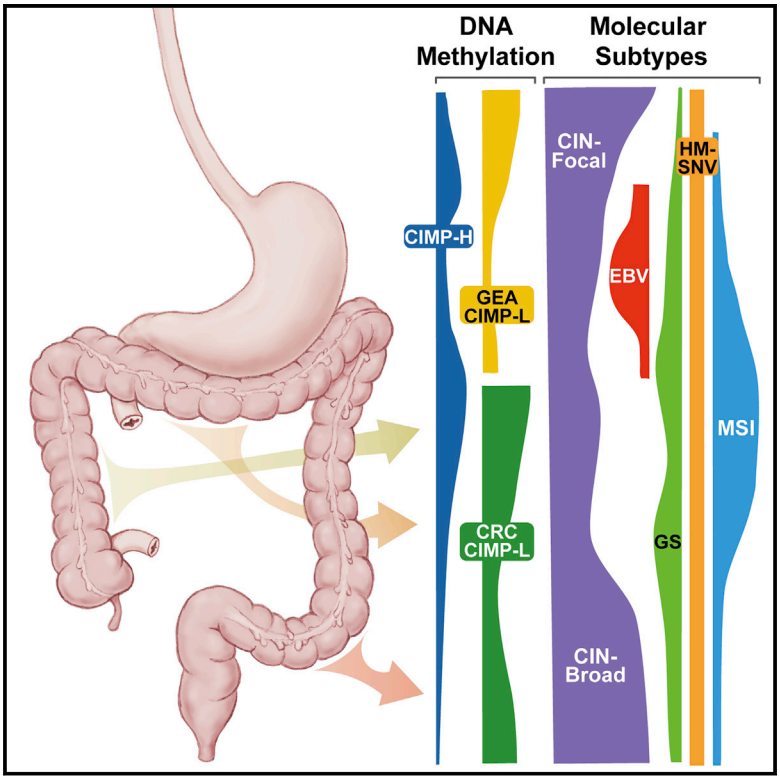
Article

Cancer Cell

# Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas

## Graphical Abstract




## Authors

Yang Liu, Nilay S. Sethi, Toshinori Hinoue, ..., Vésteinn Thorsson, Adam J. Bass, Peter W. Laird

## Correspondence

vesteinn.thorsson@systemsbiology.org (V.T.), adam_bass@dfci.harvard.edu (A.J.B.), peter.laird@vai.org (P.W.L.)

## In Brief

Liu et al. analyze 921 gastrointestinal (GI) tract adenocarcinomas and find that hypermutated tumors are enriched for insertions/deletions, upper GI tumors with chromosomal instability harbor fragmented genomes, and a group of genome-stable colorectal tumors are enriched in mutations in *SOX9* and *PCBP1*.

## Highlights

- GI adenocarcinomas comprised five molecular subtypes: EBV, MSI, HM-SNV, CIN, and GS
- Hypermutated tumors had diverse immune features varying by tissue and subtype
- CIN tumors displayed more fragmented copy-number alterations in the upper GI tract
- Genome-stable CRC subtype was enriched for recurrent mutations in *SOX9* and *PCBP1*

Liu et al., 2018, Cancer Cell *33*, 721–735
April 9, 2018 
https://doi.org/10.1016/j.ccell.2018.03.010

CellPress

Cancer Cell
Article

# Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas

Yang Liu,[1,2,22] Nilay S. Sethi,[1,2,22] Toshinori Hinoue,[3,22] Barbara G. Schneider,[4,22] Andrew D. Cherniack,[1,2] Francisco Sanchez-Vega,[5] Jose A. Seoane,[6] Farshad Farshidfar,[7] Reanne Bowlby,[8] Mirazul Islam,[1,2] Jaegil Kim,[1] Walid Chatila,[9] Rehan Akbani,[10] Rupa S. Kanchi,[10] Charles S. Rabkin,[11] Joseph E. Willis,[12] Kenneth K. Wang,[13] Shannon J. McCall,[14] Lopa Mishra,[15] Akinyemi I. Ojesina,[16,21] Susan Bullman,[2] Chandra Sekhar Pedamallu,[2] Alexander J. Lazar,[17] Ryo Sakai,[18] The Cancer Genome Atlas Research Network, Vésteinn Thorsson,[19,23,*] Adam J. Bass,[1,2,20,23,*] and Peter W. Laird[3,23,24,*]

[1]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[3]Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI 49503, USA
[4]Department of Medicine, Division of Gastroenterology, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[5]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[6]Department of Medicine, and Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[7]Department of Oncology, Cumming School of Medicine, University of Calgary, Calgarry, Canada
[8]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC V5Z 4S6, Canada
[9]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[10]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[11]Division of Cancer Epidemiology & Genetics, National Cancer Institute, Bethesda, MD 20892, USA
[12]Department of Pathology, Case Western Reserve University, Cleveland, OH 44106, USA
[13]Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN 55905, USA
[14]Department of Pathology, Duke University, Durham, NC 27710, USA
[15]Center for Translational Research, Department of Surgery, George Washington University Cancer Center, Washington, DC 20052, USA
[16]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[17]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[18]PharmiWeb Solutions, Bracknell RG12 1QB, UK
[19]Institute for Systems Biology, Seattle, WA 98109, USA
[20]Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[21]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA
[22]These authors contributed equally
[23]These authors contributed equally
[24]Lead Contact
*Correspondence: vesteinn.thorsson@systemsbiology.org (V.T.), adam_bass@dfci.harvard.edu (A.J.B.), peter.laird@vai.org (P.W.L.)
https://doi.org/10.1016/j.ccell.2018.03.010

## SUMMARY

We analyzed 921 adenocarcinomas of the esophagus, stomach, colon, and rectum to examine shared and distinguishing molecular characteristics of gastrointestinal tract adenocarcinomas (GIACs). Hypermutated tumors were distinct regardless of cancer type and comprised those enriched for insertions/deletions, representing microsatellite instability cases with epigenetic silencing of *MLH1* in the context of CpG island methylator phenotype, plus tumors with elevated single-nucleotide variants associated with mutations in *POLE*. Tumors with chromosomal instability were diverse, with gastroesophageal adenocarcinomas harboring fragmented genomes associated with genomic doubling and distinct mutational signatures. We identified a group of tumors in the colon and rectum lacking hypermutation and aneuploidy termed genome stable and enriched in DNA hypermethylation and mutations in *KRAS*, *SOX9*, and *PCBP1*.

### Significance

Adenocarcinomas of the gastrointestinal tract share not only a poor prognosis but also conserved molecular features. Hypermutated tumors display diverse immune features depending on tissue origin and molecular subtype, with implications for targeted immunotherapeutics. Upper GI tumors with chromosomal instability display a fine genome fragmentation enriched for high amplitude, focal somatic copy-number alterations associated with whole-genome doubling, specific mutational signatures, and advanced stage. We identified a genome-stable molecular subtype among colorectal cancers with an elevated frequency of recurrent mutations in *SOX9* and *PCBP1*.

## INTRODUCTION

Traditional classifications of tumors have utilized tissue of origin and histologic types. These categories have been refined with comprehensive molecular characterizations across large numbers of tumors. Adenocarcinomas of the gastrointestinal (GI) tract share similar endodermal developmental origins and exposure to common insults that promote tumor formation. We sought to evaluate molecular characteristics that distinguish GI tract adenocarcinomas (GIACs) from other cancers and to investigate the molecular features of GIACs across anatomic boundaries to provide insight into the pathogenesis of these deadly malignancies.

Approximately 1.4 million people die each year worldwide from adenocarcinomas of the esophagus, stomach, colon, or rectum (Arnold et al., 2015; Torre et al., 2016). Non-surgical treatment approaches have made only modest progress over the past half-century, inspiring efforts to better understand the biological basis of these cancers as a foundation for improving prevention, screening, and therapy. Prior studies that separately evaluated GIACs of the upper (gastroesophageal) and lower (colorectal) GI tract found subgroups such as chromosomal instability (CIN), microsatellite instability (MSI), and tumors with hypermethylation phenotypes. However, systematic efforts to characterize how shared molecular processes present differently across the GI tract have not been undertaken.

## RESULTS

The Cancer Genome Atlas Network obtained fresh frozen tissues from 921 primary GIACs (79 esophageal, 383 gastric, 341 colon, and 118 rectal cancers) without prior chemotherapy or radiotherapy. All patients provided informed consent, and collections were approved by local institutional review boards. Adjacent non-malignant tissues were obtained from 76 patients. We characterized samples by SNP array profiling for somatic copy-number alterations (SCNAs), whole-exome sequencing, array-based DNA methylation profiling, mRNA sequencing, microRNA sequencing, and, for a subset of samples, reverse-phase protein array (RPPA) profiling. Key characteristics of tumor samples are summarized in Table S1.

### Shared Features of GIACs

We investigated whether GIACs share characteristic molecular features compared with other adenocarcinomas (Table S2). Joint analysis of GIACs together with adenocarcinomas from the breast (n = 1001), endometrium (506), cervix (24), bile ducts (33), lung (240), pancreas (183), prostate (381), and ovaries (503) revealed that GIACs clustered together by DNA hypermethylation profiles (Figure S1A), mRNA (Figure S1B), and RPPA (Figure S1C). These results are consistent with integrated clustering analysis across multiple platforms of 10,000 TCGA tumors, which identified GIACs as a distinct group (Hoadley et al., 2018).

Genes mutated significantly more frequently in GIACs compared with non-GI adenocarcinomas (non-GIACs) included *FBXW7*, *SMAD2*, *SOX9*, and *PCBP1* (Figure 1A; Table S3). A GIAC-focused analysis revealed that *ATM*, *PZP*, *CACNA1C*, and *FBN3* were significantly mutated genes not previously reported in TCGA studies of single cancer types (Figure S1D; Table S3). We evaluated SCNA data to identify amplifications and deletions more common in GIACs than in non-GIACs (Figures 1B and S1E; Table S4). Arm-level gain of chromosome 13q was GIAC specific (Figure S1F), noteworthy as this region containing tumor suppressor *RB1* is often deleted in non-GIACs. *CDX2* (13q12.2) and *KLF5* (13q22.1) encoding two transcription factors in this amplified region may contribute to GIAC pathogenesis. Other genes preferentially amplified in GIACs included *CDK6* (7q21.2), *GATA6* (18q11.2), *GATA4* (8p23.1), *EGFR* (7p11.2), *CD44* (11p13), *BCL2L1* (20q11.21), *FGFR1* (8p11.22), and *IGF2* (11p15.5). *APC* and *SOX9* deletions were observed preferentially in GIACs, as were frequent mutations in these genes.

GIACs displayed markedly higher frequencies of CpG island hypermethylation than did non-GIACs (Figure 1C, upper graphs). This finding is attributable in part to the high frequency of CpG island methylator phenotype (CIMP) in GIACs, but was also evident in non-CIMP tumors. The average density of somatic mutations was also higher in GIACs. Clusters of tumors with high mutation densities were observed in gastric and colorectal GIACs as well as in breast and uterine non-GIACs (Figure 1C, middle graphs). Frequent SCNAs were observed in all GIACs, especially in esophageal adenocarcinomas (EACs), and ovarian and a subset of breast non-GIACs (Figure 1C, bottom graphs).

Gene expression analysis revealed 553 genes that were differentially expressed in GIACs compared with non-GIACs, after exclusion of genes that differed among corresponding normal tissues (Figure S1G; Table S5). Supervised multivariate orthogonal partial least-squares discriminant analysis ranked 51 of these 553 genes to have significantly higher expression in GIACs. Notably, these genes include several that have roles in gastrointestinal stem cell biology (e.g., *OLFM4*, *CD44*, and *KLF4*) and genes related to the EGFR signaling pathway (Figure S1G).

We next investigated whether genes encoding 139 transcription factors that are important in GI development (Noah et al., 2011; Sherwood et al., 2009) displayed distinct gain- or loss-of-function events in GIACs compared with non-GIACs. Amplifications were considered gain-of-function (GOF) events, while deletions, epigenetic silencing, and nonsense or indel mutations were considered loss-of-function (LOF) events (Table S6). We found 33 transcription factor genes with GOF or LOF exceeding 5% in at least one GIAC tumor type (Figure 1D). *CDX2* encodes a homeobox transcription factor expressed early in endoderm development with evidence as either a lineage-survival oncogene (Salari et al., 2012) or a tumor-suppressor gene (Bonhomme et al., 2003) in colorectal cancers (CRCs), depending on context, and is also a marker of intestinal metaplasia in Barrett's esophagus (Moons et al., 2004). Interestingly, we observed *CDX2* amplification in esophageal, colon, and rectal adenocarcinomas, but LOF in gastric cancers. Although amplifications in the genomic locus containing the stem cell transcription factor *KLF5* gene were found in all GIACs, these amplifications were associated with increased stemness only in EACs based on a gene-expression signature (Malta et al., 2018) (Figure S1H).

### Molecular Subtypes within GIACs

Other studies have relied on gene expression, oncogenic pathway, or histopathological criteria for subtype delineation among GIACs (Budinska et al., 2013; Cristescu et al., 2015;

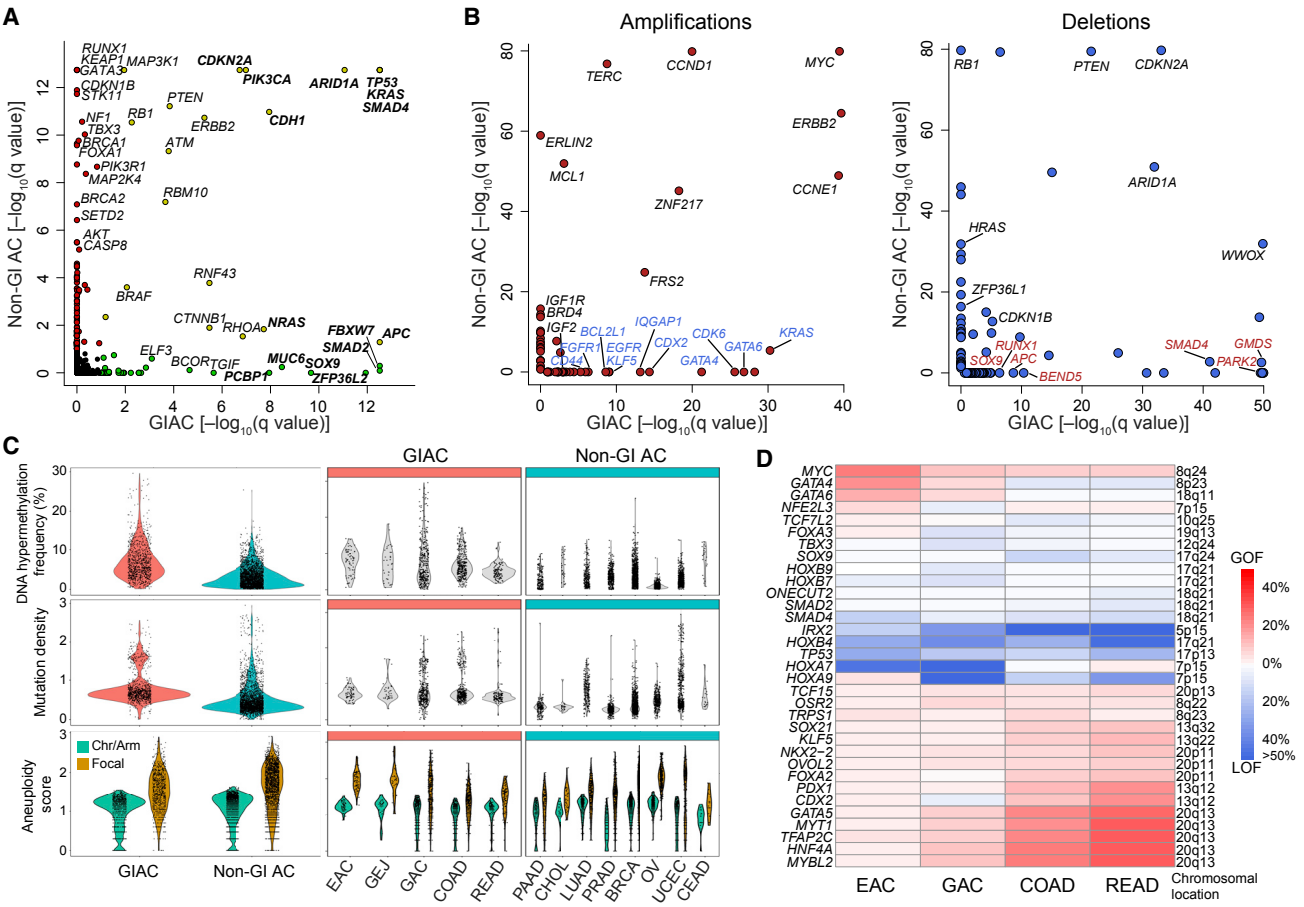



**Figure 1. Genomic Features of Gastrointestinal Adenocarcinomas**

(A) Significantly mutated genes in gastrointestinal adenocarcinomas (GIACs) indicated by green circles, significantly mutated genes identified in other adenocarcinomas (non-GIACs) indicated by red circles, and genes identified as significantly mutated in all adenocarcinomas indicated by gold circles.

(B) Genes identified as significantly recurrently amplified (left) or deleted (right) in GIACs compared with in non-GIACs.

(C) DNA hypermethylation frequency (top), mutation density (middle), and arm-level and focal copy-number events (bottom) in GIACs and non-GIACs.

(D) Percent GOF or LOF events in developmental transcription factors by cancer type. See also Figure S1 and Tables S1–S6.

Dienstmann et al., 2017; Guinney et al., 2015; Roepman et al., 2014; Tan et al., 2011). We found that unsupervised clustering of GIACs using mRNA, miRNA, and RPPA data was strongly influenced by tissue type, thus complicating defining molecular groups spanning anatomic boundaries. By contrast, evaluation of mutations, copy-number alterations, and DNA methylation patterns yielded tumor subtypes spanning tissue boundaries (Figure S1A). Our subgroups are consistent with those identified by recent genomic research across GIACs (Cancer Genome Atlas Research Network, 2012, 2014, 2017; Cristescu et al., 2015; Secrier et al., 2016; Wang et al., 2014), and rely on molecular features generally evaluable by the clinical community.

A subgroup of tumors was characterized by a high Epstein-Barr virus (EBV) burden, as previously determined via mRNA and miRNA analysis (Cancer Genome Atlas Research Network, 2014) (Figure 2A). EBV$^{+}$ tumors, found only in the stomach (n = 30), display the most extensive hypermethylation of any tumor type in TCGA (see Figure S4.6 in Cancer Genome Atlas Research Network, 2014). Hypermutated tumors (n = 157), defined by mutation density >10 per megabase (Mb) (Figure S2A) were further substratified based on the implied mechanism of replication error. MSI, arising from defective DNA mismatch repair, often yields insertion-deletion (indel) mutations in addition to single-nucleotide variants (SNVs) (Sia et al., 1997), whereas hotspot mutations in polymerase epsilon (*POLE*) are associated with SNV-predominant profiles (Cancer Genome Atlas Research Network, 2012; Palles et al., 2013; Zhou et al., 2009) (Figure 2B). Hypermutated samples with an indel density of >1 per Mb and an indel/SNV ratio >1/150 consisted of essentially all tumors with clinically defined MSI (MSI, n = 138; 54% gastroesophageal or GE; 46% colorectal or CR) (Figure S2B). All other hypermutated samples were categorized as hypermutated-SNV (HM-SNV), (n = 19 [n = 11 with *POLE* mutations]; 47% GE; 53% CR) (Figure 2B and S2B). The remaining two groups were distinguished by presence or absence of extensive SCNAs (Figure S2C). Chromosomal instability (CIN) tumors (n = 625, 48% GE; 52% CR) exhibited marked aneuploidy, defined by a clonal deletion score (CDS), (see STAR Methods) > 0.0249, which is largely determined by chromosome- and arm-level losses. By contrast, genome

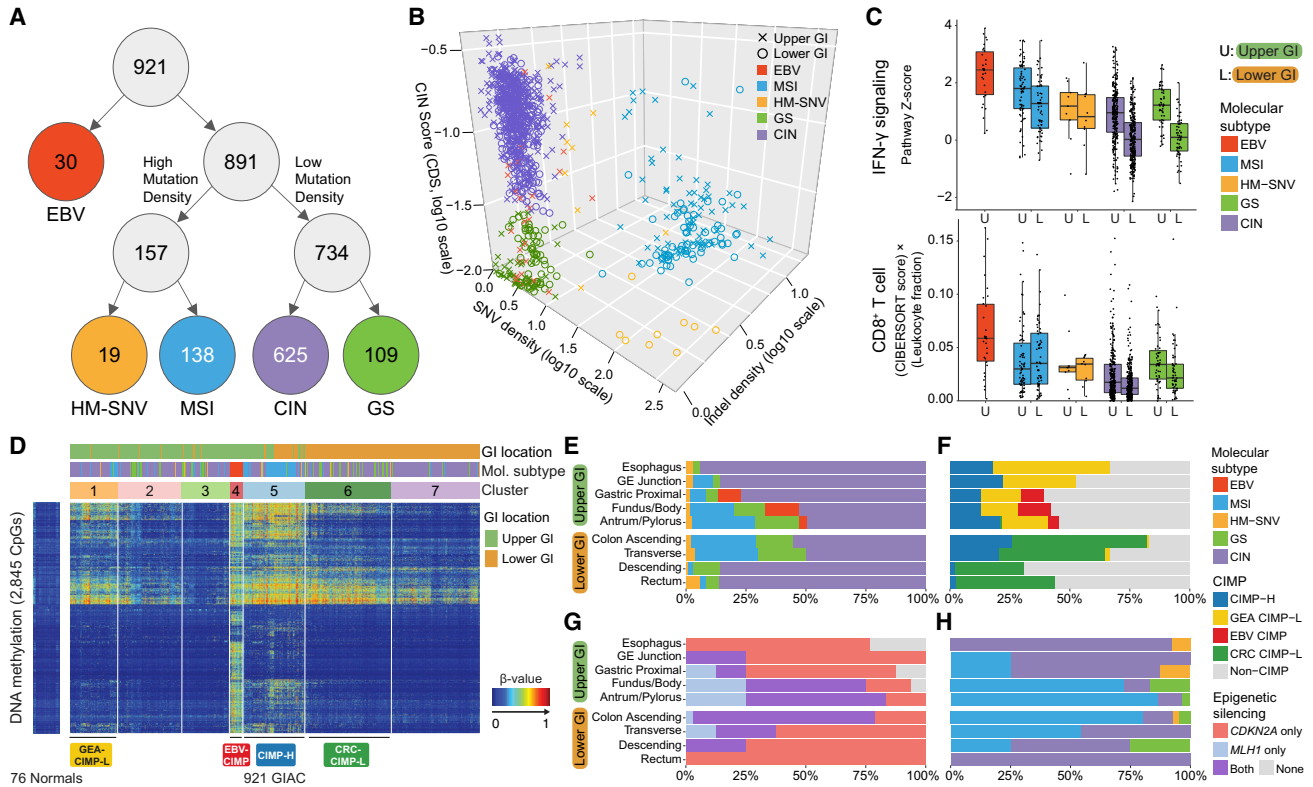



**Figure 2. Molecular Subtypes of GIACs**

(A) Flowchart of molecular subtypes: Epstein-Barr virus (EBV)-positive (red); hypermutated-single-nucleotide variant predominant (HM-SNV) (gold); microsatellite instability (MSI) (blue); chromosomal instability (CIN) (purple); and genomically stable (GS) (green).

(B) 3D plot of GIACs by SNV density, indel density, and clonal deletion score (CDS). Tumors annotated as upper GI (crosses) and lower GI (circles) and color coded by subtypes.

(C) IFN-γ pathway score (top) and $CD8^+$ T cell score (adjusted for total leukocytes; bottom) by subtypes stratified by upper versus lower GI. Horizontal bars indicate median values, boxes represent interquartile range, and whiskers indicate values within 1.5 times interquartile range.

(D) Unsupervised analysis of DNA methylation across GIACs.

(E and F) Distribution of subtypes (E) and CIMP subgroups (F) across anatomic regions.

(G and H) Distribution of *MLH1*/*CDKN2A* silencing (G) and subtypes (H) in CIMP-H tumors by anatomic region. See also Figure S2.

stable (GS) (n = 109, 47% GE; 53% CR) samples lacked such aneuploidy (Figures 2B and S2B).

We evaluated the relationship between our molecular subtypes and consensus molecular subtypes (CMSs), which have been established for CRC based primarily on gene expression (Guinney et al., 2015). We applied the CMS classification system to the lower GI tumors in our study and found a significant association between the two groupings ($p < 2.2 \times 10^{-16}$), but with noteworthy differences (Figure S2D). The CMS1-MSI immune grouping did not discriminate MSI tumors from the HM-SNV tumors (Figure S2B). A substantial fraction of GS CRCs were represented in the CMS3-metabolic subtype ($p = 1.6 \times 10^{-6}$), but the CMS system appeared to be largely unable to distinguish CIN and GS (Figure S2D).

Our molecular groupings also correlated with key immune features of GIACs (Thorsson et al., 2018) (Figures 2C and S2E). As previously reported, $EBV^+$ tumors possessed the highest gene expression scores for $CD8^+$ T cells, M1-macrophages, and interferon-γ (IFN-γ) signatures (Figures 2C and S2E) (Derks et al., 2016; Koh et al., 2017). MSI tumors showed the next greatest IFN-γ signature, consistent with reported immunogenicity of MSI tumors (Le et al., 2017). Moreover, MSI tumors displayed diverse immune signatures depending on their tissue of origin (Figures S2F and S2G); for example, checkpoint protein CD276 was significantly enriched in MSI CRC, whereas *ENTPD1* was preferentially expressed in MSI gastroesophageal adenocarcinomas (GEAs) (Figure S2G). HM-SNV also demonstrated heterogeneity in immune signature expression when comparing the upper and lower GI tract (Figure S2F). Of translational importance, an attenuation in HLA/antigen presentation (Figure S2F) and significant elevation in natural killer (NK) cell gene expression was found in HM-SNV CRC (Figure S2H), suggesting that NK cells are found in a subset of tumors and are capable of anti-tumor responses (Wagner et al., 2017). The cytotoxic activity of NK cells is finely regulated by the integration of activating and inhibitor cues (Ljunggren and Malmberg, 2007), and cells lacking MHC expression often are subjected to NK cell cytotoxicity due to the absence of inhibitory cues mediated by killer cell immunoglobulin-like receptor. These data suggest agents to enhance NK activity may be a therapeutic option for HM-SNV tumors.

Unsupervised clustering of DNA methylation data across GIACs using cancer-associated methylated sites (excluding

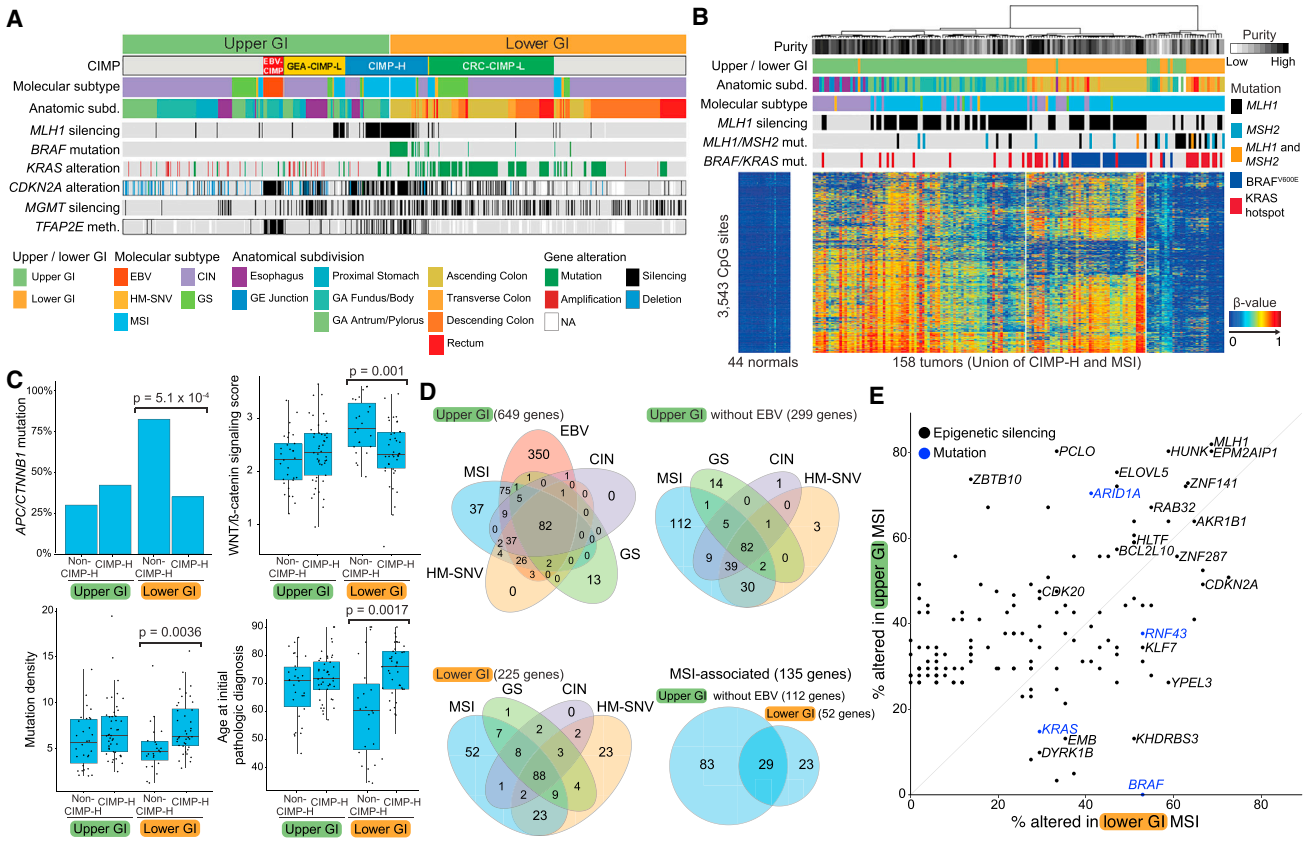



**Figure 3. Analysis of MSI and CIMP Tumors**

(A) Methylation subtypes with four CIMPs: EBV-CIMP (red), CIMP-high (blue), GEA-CIMP low (yellow), and CRC-CIMP low (green) with alterations of indicated genes.

(B) Methylation profiles of union of CIMP-high and MSI tumors with *MLH1* silencing, *KRAS, BRAF, MLH1*, and *MSH2* mutations.

(C) Features of MSI tumors stratified by upper versus lower GI and by CIMP-high status. Horizontal bars indicate median values, boxes represent interquartile range, and whiskers indicate values within 1.5 times interquartile range.

(D) Unique and overlapping epigenetically silenced genes (>25%) in upper GI (top left), upper GI tumors excluding EBV$^{+}$ (top right), lower GI (bottom, left), and MSI (bottom, right).

(E) Frequency of silencing (black) and mutation (blue) of select genes in upper GI MSI (vertical axis) versus lower GI MSI tumors (horizontal axis). See also Figure S3 and Table S7.

CpGs with tissue-specific methylation) revealed extensive CpG island methylation in the EBV$^{+}$ gastric cancers, distinguishing these tumors (Figure 2D), as previously noted by us and others (Cancer Genome Atlas Research Network, 2014; Matsusaka et al., 2011; Wang et al., 2014). The remaining tumors could be characterized as those lacking a CIMP (non-CIMP) and those displaying a low or high frequency of DNA hypermethylation (CIMP-L and CIMP-H, respectively).

Hypermutated tumors were located primarily in the central part of the GI tract, the distal stomach, and proximal colon, whereas CIN tumors were more prevalent in the anatomic extremes, the esophagus and distal colon/rectum (Figure 2E) (Budinska et al., 2013). Although CIMP-H occurred throughout the upper GI tract and proximal colon (Figure 2F), epigenetic silencing of *MLH1*, responsible for MSI, was observed primarily in the distal stomach and proximal colon (Figure 2G). Within the proximal stomach and esophagus, only 4/29 (14%) of CIMP-H tumors exhibited *MLH1* epigenetic silencing and MSI, while 23 of the 29 (79%) were microsatellite stable (MSS) and displayed the CIN phenotype (Figure 2H). In the lower GI tract, CIMP-H and MSI were largely absent in the descending colon and rectum (Figures 2E and 2H).

## Analysis of Hypermethylation and Hypermutation

MSI tumors exhibited distinct expression features independent of tissue of origin, implying common biological features of this class of tumors (Figures S3A and S3B). Most sporadic MSI cases in both colorectal and gastric cancer arise as a consequence of epigenetic silencing of *MLH1* by promoter DNA hypermethylation (Herman et al., 1998; Leung et al., 1999) in the context of CIMP-H (Weisenberger et al., 2006). MSI tumors with *MLH1* methylation were associated with *BRAF*$^{V600E}$ mutation only in the colon, not the stomach (Figure 3A). *KRAS* mutations were found primarily in CIMP-L tumors of the lower GI tract, whereas *KRAS* amplification was observed in upper GI tumors (Figure 3A). *TFAP2E* promoter methylation, which is associated with non-response to chemotherapy in CRC (Ebert et al., 2012), was found in a substantial fraction of CIMP-H tumors and in almost all EBV$^{+}$

gastric cancers (Figure 3A). CIMP-H tumors showed near-ubiquitous methylation of the tumor suppressor *CDKN2A* in gastric and colon MSI tumors (Figures 3A and 3B). However, 39% of the CIMP-H tumors lacked *MLH1* silencing and MSI, and instead included other classes of GIACs, most commonly CIN tumors in the proximal stomach/esophagus or rectum/descending colon (Figures 2H and 3B).

Given the tight associations between CIMP-H and MSI and their heterogeneity across anatomic boundaries, we studied the collection of tumors containing either of these features in more detail (Figure 3B). A portion of MSI cases lacking both *MLH1* methylation and the CIMP contained somatic mutations in *MLH1* or *MSH2,* indicating an alternative route to loss of DNA mismatch repair (Figure 3B, right side). These tumors were preferentially associated with mutations in *KRAS* rather than *BRAF.* A small number of MSI tumors (n = 8) could not be explained by genetic or epigenetic inactivation of a mismatch repair gene.

Broadly, the MSI group of CRCs harbored lower WNT signatures than did other CRCs (Figures S3C and S3D), a finding that may be attributable to a reduced reliance of CIMP-H tumors on WNT signaling. Among MSI CRCs, those arising in the context of CIMP-H have a lower percentage of *APC* mutation (28%) than those arising in either CIMP-L (78%) or non-CIMP (58%) (Fisher's exact test p = 0.0091). This finding holds true for MSS CIMP-H tumors as well, and is discussed in the GS subtype section below. CIMP-H MSI CRC showed a reduced combined frequency of either *APC* or *CTNNB1* mutations and decreased WNT gene expression signatures compared with non-CIMP-H MSI cases, and were more similar to upper GI MSI tumors in their lower reliance on WNT activation (Figure 3C). Despite the reduced frequency of *APC* and *CTNNB1* mutations, MSI CIMP-H tumors displayed overall greater mutational densities and arose at an older age of onset than did non-CIMP-H MSI cases or upper GI MSI cases (Figure 3C).

We investigated the genes silenced by promoter hypermethylation in the molecular subgroups (Figures 3D and 3E; Table S7). Pathway analysis of epigenetically silenced genes among all subgroups revealed enrichment for genes encoding DNA binding proteins and transcription factors, consistent with previous findings of enrichment for stem cell polycomb target genes (Widschwendter et al., 2007). We identified 135 genes silenced in at least 25% of the upper or lower GI MSI tumors and compared their relative frequency of silencing and frequency of several key gene mutations (Figure 3E). *HUNK*, a negative regulator of intestinal cell proliferation (Reed et al., 2015), was found to be frequently silenced in MSI tumors. Another frequently silenced gene, *ELOVL5,* lies within the locus with germline variants most significantly linked to survival of CRC patients (Phipps et al., 2016).

## Molecular Features of the CIN Subtype

The landscape of SCNAs revealed a more finely fragmented genome in GEA compared with CRC, despite an overall similar pattern of affected regions of the genome (Figure 4A). Evaluation of SCNA distribution, categorized by both focality and intensity, revealed higher prevalence of focal copy-number events within the CIN GEA population (Figures 4B and S4A). The difference between the upper and lower GI was greater for focal amplifications than for deletions (Figure 4B), primarily evident in high-amplitude focal amplifications (Figure S4A). We developed a score that captures the quantity and intensity of focal high-level amplicons (see STAR Methods). CIN tumors with a higher score were designated CIN-Focal (CIN-F), whereas those with a lower score, and therefore low-amplitude, broader amplicons, were called CIN-Broad (CIN-B) (Figure 4C). The distribution of these two classes of CIN differed between upper and lower GIACs, with CIN GEAs displaying 74% CIN-F and 26% CIN-B, and CRCs showing reversed proportions consisting of 22% CIN-F and 78% CIN-B (Figure 4C). Despite this difference between upper and lower GI tumors, the ratios of CIN-B/CIN-F did not vary anatomically within upper GI tumors or within the lower GI tract tumors (Figure S4B). Notably, in addition to the higher prevalence of CIN-F in upper GIACs, such CIN-F GEAs displayed a higher intensity in the focal-amplification score compared with their CIN-F CRC counterparts (Figure S4C). CIN-F GEA was associated with advanced tumor stage, underscoring its potential clinical significance (Figure 4D).

Although CIMP frequency displayed an anatomic gradient within the upper GI (Figure S4D), we found no correlation of CIMP class with arm-level or focal SCNAs in CIN (Figure S4E). CIN-F GEAs demonstrated significantly more whole-genome duplication (WGD) than did CIN-B GEAs, 68% versus 42% (Figures 4E and S4F), with evidence of two or more genome doublings (WGD2) in 18% of CIN-F compared with 7% of CIN-B in upper GI CIN tumors. WGD2 was associated with poor survival in GEA, independent of age and stage (Figure S4G). However, the strong association of genomic doubling and CIN-F was not observed in CRC, despite similar total rates of genome duplication (Figure S4H; 59% in lower GI and 61% in the upper GI).

CIN-F GEAs sustained significantly more frequent focal amplification of genes encoding receptor tyrosine kinases, KRAS and cell-cycle mediators (Figure 4F). In contrast, CIN-B GEAs more commonly sustained activating mutations of oncogenes (e.g., *KRAS* and *ERBB2*) than did GEA-CIN-F tumors (Figure 4F). *ERBB2* amplifications significantly co-occurred with *CCNE1* amplifications (p = 0.039) and trended toward co-occurrence with gains in chr.20q/*SRC* (p = 0.0692). Intriguingly, activating mutations in *ERBB2* co-occurred with *ERBB2* amplifications (p = 0.0087). CIN-B GEAs harbored more frequent somatic inactivation of tumor suppressors related to cell-cycle regulation (e.g., *CDKN2A*), WNT pathway activation (e.g., *APC*), and transforming growth factor β (TGF-β) regulation (e.g., *SMAD2* and *SMAD4*) than CIN GEA-F. By contrast, CIN-F GEA showed a higher frequency of *TP53* mutations (Figure 4F; 76% versus 54%) and higher rates of oncogene amplifications (Figure 4G).

Among lower GI CIN tumors, the differences in somatic mutations and copy-number alterations found in CIN-F and CIN-B tumors were modest (Figure S4I), although CIN-F did associate with poorer survival in CRC (Cox regression p = 0.0053, adjusted for stage, age, and molecular subtype). We identified amplifications including *CDX2*, *ERBB2*, and *CCND2* enriched in these tumors. Consistent with the different patterns of CIN between upper and lower GI cancers, we found that ERBB2$^{+}$ CRC not only harbor lower CIN-F scores (Figure S4J), but also fewer co-occurring genomic alterations than ERBB2$^{+}$ GEA (Figure S4K). These findings are consistent with efficacy in CRC of ERBB2 therapy without chemotherapy (Sartore-Bianchi et al., 2016),

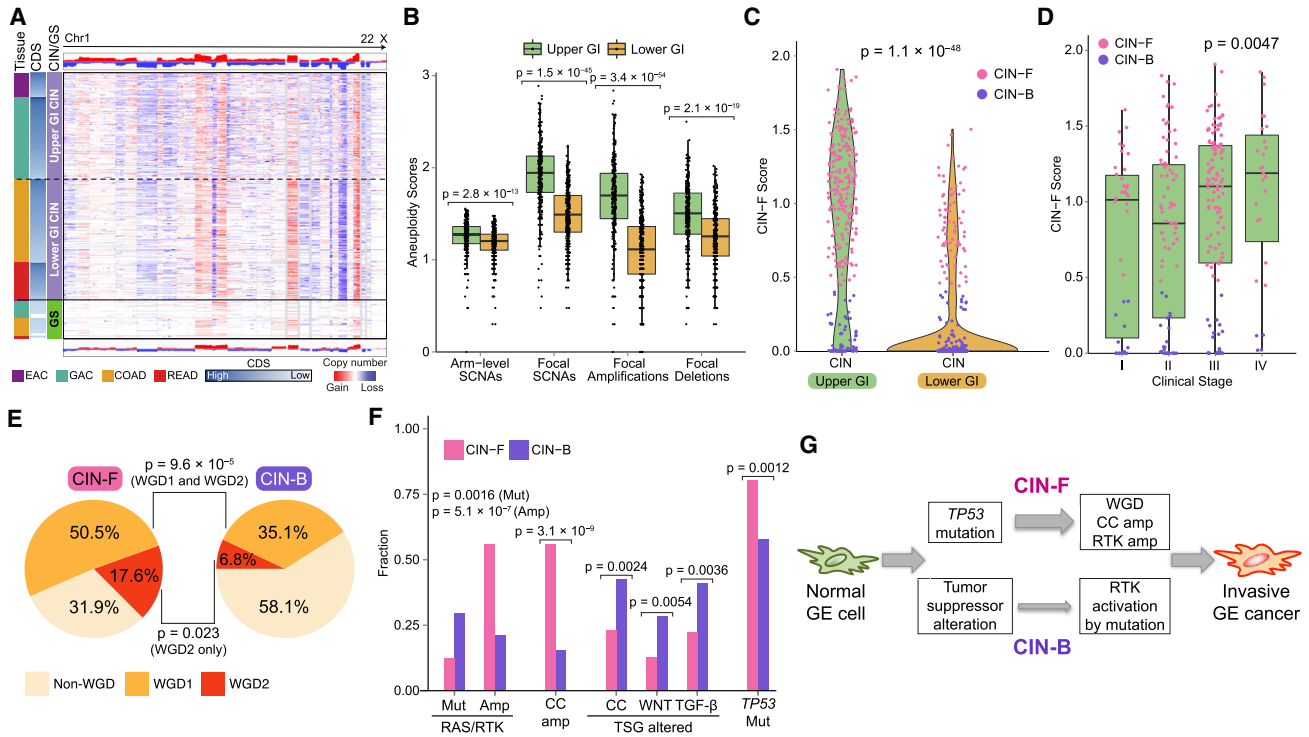



**Figure 4. Molecular Features of the CIN Subtype in Upper GI**

(A) Copy-number heatmap of non-hypermutated GIACs with amplification (red) and deletion (blue) with upper GI CIN tumors (top), CIN CRC (middle), and GS (bottom).

(B) Plots of arm-level and focal copy-number events in CIN tumors by the upper and lower GI tract. Horizontal bars indicate median values, boxes represent interquartile range, and whiskers indicate values within 1.5 times interquartile range.

(C) Distribution of CIN-F (CIN-Focal) score by upper and lower GI CIN tumors. CIN-B denotes CIN-Broad.

(D) Distribution of CIN-F score by clinical stage in the upper GI. Horizontal bars indicate median values, boxes represent interquartile range, and whiskers indicate values within 1.5 times interquartile range.

(E) Whole-genome doubling (WGD) in CIN-F and CIN-B tumors in the upper GI tract; WGD1 indicates one WGD, and WGD2 indicates > one WGD.

(F) Frequency of distinct classes of somatic alterations in RAS and receptor tyrosine kinases (RTK)(*KRAS, PIK3CA, BRAF, ERBB3, ERBB2, NRAS, EGFR, FGFR1,* and *FGFR2*), cell-cycle (CC; *FBXW7, CCNE1, CDK6, CDKN2A, CDKN1B, CCND1,* and *CCND2*), and tumor suppressor genes (TSG) including WNT (*APC, RNF43, SOX9, TCF7L2,* and *CTNNB1),* TGF-β*: TGFBR2, ACVR2A, ACVR1B, SMAD4, SMAD2,* and *SMAD*), and *TP53* in upper GI CIN-F and CIN-B tumors.

(G) Schematic model of CIN-F and CIN-B pathogenesis in the upper GI. See also Figure S4.

compared with ERBB2$^+$ GEA, which often carry co-occurring amplified oncogenes implicated in *de novo* resistance (Janjigian et al., 2018; Kim et al., 2014).

CIN-B and CIN-F CRC displayed comparable rates of *APC* and *KRAS* mutations (Figure S4I; *APC*: 79% versus 87%; *KRAS*: 35% versus 44%). However, *PIK3CA* mutations and TGF-β pathway alterations were more common in CIN-B CRC than in CIN-F CRC (Figure S4I). Both groups of CIN CRCs had somatic patterns more closely resembling the CIN-B GEA group, in which oncogenes were activated more commonly by mutation than by amplification. These data suggest that the preponderance of early *APC* loss and selection for mutational activation of oncogenes like *KRAS* may precede a form of aneuploidy and transformation distinct from the catastrophic aneuploidy and resulting oncogene amplification occurring in GEA (Figure 4G).

Among CIN CRCs, we observed more frequent CIMP, primarily CIMP-L, in proximal, right-sided CIN tumors and less frequent CIMP in distal, left-sided ones (Figure S5A). Arm-level SCNAs were significantly less frequent in CIMP$^+$ CIN CRCs (Wilcoxon $p = 2.7 \times 10^{-9}$), despite the lack of an overall difference in focal alterations (Figure S5B). Among chromosome arms, gain of 20q was most enriched in non-CIMP CIN CRC, with a mean copy-number gain of 1.8 (ploidy-adjusted), compared with 1.1 in CIMP$^+$ CIN CRC (Figure S5C). By contrast, except for *TP53*, which was more frequently mutated in non-CIMP CIN tumors, the frequency of somatic mutations was significantly higher in CIMP$^+$ CIN CRC (Figure S5D), notably affecting the TGF-β pathway and key oncogenes including *KRAS*/*NRAS*/*BRAF* and *PIK3CA*. Dichotomizing CIN CRC tumors by CIMP status thus showed parallels to the division of upper GI CIN tumors by CIN-F/CIN-B status. CIMP$^+$ CIN CRC, like CIN-B GEA, harbored more oncogene mutations (Figure 5A). Taken together, these data suggest that CIMP status may play an important role in shaping evolution of CIN tumors in the lower GI tract, and to a lesser extent in the upper GI tract.

## Molecular Features of the GS Subtype

Although CRCs are classically divided between hypermutated/MSI and CIN (Bijlsma et al., 2017), we detected a population of CRCs lacking both aneuploidy/CIN and hypermutation, a group

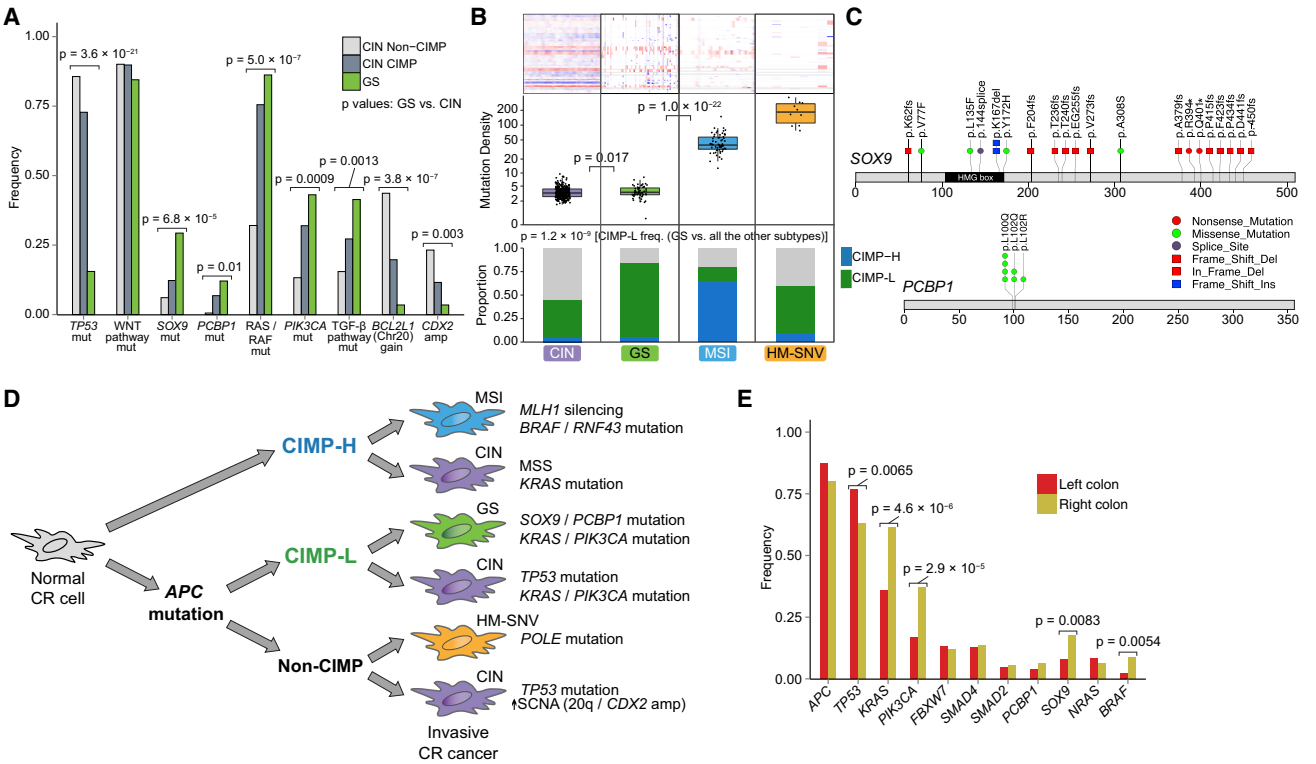



**Figure 5. Molecular Features of CIN and GS Colorectal Cancer**

(A) Frequency of somatic alterations in indicated genes or pathways in non-CIMP CIN, CIMP-H/L CIN, and GS lower GI tumors.

(B) SCNAs (top), mutation density (middle), and CIMP classes (bottom) across subtypes in the lower GI tract. Horizontal bars indicate median values, boxes represent interquartile range, and whiskers indicate values within 1.5 times interquartile range.

(C) Distribution of somatic mutations in *SOX9* and *PCBP1* in lower GI GS.

(D) Schematic model of pathogenesis of molecular subtypes in lower GI.

(E) Frequency of mutations in indicated genes in lower GI CIN/GS stratified anatomically. See also Figure S5.

we classified as GS (Figure 5B). Unlike with MSI, these GS CRCs shared few underlying biologic features with GS in upper GIACs. As we reported previously (Cancer Genome Atlas Research Network, 2014), upper GI GS tumors are enriched for the diffuse-type gastric cancer (65.7%) and commonly harbor mutations in *CDH1* and *RHOA* (Figure S5E). Thus, upper GI GS, such as EBV$^{+}$ tumors, represent an essentially unique entity confined to the stomach.

GS CRCs shared features of other CRCs; like the CIN CRCs, GS CRCs shared a predilection for loss of *APC* (GS 81% versus CIN 85%, Figure S5F). GS CRCs were more common in ascending and transverse colon (Figure 2E) and when compared with the CIN CRCs, showed significant enrichment for the CIMP-L phenotype (79% versus 40%, $p = 1.2 \times 10^{-9}$, Figure 5B) and for the CMS3 metabolic subtype ($p = 1.6 \times 10^{-6}$, Figure S2D) (Guinney et al., 2015). Despite having fewer SCNAs, a subset of GS CRCs showed amplifications of *IGF2* ($q < 0.05$) (Figure S5G). Mitogen-activated protein kinase pathway mutations were more common in these tumors, with *KRAS, NRAS*, or *BRAF* mutated in 69%, 10%, and 9% of tumors, respectively, and with *PIK3CA* mutations present in 43%, compared with 22% of CIN CRCs (Figure 5A). Consistent with the relative lack of aneuploidy, *TP53* mutations were less common (16%) in GS compared with CIN tumors (80%) (Figure S5E). However, we observed enrichment for somatic mutations in *SOX9*, which encodes a transcription factor, and in *PCBP1*, which encodes an RNA-binding protein that regulates splicing, mRNA stability, and translation (Leffers et al., 1995) (Figures 5A, 5C, and S5H). *SOX9*, mutated in 29% of GS CRCs, encodes a WNT-regulated transcription factor that controls cell fate and crypt homeostasis in intestinal development (McConnell et al., 2011; Nandan et al., 2014). GS CRCs with mutations in *SOX9* also had more frequent somatic mutations in the TGF-β pathway genes, including *PCBP1* (Figure S5I). Our mutation analysis within GS revealed highly clustered missense mutations in the KH domain of *PCBP1* in 13% of GS CRCs, raising the potential for a GOF event (Figure 5C). Interestingly, overexpression of wild-type *PCBP1* was associated with oxaliplatin resistance in CRC (Guo et al., 2010).

Overall, GS CRCs had more frequent mutations in the TGF-β pathway, RAS/RAF genes, and *PIK3CA* than did CIN CRCs (Figure S5F). Comparison of GS CRCs with CIN CRCs revealed a progressive gradation of features between non-CIMP CIN CRCs, CIMP-H or CIMP-L CIN CRCs, and GS CRCs (Figure 5A). These data suggest a pathway to cancer in the colorectum in which APC mutant cells, typically containing the CIMP-L phenotype, are able to undergo transformation by sustaining additional pathogenic mutations without the need for p53 loss or aneuploidy (Figure 5D).

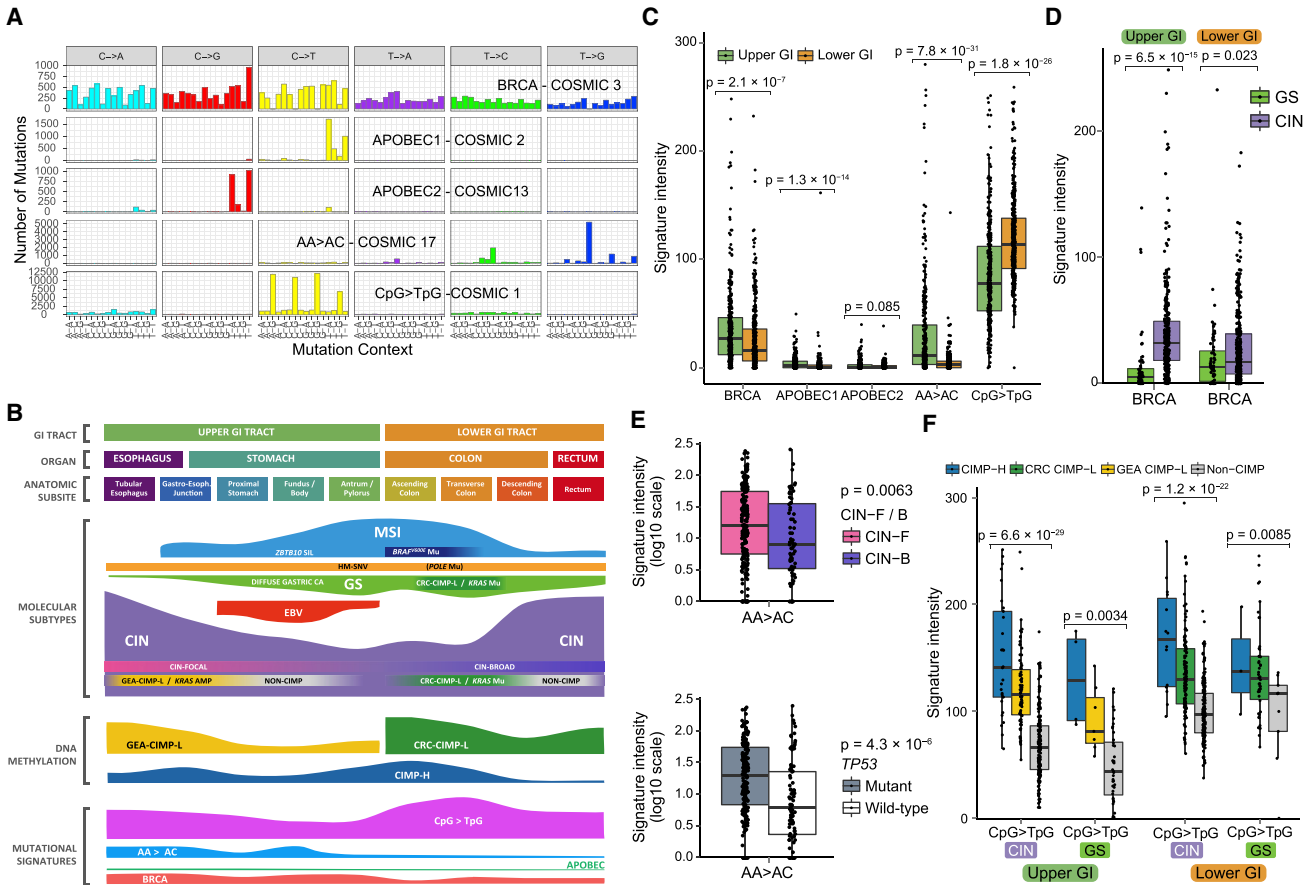



**Figure 6. Gastrointestinal Adenocarcinoma Mutational Signatures**

(A) Mutation signatures in non-hypermutated GIACs displayed by substitution class and sequence immediately 3′ and 5′ to the mutated base.

(B) Key molecular features of GIACs by anatomical distribution.

(C) Intensities of mutational signatures in CIN and GS subtypes by the upper and lower GI.

(D) BRCA signature in CIN and GS tumors in the upper and lower GI tract.

(E) AA > AC signature stratified by CIN-F and CIN-B (top) and *TP53* mutation (bottom) in upper GI CIN tumors.

(F) CpG > TpG signature in CIN and GS tumors in the upper and lower GI stratified by CIMP status. For all boxplots, horizontal bars indicate median values, boxes represent interquartile range, and whiskers indicate values within 1.5 times interquartile range. See also Figure S6.

We had noted earlier that CIMP-H MSI tumors appeared to rely less on WNT signaling. CIMP-H MSS tumors also displayed reduced rates of *APC* mutation (47%) compared with CIMP-L (87%) or non-CIMP (86%) (Fisher's exact test p = 0.00066). These findings suggest an alternative CRC pathway that is not initiated by mutation of *APC*, but rather by an epigenetic aberration causing CIMP-H. If *MLH1* is silenced in the context of CIMP-H, then the tumor would become MSI, whereas, if *MLH1* is not affected, the tumor would develop along the CIN pathway to give rise to CIMP-H MSS CIN tumors (Figure 5D). Non-hypermutated CRCs from the right-sided (ascending/transverse) colon revealed significantly higher rates of *KRAS, PIK3CA*, and *SOX9* mutation than those from the left-sided (descending) colon/rectum (Figure 5E).

## Mutational Signatures in GIACs

MSI and POLE signatures dominated the total mutational signature scores among GIACs as a consequence of the high mutational burden in MSI and POLE-deficient tumors (Figures S6A and S6B). Signature discovery following removal of hypermutated cases revealed a BRCA signature (COSMIC signature 3), two APOBEC signatures, a signature resembling COSMIC signature 17 with common AA > AC transversions, and a signature dominated by C > T transitions at CpG dinucleotides (COSMIC signature 1) (Figures 6A–6F) (Alexandrov et al., 2013; Bignell et al., 2010). The APOBEC signatures contributed minimally to the mutational profile across GIACs (Figures 6B, 6C, and S6C), but the other three signatures had substantial activity in non-hypermutated GIACs with the AA > AC signature limited to upper GIACs (Figures 6B, 6C, 6E, and S6D). A recent study identified the existence of the BRCA signature in gastric cancers that lacked mutations in *BRCA1* and *BRCA2* (Alexandrov et al., 2015). We confirmed the presence of BRCA signature activity in GIACs, with significant enrichment of somatic and germline mutations in several homologous recombination genes such as *BRCA1, BRCA2*, and *PALB2* (Figure S6E). BRCA signature activity was also significantly enriched in tumors with epigenetic silencing of *BRCA1* or *RAD51C,* especially within EBV$^{+}$ GCs

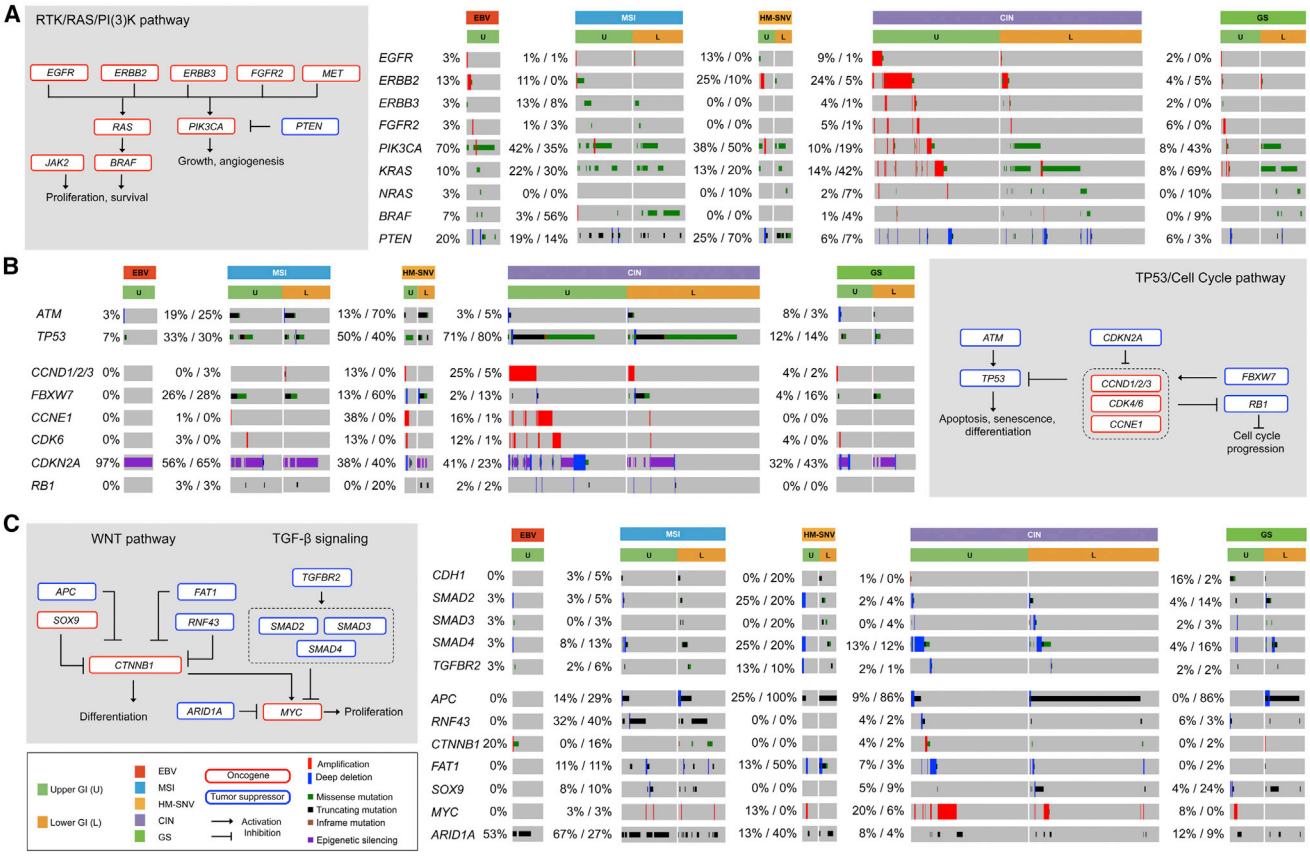



**Figure 7. Integrated Molecular Comparison of Somatic Alterations across GIAC Molecular Subtypes**

(A–C) Alterations in select genes and pathways including RTK/RAS/PI3-K (A), *TP53*, cell cycle (B), and WNT/TGF-β (C). Deep deletions representing loss of more than half of the gene copies for the given ploidy of the tumor, blue; amplifications, red; missense mutations in the COSMIC repository, green; nonsense or frameshift mutations, black. Percentage of somatic alteration is indicated by numbers to the left of each gene box and divided by the upper (U) and lower GI (L).

(Figure S6F). We observed a significant association between BRCA signature activity and upper GI cancers, particularly the CIN subtype (Figure 6D). The BRCA signature was associated only with focal SCNA events (Figure S6G), which are likely initiated by double-strand breaks. The AA > AC signature was also enriched in upper-GI CIN (Figures 6E, S6C, and S6D), most notably in the tubular esophagus (Figure S6D). Moreover, this mutational signature was enriched in CIN-F and *TP53*-mutated upper GI CIN tumors (Figure 6E). The AA > AC signature lacks a known etiology, but its association with GEA and its correlation with higher CIN-F scores raises the possibility that this signature reflects a process that contributes to greater focal aneuploidy observed in GEAs compared with CRCs and differences in oncogene profiles between upper and lower GIACs (Figure 7).

The CpG > TpG pattern, often termed the "aging signature", was the most common signature among all tumors, but it was especially frequent in right-sided CRC (Figure S6C). This signature is thought to arise from spontaneous hydrolytic deamination of 5-methylcytosine, and is consolidated as a persistent mutation if it occurs during DNA replication. Hence, this signature tracks the cumulative number of cell divisions and aging. Although we observed an association with CIMP status (Figure 6F), we do not believe that this is explained by a simple quantitative difference in DNA methylation. The CIMP hypermethylation is measured primarily at promoter CpG islands, which are unmethylated in normal cells and thus do not sustain many CpG > TpG mutations prior to acquisition of methylation and clonal expansion, whereas the mutation signatures were obtained by exome sequencing of gene bodies, which are generally highly methylated. The association between CIMP status and CpG > TpG signature may reflect the fact that CIMP tumors require more cell divisions to progress and thus acquire more CpG > TpG mutations over time.

## DISCUSSION

GIACs originate from columnar epithelium with a shared endodermal origin and display a spectrum of common molecular features, such as aneuploidy and MSI, which span anatomic boundaries. GIACs are enriched for activation of the WNT signaling pathway, particularly in the lower GI tract, consistent with the importance of WNT in GI development (Schepers and Clevers, 2012). We found that CIMP-H CRC appeared less dependent on canonical WNT signaling mutations and pathways. GIACs also displayed a predisposition for disruptions in TGF-β and SMAD signaling components. TGF-β signaling helps to maintain

intestinal stem cell equilibrium, promoting growth during development, but controlling self-renewal in adult epithelium (Mishra et al., 2005).

The vast majority of sporadic MSI tumors arise as a consequence of promoter methylation of *MLH1* in the context of CIMP-H. Methylation profiles of CIMP-H tumors are quite consistent throughout the GI tract. However, *MLH1* silencing within CIMP-H is much more anatomically restricted, primarily observed in the distal stomach and proximal (ascending and transverse) colon, but notably uncommon in proximal upper GI cancers. The epithelia of the distal stomach and proximal colon appear more susceptible to oncogenic effects of *MLH1* silencing. High rates of epithelial cell turnover with accompanying DNA replication may more effectively consolidate replication-associated errors in these sections of the GI tract. This hypothesis is consistent with the tumor spectrum observed with germline mutations in mismatch repair genes, leading to increased risk of cancers arising in highly proliferative tissues (Lynch et al., 2015). In this scenario, stochastic promoter methylation of *MLH1* from CIMP-H would provide less selective advantage when arising in the less proliferative sections of the GI tract.

CIMP-H GIACs possessed other differences in molecular features between various anatomic locations. $BRAF^{V600E}$ mutations occurred almost exclusively in CIMP-H tumors of the ascending colon and were absent from otherwise similar CIMP-H GEAs. In addition, some colorectal CIMP-H tumors with similar DNA methylation profiles lacked $BRAF^{V600E}$ mutations, a finding inconsistent with the proposed role for $BRAF^{V600E}$ mutation as a cause of CIMP-H (Fang et al., 2016). Alternatively, CIMP may provide a permissive environment for $BRAF^{V600E}$ mutation, perhaps by silencing pathways involved in oncogene-induced senescence and apoptosis (Hinoue et al., 2009). Despite the large overlap of CIMP-H and MSI in GIACs, our data revealed that this co-occurrence occurs predominantly in the distal stomach and ascending colon. The etiology for CIMP-H tumors commonly progressing via a CIN pathway in proximal GE and distal CRC is not established.

CIN is a common feature of GIACs and other tumors (Cancer Genome Atlas Research Network, 2012; 2014, 2017; Hoadley et al., 2014). Despite the deleterious effect on cellular and organismal fitness (Sheltzer et al., 2011, 2017; Torres et al., 2007; Williams et al., 2008), CIN with its resultant aneuploidy remains the predominant molecular subtype among GIACs, found most frequently in the proximal upper and distal lower GI tract (Dulak et al., 2012). Unlike tumors with MSI, CIN tumors had more discrepant molecular features between upper and lower GI cancers. Most striking was the preponderance of focal, high-amplitude SCNAs, especially amplifications, in GEAs. Within CIN GEAs, we found that tumors with high CIN-F scores had a strong association with prior genome doubling, a process associated with CIN (Ganem et al., 2007). Amplifications in CIN-F GEAs commonly targeted mitogen pathway components, cell-cycle regulators, and lineage survival transcription factors, whereas CIN-B and GS tumors more frequently carried activating mutations in these pathways.

A notable finding was the predilection in CIN-B GEAs for alterations in tumor suppressors such as *CDKN2A, APC*, and *SMAD4*. These findings suggest that the marked aneuploidy found within the CIN-F GEAs is less apt to occur in precursors with pathogenic alterations other than *TP53*. One explanation is that precursors with altered oncogenes/tumor suppressors other than *TP53* have less requirement for more "catastrophic" aneuploidy to simultaneously abrogate multiple such checkpoints. By contrast, such marked instability could facilitate transformation in precursors with p53 loss without as many other preexisting pathogenic alterations. Indeed, although p53 loss alone is not sufficient to promote aneuploidy (Bunz et al., 2002), several lines of evidence support its necessity, most likely by circumventing p53-dependent cell-cycle arrest in response to damage by reactive oxygen species (Guo et al., 2010), to mutations in ataxia telangiectasia (Li et al., 2010), or to spindle assembly checkpoint activation (Thompson and Compton, 2010). Given these data, the lesser rates of CIN-F in lower GI CIN tumors (compared with CIN tumors of the upper GI tract) may be a consequence of *APC* loss as an early event in colorectal neoplasia, leading to mutations in *TP53* rarely occurring in the absence of a prior *APC* loss. Instead, we noted that CIMP status likely has a stronger influence on the features of CIN CRC, with CIMP being associated with mutations in *KRAS* and in tumor suppressor pathways such as TGF-β. Aneuploid CIMP tumors in the lower GI tract showed lower rates of SCNA, but a greater number of oncogenic mutations compared with non-CIMP. Both upper and lower GI CIN tumors were also associated with the BRCA mutational signature. However, the propensity for greater CIN-F in upper GIACs correlated with the AA > AC mutational signature, a signature of unknown etiology, previously reported in upper GI tumors (Dulak et al., 2013).

Our exploration of the role of CIMP in shaping the features of CIN in CRC became linked with our finding of a GS CRC subtype falling outside the classic CIN/MSI CRC dichotomy. This GS subtype may partially overlap with the previously identified microsatellite and chromosome stable (MACS) CRCs (Chan et al., 2001), while showing important differences. The MACS phenotype is an independent predictor of poor outcome (Banerjea et al., 2009), in contrast, GS CRCs are enriched for earlier stage tumors. MACS tumors have an elevated proportion of early onset cases (Chan et al., 2001), whereas GS CRCs have a higher mean age at diagnosis than CIN cases. Like MACS, HM-SNV cases are microsatellite and genome stable, and also arise in younger patients, so it is possible that some early-onset MACS tumors may have represented unrecognized HM-SNV tumors. The GS CRCs overlap with a subgroup identified by gene expression clustering as CMS3 (Guinney et al., 2015) and commonly displaying CIMP-L. Many features enriched in CIMP CIN CRCs compared with non-CIMP CIN CRCs were even more prevalent in GS CRCs. Moreover, we found these tumors to have recurrent mutations in *SOX9* and *PCBP1*. While the presence of frameshift mutations of *SOX9* implies LOF, truncating mutations in *SOX9* are overexpressed in primary tumor specimens (Javier et al., 2016), making their functional significance unclear. GS CRCs with mutations in *SOX9* also had more frequent somatic mutations in TGF-β pathway genes, including *PCBP1*, which impacts TGF-β signaling by regulating Smad3-associated alternative splicing (Tripathi et al., 2016). Given the strikingly low frequency of *TP53* mutations in GS CRCs, the presence of *SOX9* and *PCBP1* mutations may co-operate with *APC* and *KRAS* mutation

to facilitate transformation, despite lack of hypermutation and low levels of aneuploidy.

Our findings also bear some relevance to the evolving field of immunotherapy, which already has established efficacy in MSI tumors. The HM-SNV tumors, which display a large SNV burden in the setting of *POLE* mutations, did not harbor the equivalent CD8 or IFN-γ signatures as did the MSI tumors, perhaps suggesting that indel mutations may better generate neoantigens than SNVs. The strong signatures in $EBV^+$ tumors suggest a potential for immune checkpoint inhibition in this subset. The reason for consistently higher IFN-γ signatures in upper GI compared with lower GI adenocarcinomas when stratified by molecular subtype is less obvious and may simply indicate that GEAs are more immunogenic than CRC, results consistent with the presence of clinical responses to PD1 inhibitor monotherapy in MSS GEAs, but not in CRCs (Jin and Yoon, 2016; Muro et al., 2016).

In summary, these results highlight how processes such as DNA hypermethylation and CIN can manifest themselves in different ways across related tissues. In some instances, as with DNA hypermethylation in upper-GI versus lower-GI MSI tumors, such differences can be subtle. However, as the exploration of CIN indicates, other processes can lead to substantially different molecular outcomes across these regions. Provision of greater detail in the various manifestations of molecular defects may reveal new opportunities for targeted therapies for these cancers. Furthermore, these data highlight how consideration of molecular subtypes as well as organ of origin will be essential in the study and treatment of cancer.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:



## SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.ccell.2018.03.010.

## ACKNOWLEDGMENTS

We thank all patients who contributed to this study. This work was supported by the Intramural Research Program and the following grants from the United States NIH: U24 CA143799, U24 CA143835, U24 CA143840, U24 CA143843, U24 CA143845, U24 CA143848, U24 CA143858, U24 CA143866, U24 CA143867, U24 CA143882, U24 CA143883, U24 CA144025, U54 HG003067, U54 HG003079, U54 HG003273, and P30 CA16672.

## AUTHOR CONTRIBUTIONS

Conceptualization: YL, NSS, TH, ADC, JAS, FF, RB, VT, AJB, PWL.

Data Analysis: YL, NSS, TH, ADC, FSV, JAS, FF, RB, MI, JK, WC, RA, RSK, VT, Clinicopathologic Analysis: CSR, JEW, KKW, SJM, LM, AIO, SB, CSP, AJL.

Critical Thinking: YL, NSS, TH, ADC, JAS, FF, RB, VT, AJB, PWL.

Writing - Original Draft: YL, NSS, TH, BGS, ADC, FF, VT, AJB, PWL.

Visualization: YL, NSS, TH, ADC, RB, MI, JK, WC, RA, RSK, RS, VT.

Writing - Review & Editing: YL, NSS, TH, BGS, ADC, VT, AJB, PWL.

## DECLARATION OF INTERESTS

Michael Seiler, Peter G. Smith, Ping Zhu, Silvia Buonamici, and Lihua Yu are employees of H3 Biomedicine, Inc. Parts of this work are the subject of a patent application: WO2017040526 titled "Splice variants associated with neomorphic sf3b1 mutants." Shouyoung Peng, Anant A. Agrawal, James Palacino, and Teng Teng are employees of H3 Biomedicine, Inc. Andrew D. Cherniack, Ashton C. Berger, and Galen F. Gao receive research support from Bayer Pharmaceuticals. Gordon B. Mills serves on the External Scientific Review Board of Astrazeneca. Anil Sood is on the Scientific Advisory Board for Kiyatec and is a shareholder in BioPath. Jonathan S. Serody receives funding from Merck, Inc. Kyle R. Covington is an employee of Castle Biosciences, Inc. Preethi H. Gunaratne is founder, CSO, and shareholder of NextmiRNA Therapeutics. Christina Yau is a part-time employee/consultant at NantOmics. Franz X. Schaub is an employee and shareholder of SEngine Precision Medicine, Inc. Carla Grandori is an employee, founder, and shareholder of SEngine Precision Medicine, Inc. Robert N. Eisenman is a member of the Scientific Advisory Boards and shareholder of Shenogen Pharma and Kronos Bio. Daniel J. Weisenberger is a consultant for Zymo Research Corporation. Joshua M. Stuart is the founder of Five3 Genomics and shareholder of NantOmics. Marc T. Goodman receives research support from Merck, Inc. Andrew J. Gentles is a consultant for Cibermed. Charles M. Perou is an equity stock holder, consultant, and Board of Directors member of BioClassifier and GeneCentric Diagnostics and is also listed as an inventor on patent applications on the Breast PAM50 and Lung Cancer Subtyping assays. Matthew Meyerson receives research support from Bayer Pharmaceuticals; is an equity holder in, consultant for, and Scientific Advisory Board chair for OrigiMed; and is an inventor of a patent for EGFR mutation diagnosis in lung cancer, licensed to LabCorp. Eduard Porta-Pardo is an inventor of a patent for domainXplorer. Han Liang is a shareholder and scientific advisor of Precision Scientific and Eagle Nebula. Da Yang is an inventor on a pending patent application describing the use of antisense oligonucleotides against specific lncRNA sequence as diagnostic and therapeutic tools. Yonghong Xiao was an employee and shareholder of TESARO, Inc. Bin Feng is an employee and shareholder of TESARO, Inc. Carter Van Waes received research funding for the study of IAP inhibitor

ASTX660 through a Cooperative Agreement between NIDCD, NIH, and Astex Pharmaceuticals. Raunaq Malhotra is an employee and shareholder of Seven Bridges, Inc. Peter W. Laird serves on the Scientific Advisory Board for AnchorDx. Joel Tepper is a consultant at EMD Serono. Kenneth Wang serves on the Advisory Board for Boston Scientific, Microtech, and Olympus. Andrea Califano is a founder, shareholder, and advisory board member of DarwinHealth, Inc. and a shareholder and advisory board member of Tempus, Inc. Toni K. Choueiri serves as needed on advisory boards for Bristol-Myers Squibb, Merck, and Roche. Lawrence Kwong receives research support from Array BioPharma. Sharon E. Plon is a member of the Scientific Advisory Board for Baylor Genetics Laboratory. Beth Y. Karlan serves on the Advisory Board of Invitae.

Received: July 21, 2017
Revised: January 25, 2018
Accepted: March 7, 2018
Published: April 2, 2018

## REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Nik-Zainal, S., Siu, H.C., Leung, S.Y., and Stratton, M.R. (2015). A mutational signature in gastric cancer suggests therapeutic strategies. Nat. Commun. *6*, 8683.

Arnold, M., Soerjomataram, I., Ferlay, J., and Forman, D. (2015). Global incidence of oesophageal cancer by histological subtype in 2012. Gut *64*, 381–387.

Banerjea, A., Hands, R.E., Powar, M.P., Bustin, S.A., and Dorudi, S. (2009). Microsatellite and chromosomal stable colorectal cancers demonstrate poor immunogenicity and early disease recurrence. Colorectal Dis. *11*, 601–608.

Bignell, G.R., Greenman, C.D., Davies, H., Butler, A.P., Edkins, S., Andrews, J.M., Buck, G., Chen, L., Beare, D., Latimer, C., et al. (2010). Signatures of mutation and selection in the cancer genome. Nature *463*, 893–898.

Bijlsma, M.F., Sadanandam, A., Tan, P., and Vermeulen, L. (2017). Molecular subtypes in cancers of the gastrointestinal tract. Nat. Rev. Gastroenterol. Hepatol. *14*, 333–342.

Bonhomme, C., Duluc, I., Martin, E., Chawengsaksophak, K., Chenard, M.P., Kedinger, M., Beck, F., Freund, J.N., and Domon-Dell, C. (2003). The Cdx2 homeobox gene has a tumour suppressor function in the distal colon in addition to a homeotic role during gut development. Gut *52*, 1465–1471.

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K.O., Di Narzo, A.F., Yan, P., Hodgson, J.G., Weinrich, S., et al. (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J. Pathol. *231*, 63–76.

Bunz, F., Fauth, C., Speicher, M.R., Dutriaux, A., Sedivy, J.M., Kinzler, K.W., Vogelstein, B., and Lengauer, C. (2002). Targeted inactivation of p53 in human cells does not result in aneuploidy. Cancer Res. *62*, 1129–1133.

Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Cancer Genome Atlas Research Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209.

Cancer Genome Atlas Research Network. (2017). Integrated genomic characterization of oesophageal carcinoma. Nature *541*, 169–175.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.

Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: a precision oncology knowledge base. JCO Precis. Oncol. https://doi.org/10.1200/PO.17.00011.

Chan, T.L., Curtis, L.C., Leung, S.Y., Farrington, S.M., Ho, J.W., Chan, A.S., Lam, P.W., Tse, C.W., Dunlop, M.G., Wyllie, A.H., et al. (2001). Early-onset colorectal cancer with stable microsatellite DNA and near-diploid chromosomes. Oncogene *20*, 4871–4876.

Chu, J., Sadeghi, S., Raymond, A., Jackman, S.D., Nip, K.M., Mar, R., Mohamadi, H., Butterfield, Y.S., Robertson, A.G., and Birol, I. (2014). BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. Bioinformatics *30*, 3402–3404.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics *27*, 2601–2602.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. *41*, e67.

Cristescu, R., Lee, J., Nebozhyn, M., Kim, K.M., Ting, J.C., Wong, S.S., Liu, J., Yue, Y.G., Wang, J., Yu, K., et al. (2015). Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat. Med. *21*, 449–456.

Daily, K., Ho Sui, S.J., Schriml, L.M., Dexheimer, P.J., Salomonis, N., Schroll, R., Bush, S., Keddache, M., Mayhew, C., Lotia, S., et al. (2017). Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives. Sci. Data *4*, 170030.

Davoli, T., Uno, H., Wooten, E.C., and Elledge, S.J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science *355*, https://doi.org/10.1126/science.aaf8399.

Derks, S., Liao, X., Chiaravalli, A.M., Xu, X., Camargo, M.C., Solcia, E., Sessa, F., Fleitas, T., Freeman, G.J., Rodig, S.J., et al. (2016). Abundant PD-L1 expression in Epstein-Barr virus-infected gastric cancers. Oncotarget *7*, 32925–32932.

Dienstmann, R., Vermeulen, L., Guinney, J., Kopetz, S., Tejpar, S., and Tabernero, J. (2017). Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat. Rev. Cancer *17*, 79–92.

Dulak, A.M., Schumacher, S.E., van Lieshout, J., Imamura, Y., Fox, C., Shim, B., Ramos, A.H., Saksena, G., Baca, S.C., Baselga, J., et al. (2012). Gastrointestinal adenocarcinomas of the esophagus, stomach, and colon exhibit distinct patterns of genome instability and oncogenesis. Cancer Res. *72*, 4383–4393.

Dulak, A.M., Stojanov, P., Peng, S., Lawrence, M.S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S.E., Shefler, E., et al. (2013). Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat. Genet. *45*, 478–486.

Ebert, M.P., Tanzer, M., Balluff, B., Burgermeister, E., Kretzschmar, A.K., Hughes, D.J., Tetzner, R., Lofton-Day, C., Rosenberg, R., Reinacher-Schick, A.C., et al. (2012). TFAP2E-DKK4 and chemoresistance in colorectal cancer. N. Engl. J. Med. *366*, 44–53.

Fang, M., Hutchinson, L., Deng, A., and Green, M.R. (2016). Common BRAF(V600E)-directed pathway mediates widespread epigenetic silencing in colorectal cancer and melanoma. Proc. Natl. Acad. Sci. USA *113*, 1250–1255.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945–D950.

Ganem, N.J., Storchova, Z., and Pellman, D. (2007). Tetraploidy, aneuploidy and cancer. Curr. Opin. Genet. Dev. *17*, 157–162.

GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis

Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA, Biospecimen Collection Source Site—NDRI; et al (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Guinney, J., Dienstmann, R., Wang, X., de Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356.

Guo, Z., Kozlov, S., Lavin, M.F., Person, M.D., and Paull, T.T. (2010). ATM activation by oxidative stress. Science *330*, 517–521.

Herman, J.G., Umar, A., Polyak, K., Graff, J.R., Ahuja, N., Issa, J.P., Markowitz, S., Willson, J.K., Hamilton, S.R., Kinzler, K.W., et al. (1998). Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc. Natl. Acad. Sci. USA *95*, 6870–6875.

Hinoue, T., Weisenberger, D.J., Pan, F., Campan, M., Kim, M., Young, J., Whitehall, V.L., Leggett, B.A., and Laird, P.W. (2009). Analysis of the association between CIMP and BRAF in colorectal cancer by DNA methylation profiling. PLoS One *4*, e8357.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D., Niu, B., McLellan, M.D., Uzunangelov, V., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell, https://doi.org/10.1016/j.cell.2018.03.022.

International Agency for Research on Cancer. (2010). WHO Classification of Tumours of the Digestive System, Fourth Edition (WHO).

Isella, C., Terrasi, A., Bellomo, S.E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C., et al. (2015). Stromal contribution to the colorectal cancer transcriptome. Nat. Genet. *47*, 312–319.

Janjigian, Y.Y., Sanchez-Vega, F., Jonsson, P., Chatila, W.K., Hechtman, J.F., Ku, G.Y., Riches, J.C., Tuvy, Y., Kundra, R., Bouvier, N., et al. (2018). Genetic predictors of response to systemic therapy in esophagogastric cancer. Cancer Discov. *8*, 49–58.

Javier, B.M., Yaeger, R., Wang, L., Sanchez-Vega, F., Zehir, A., Middha, S., Sadowska, J., Vakiani, E., Shia, J., Klimstra, D., et al. (2016). Recurrent, truncating SOX9 mutations are associated with SOX9 overexpression, KRAS mutation, and TP53 wild type status in colorectal carcinoma. Oncotarget *7*, 50875–50882.

Jin, Z., and Yoon, H.H. (2016). The promise of PD-1 inhibitors in gastro-esophageal cancers: microsatellite instability vs. PD-L1. J. Gastrointest. Oncol. *7*, 771–788.

Kim, J., Fox, C., Peng, S., Pusung, M., Pectasides, E., Matthee, E., Hong, Y.S., Do, I.G., Jang, J., Thorner, A.R., et al. (2014). Preexisting oncogenic events impact trastuzumab sensitivity in ERBB2-amplified gastroesophageal adenocarcinoma. J. Clin. Invest. *124*, 5145–5158.

Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., Tiao, G., Kwiatkowski, D.J., Rosenberg, J.E., Van Allen, E.M., D'Andrea, A.D., et al. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. *48*, 600–606.

Koh, J., Ock, C.Y., Kim, J.W., Nam, S.K., Kwak, Y., Yun, S., Ahn, S.H., Park, D.J., Kim, H.H., Kim, W.H., et al. (2017). Clinicopathologic implications of immune classification by PD-L1 expression and CD8-positive tumor-infiltrating lymphocytes in stage II and III gastric cancer patients. Oncotarget *8*, 26356–26367.

Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K., et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. *40*, 1253–1260.

Kostic, A.D., Ojesina, A.I., Pedamallu, C.S., Jung, J., Verhaak, R.G., Getz, G., and Meyerson, M. (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat. Biotechnol. *29*, 393–396.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Le, D.T., Durham, J.N., Smith, K.N., Wang, H., Bartlett, B.R., Aulakh, L.K., Lu, S., Kemberling, H., Wilt, C., Luber, B.S., et al. (2017). Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science *357*, 409–413.

Leffers, H., Dejgaard, K., and Celis, J.E. (1995). Characterisation of two major cellular poly(rC)-binding human proteins, each containing three K-homologous (KH) domains. Eur. J. Biochem. *230*, 447–453.

Leung, S.Y., Yuen, S.T., Chung, L.P., Chu, K.M., Chan, A.S., and Ho, J.C. (1999). hMLH1 promoter methylation and lack of hMLH1 expression in sporadic gastric carcinomas with high-frequency microsatellite instability. Cancer Res. *59*, 159–164.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, M., Fang, X., Baker, D.J., Guo, L., Gao, X., Wei, Z., Han, S., van Deursen, J.M., and Zhang, P. (2010). The ATM-p53 pathway suppresses aneuploidy-induced tumorigenesis. Proc. Natl. Acad. Sci. USA *107*, 14188–14193.

Ljunggren, H.G., and Malmberg, K.J. (2007). Prospects for the use of NK cells in immunotherapy of human cancer. Nat. Rev. Immunol. *7*, 329–339.

Lynch, H.T., Snyder, C.L., Shaw, T.G., Heinen, C.D., and Hitchins, M.P. (2015). Milestones of Lynch syndrome: 1895–2015. Nat. Rev. Cancer *15*, 181–194.

Malta, T.M., Sokolov, A.,Gentles, A.J., Burzykowski, T.,Poisson, L., Weinstein, J.N., Kamińska, B., Huelsken,J., Omberg,L., Gevaert, O. (2018). Machine learning identifiesstemness features associated with oncogenic dedifferentiation. Cell, https://doi.org/10.1016/j.cell.2018.03.034.

Matsusaka, K., Kaneda, A., Nagae, G., Ushiku, T., Kikuchi, Y., Hino, R., Uozaki, H., Seto, Y., Takada, K., Aburatani, H., et al. (2011). Classification of Epstein-Barr virus-positive gastric cancers by definition of DNA methylation epigenotypes. Cancer Res. *71*, 7187–7197.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166–1174.

McConnell, B.B., Kim, S.S., Yu, K., Ghaleb, A.M., Takeda, N., Manabe, I., Nusrat, A., Nagai, R., and Yang, V.W. (2011). Kruppel-like factor 5 is important for maintenance of crypt architecture and barrier function in mouse intestine. Gastroenterology *141*, 1302–1313, 1313.e1–6.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Mishra, L., Shetty, K., Tang, Y., Stuart, A., and Byers, S.W. (2005). The role of TGF-beta and Wnt signaling in gastrointestinal stem cells and cancer. Oncogene *24*, 5775–5789.

Moons, L.M., Bax, D.A., Kuipers, E.J., Van Dekken, H., Haringsma, J., Van Vliet, A.H., Siersema, P.D., and Kusters, J.G. (2004). The homeodomain protein CDX2 is an early marker of Barrett's oesophagus. J. Clin. Pathol. *57*, 1063–1068.

Muro, K., Chung, H.C., Shankaran, V., Geva, R., Catenacci, D., Gupta, S., Eder, J.P., Golan, T., Le, D.T., Burtness, B., et al. (2016). Pembrolizumab for patients with PD-L1-positive advanced gastric cancer (KEYNOTE-012): a multicentre, open-label, phase 1b trial. Lancet Oncol. *17*, 717–726.

Nandan, M.O., Ghaleb, A.M., Liu, Y., Bialkowska, A.B., McConnell, B.B., Shroyer, K.R., Robine, S., and Yang, V.W. (2014). Inducible intestine-specific deletion of Kruppel-like factor 5 is characterized by a regenerative response in adult mouse colon. Dev. Biol. *387*, 191–202.

Noah, T.K., Donahue, B., and Shroyer, N.F. (2011). Intestinal development and differentiation. Exp. Cell Res. *317*, 2702–2710.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557–572.

Palles, C., Cazier, J.B., Howarth, K.M., Domingo, E., Jones, A.M., Broderick, P., Kemp, Z., Spain, S.L., Guarino, E., Salguero, I., et al. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat. Genet. *45*, 136–144.

Phipps, A.I., Passarelli, M.N., Chan, A.T., Harrison, T.A., Jeon, J., Hutter, C.M., Berndt, S.I., Brenner, H., Caan, B.J., Campbell, P.T., et al. (2016). Common genetic variation and survival after colorectal cancer diagnosis: a genome-wide analysis. Carcinogenesis *37*, 87–95.

Ramos, A.H., Lichtenstein, L., Gupta, M., Lawrence, M.S., Pugh, T.J., Saksena, G., Meyerson, M., and Getz, G. (2015). Oncotator: cancer variant annotation tool. Hum. Mutat. *36*, E2423–E2429.

Ratan, A., Olson, T.L., Loughran, T.P., Jr., and Miller, W. (2015). Identification of indels in next-generation sequencing data. BMC Bioinformatics *16*, 42.

Reed, K.R., Korobko, I.V., Ninkina, N., Korobko, E.V., Hopkins, B.R., Platt, J.L., Buchman, V., and Clarke, A.R. (2015). Hunk/Mak-v is a negative regulator of intestinal cell proliferation. BMC Cancer *15*, 110.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat. Genet. *38*, 500–501.

Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M.H., Chresta, C.M., Rosenberg, R., Nitsche, U., et al. (2014). Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int. J. Cancer *134*, 552–562.

Salari, K., Spulak, M.E., Cuff, J., Forster, A.D., Giacomini, C.P., Huang, S., Ko, M.E., Lin, A.Y., van de Rijn, M., and Pollack, J.R. (2012). CDX2 is an amplified lineage-survival oncogene in colorectal cancer. Proc. Natl. Acad. Sci. USA *109*, E3196–E3205.

Salomonis, N., Dexheimer, P.J., Omberg, L., Schroll, R., Bush, S., Huo, J., Schriml, L., Ho Sui, S., Keddache, M., Mayhew, C., et al. (2016). Integrated genomic analysis of diverse induced pluripotent stem cells from the Progenitor Cell Biology Consortium. Stem Cell Reports *7*, 110–125.

Sartore-Bianchi, A., Trusolino, L., Martino, C., Bencardino, K., Lonardi, S., Bergamo, F., Zagonel, V., Leone, F., Depetris, I., Martinelli, E., et al. (2016). Dual-targeted therapy with trastuzumab and lapatinib in treatment-refractory, KRAS codon 12/13 wild-type, HER2-positive metastatic colorectal cancer (HERACLES): a proof-of-concept, multicentre, open-label, phase 2 trial. Lancet Oncol. *17*, 738–746.

Schepers, A., and Clevers, H. (2012). Wnt signaling, stem cells, and cancer of the gastrointestinal tract. Cold Spring Harb. Perspect. Biol. *4*, a007989.

Secrier, M., Li, X., de Silva, N., Eldridge, M.D., Contino, G., Bornschein, J., MacRae, S., Grehan, N., O'Donovan, M., Miremadi, A., et al. (2016). Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat. Genet. *48*, 1131–1141.

Sheltzer, J.M., Blank, H.M., Pfau, S.J., Tange, Y., George, B.M., Humpton, T.J., Brito, I.L., Hiraoka, Y., Niwa, O., and Amon, A. (2011). Aneuploidy drives genomic instability in yeast. Science *333*, 1026–1030.

Sheltzer, J.M., Ko, J.H., Replogle, J.M., Habibe Burgos, N.C., Chung, E.S., Meehl, C.M., Sayles, N.M., Passerini, V., Storchova, Z., and Amon, A. (2017). Single-chromosome gains commonly function as tumor suppressors. Cancer Cell *31*, 240–255.

Sherwood, R.I., Chen, T.Y., and Melton, D.A. (2009). Transcriptional dynamics of endodermal organ formation. Dev. Dyn. *238*, 29–42.

Sia, E.A., Kokoska, R.J., Dominska, M., Greenwell, P., and Petes, T.D. (1997). Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol. Cell Biol. *17*, 2851–2858.

Sokolov, A., Paull, E.O., and Stuart, J.M. (2016). One-class detection of cell States in tumor subtypes. Pac. Symp. Biocomput. *21*, 405–416.

Tan, I.B., Ivanova, T., Lim, K.H., Ong, C.W., Deng, N., Lee, J., Tan, S.H., Wu, J., Lee, M.H., Ooi, C.H., et al. (2011). Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. Gastroenterology *141*, 476–485, 485.e1–11.

Thompson, S.L., and Compton, D.A. (2010). Proliferation of aneuploid human cells is limited by a p53-dependent mechanism. J. Cell Biol. *188*, 369–381.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.-H., Porta Pardo, E., Gao, G., Eddy, J.A., Plaisier, C.L., et al. (2018). The ilmmune landscape of cancer. Immunity, https://doi.org/10.1016/j.immuni.2018.03.023.

Torre, L.A., Siegel, R.L., Ward, E.M., and Jemal, A. (2016). Global cancer incidence and mortality rates and trends–an update. Cancer Epidemiol. Biomarkers Prev. *25*, 16–27.

Torres, E.M., Sokolsky, T., Tucker, C.M., Chan, L.Y., Boselli, M., Dunham, M.J., and Amon, A. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science *317*, 916–924.

Tripathi, V., Sixt, K.M., Gao, S., Xu, X., Huang, J., Weigert, R., Zhou, M., and Zhang, Y.E. (2016). Direct regulation of alternative splicing by SMAD3 through PCBP1 is essential to the tumor-promoting role of TGF-beta. Mol. Cell *64*, 1010.

Wagner, J.A., Rosario, M., Romee, R., Berrien-Elliott, M.M., Schneider, S.E., Leong, J.W., Sullivan, R.P., Jewell, B.A., Becker-Hapak, M., Schappe, T., et al. (2017). CD56bright NK cells exhibit potent antitumor responses following IL-15 priming. J. Clin. Invest. *127*, 4042–4058.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Wang, K., Yuen, S.T., Xu, J., Lee, S.P., Yan, H.H., Shi, S.T., Siu, H.C., Deng, S., Chu, K.M., Law, S., et al. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat. Genet. *46*, 573–582.

Weisenberger, D.J., Siegmund, K.D., Campan, M., Young, J., Long, T.I., Faasse, M.A., Kang, G.H., Widschwendter, M., Weener, D., Buchanan, D., et al. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat. Genet. *38*, 787–793.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D.J., Campan, M., Young, J., Jacobs, I., et al. (2007). Epigenetic stem cell signature in cancer. Nat. Genet. *39*, 157–158.

Williams, B.R., Prabhu, V.R., Hunter, K.E., Glazier, C.M., Whittaker, C.A., Housman, D.E., and Amon, A. (2008). Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. Science *322*, 703–709.

Zhou, Q., Talvinen, K., Sundstrom, J., Elzagheid, A., Pospiech, H., Syvaoja, J.E., and Collan, Y. (2009). Mutations/polymorphisms in the 55 kDa subunit of DNA polymerase epsilon in human colorectal cancer. Cancer Genomics Proteomics *6*, 297–304.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Antibodies | | |
| RPPA antibodies | RPPA Core Facility, MD Anderson Cancer Center | https://www.mdanderson.org/research/research-resources/core-facilities/functional-proteomics-rppa-core.html |
| Biological Samples | | |
| Tumor and normal tissue and blood samples | TCGA Network | https://portal.gdc.cancer.gov/legacy-archive/ |
| Critical Commercial Assays | | |
| DNA/RNA AllPrep kit | Qiagen | Cat# 80204 |
| mirVana miRNA Isolation kit | Ambion | Cat# AM1560 |
| QiaAmp blood midi kit | Qiagen | Cat# 51185 |
| AmpFISTR Identifiler kit | Applied Biosystems | Cat# A30737 |
| RNA6000 nano Assay | Agilent | Cat# 5067-1511 |
| SureSelect Human All Exon 50 Mb | Agilent | Cat# G3370J |
| Genome-Wide Human SNP Array 6.0 | Affymetrix | Cat# 901150 |
| Illumina Barcoded Paired-End Library Preparation kit | Illumina | http://www.hgsc.bcm.edu/sites/-default/files/documents/-Illumina_Barcoded_Paired-End_Capture_Library_Preparation.pdf |
| TruSeq PE Cluster Generation kit | Illumina | PE-401-3001 |
| Phusion PCR Supermix HiFi (2X) | New England Biolabs | Cat# M0531L |
| HumanMethylation450 | Infinium | Cat# WG-314-1002 |
| HumanMethylation450 | Infinium | Cat# WG-311-2201 |
| mRNA TruSeq kit | Illumina | Cat# RS-122-2001 |
| Deposited Data | | |
| Raw genomic and clinical data | NCI Genomic Data Commons | https://portal.gdc.cancer.gov/legacy-archive/ |
| MC3 mutation annotation file | NCI Genomic Data Commons | https://gdc.cancer.gov/about-data/publications/mc3-2017 |
| Processed data files | NCI Genomic Data Commons | https://gdc.cancer.gov/about-data/publications/pancanatlas |
| Software and Algorithms | | |
| Broad Institute QC on BAM files - ContEst | (Cibulskis et al., 2011) | PMID: 21803805 |
| Broad Institute Mutation Calling - MuTect | (Cibulskis et al., 2013) | PMID: 23396013 |
| Broad Institute small indel Calling - Indelocator | | https://www.broadinstitute.org/cancer/cga/indelocator |
| Broad Institute Mutation/Indel Annotation - Oncotator | (Ramos et al., 2015) | PMID: 25703262 |
| Mutation Significance Analysis - MutSigCV | (Lawrence et al., 2014) | PMID: 24390350 |
| RNA,DNaseq classifier - BioBloomTools(v1.2.4.b) | (Chu et al., 2014) | PMID: 25143290 |
| Broad Institute - PathSeq | (Kostic et al., 2011) | PMID: 21552235 |
| RNA read assembly – MapSplice 0.7.4 | (Wang et al., 2010) | PMID: 20802226 |
| Gene expression quantification - RSEM | (Li and Dewey, 2011) | PMID: 21816040 |
| Copy number estimation | NA | http://archive.broadinstitute.org/cancer/cga/copynumber_pipeline |
| Significant focal copy number change – GISTIC 2.0 | (Mermel et al., 2011) | http://software.broadinstitute.org/software/cprg/?q=node/31 |
| Purity, ploidy, genome doubling - ABSOLUTE | (Carter et al., 2012) | http://archive.broadinstitute.org/cancer/cga/absolute |
| Cluster analysis - ConsensusClusterPlus | | http://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html |
| Mbatch (EB++) | NA | http://bioinformatics.mdanderson.org/main/TCGABatchEffects:Overview |

## CONTACT FOR RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Peter W. Laird (Peter.Laird@vai.org). Sequence data hosted at the GDC is under controlled access. Details for gaining access can be found at (https://gdc.cancer.gov/access-data/data-access-processes-and-tools).

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Human Subjects and Tumor Data Selection

Molecular data were obtained as part of the Cancer Genome Atlas Project, from patients untreated by chemo- or radiation therapy and who provided informed consent; tissue collection was approved by the local Institutional Review Boards (IRBs) as noted below. GIAC cases (n=921) were selected as follows. Of the 559 Upper GI cases (171 ESCA and 388 STAD) in (Cancer Genome Atlas Research Network, 2017), 90 were excluded as ESCC and two as undifferentiated (TCGA-2H-A9GQ, TCGA-VR-A8Q7). Of the remaining 467 Upper GI adenocarcinomas, 462 (79 ESCA, 383 STAD) cases have molecular data available from the five TCGA core platforms (RNASeq, miRNASeq, DNA Methylation, SNP6, and mutation calls). We used germline DNA from blood or non-malignant gastrointestinal tissue as a reference for detecting somatic alterations. For lower GI, all available TCGA COAD and READ cases were considered, but cases bearing the BCR annotation ''Redacted'' were excluded, as were cases with Notification: 'Unacceptable Prior Treatment' or 'Item does not meet study protocol'. Review of COAD and READ pathology reports led to the exclusion of three additional COAD cases from this study (TCGA-AA-A022: Pathology report indicates poorly-differentiated carcinoma of the neuroendocrine type; TCGA-AA-A02R: Pathology report shows poorly-differentiated carcinoma with positivity for both S-100 and chromogranin, and focal synaptophysin; and TCGA-AZ-6607: Pathology report indicates this is likely to be a pancreaticobiliary primary tumor metastasizing to colon. The remaining 459 lower GI cases (341 COAD and 118 READ) with molecular data available for the five platforms were retained.

A group of 2,871 non-GIAC cases was constructed from TCGA tumor types BRCA, CESC, CHOL, LUAD, OV, PAAD, PRAD and UCEC, comprising all cases meeting the established criteria of the PanCancer Atlas Consortium (exclusion of Redacted,'Unacceptable Prior Treatment' or 'Item does not meet study protocol' and cases with no molecular data). For BRCA, CHOL, PRAD, and OV, and UCEC cases annotated as problematic by Expert Pathology Review (marked as AWG_excluded_because_of_pathology in the PanCancerAtlas Merged Annotation File) were excluded. For CESC, LUAD, and PAAD, further exclusions were made based on case review, as follows: CESC, retain only adenocarcinomas; LUAD, exclude samples without histology; PAAD, exclude samples with cellularity < 20%.

Demographic data for patients are as follows: GIAC (60.3% male, median age 68 years, range 29-90 years); Non-GIAC (21.3% male; median age 61 years, range 25 to 90 years).

TCGA Project Management collected necessary human subject documentation to ensure the project complies with 45-CFR-46 (the ''Common Rule''). The program has obtained documentation from every contributing clinical site to verify that IRB approval has been obtained to participate in TCGA. Such documented approval may include one or more of the following:

- An IRB-approved protocol with Informed Consent specific to TCGA or a substantially similar program. In the latter case, if the protocol was not TCGA-specific, the clinical site PI provided a further finding from the IRB that the already-approved protocol is sufficient to participate in TCGA.
- A TCGA-specific IRB waiver has been granted.
- A TCGA-specific letter that the IRB considers one of the exemptions in 45-CFR-46 applicable. The two most common exemptions cited were that the research falls under 46.102(f)(2) or 46.101(b)(4). Both exempt requirements for informed consent, because the received data and material do not contain directly identifiable private information.
- A TCGA-specific letter that the IRB does not consider the use of these data and materials to be human subjects research. This was most common for collections in which the donors were deceased.

## METHOD DETAILS

### Sample Processing

RNA and DNA were extracted from tumor and adjacent normal tissue specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a mirVana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp blood midi kit (Qiagen). Each specimen was quantified by measuring Abs260 with a UV spectrophotometer or by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify tumor DNA and germline DNA were derived from the same patient. Five hundred nanograms of each tumor and normal DNA were sent to Qiagen for REPLI-g whole genome amplification using a 100 μg reaction scale. Only specimens yielding a minimum of 6.9 μg of tumor DNA, 5.15 μg RNA,

and 4.9 μg of germline DNA were included in this study. RNA was analyzed via the RNA6000 Nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only the cases with RIN >7.0 were included in this study.

## Pathology Review

All samples were systematically evaluated by gastroenterological pathologists to confirm the histopathologic diagnosis and any variant histology according to the most recent World Health Organization (WHO) classification(International Agency for Research on Cancer, 2010). All tumor samples were assessed for tumor content (percent tumor nuclei), Tumor samples were evaluated for the presence and extent of inflammatory infiltrate as well as the type of the infiltrating cells in the tumor microenvironment (lymphocytes, neutrophils, eosinophils, histiocytes, plasma cells). Any non-concordant diagnoses among the pathologists were re-reviewed and resolution achieved after discussion.

## DNA Sequencing Data

Exome capture was performed using Agilent SureSelect Human All Exon 50 Mb according to the manufacturers' instructions. Briefly, 0.5–3 micrograms of DNA from each sample were used to prepare the sequencing library through shearing of the DNA followed by ligation of sequencing adaptors. All whole exome (WES) and whole genome (WGS) sequencing was performed on the Illumina HiSeq platform. Paired-end sequencing (2 x 101 bp for WGS and 2 x 76 bp for WE) was carried out using HiSeq sequencing instruments; the resulting data were analyzed with the current Illumina pipeline. Basic alignment and sequence QC were done with the Picard and Firehose pipelines at the Broad Institute. Sequencing data were processed using two consecutive pipelines: **(1) Sequencing data processing pipeline ("Picard pipeline")**. Picard (http://picard.sourceforge.net/) uses the reads and qualities produced by the Illumina software for all lanes and libraries generated for a single sample (either tumor or normal) and produces a single BAM file (http://samtools.sourceforge.net/SAM1.pdf) representing the sample. The final BAM file stores all reads and calibrated qualities along with their alignments to the genome.

**(2) Cancer genome analysis pipeline ("Firehose pipeline")**. Firehose (http://www.broadinstitute.org/cancer/cga/Firehose) takes the BAM files for the tumor and patient-matched normal samples and performs analyses including quality control, local realignment, mutation calling, small insertion and deletion identification, rearrangement detection, coverage calculations and others as described briefly below. The pipeline represents a set of tools for analyzing massively parallel sequencing data for both tumor DNA samples and their patient-matched normal DNA samples. Firehose uses GenePattern (Reich et al., 2006) as its execution engine for pipelines and modules based on input files specified by Firehose. The pipeline contains the following steps:

### *Quality Control*

This step confirms identity of individual tumor and normal to avoid mix-ups between tumor and normal data for the same individual.

### *Local Realignment of Reads*

This step realigns reads at sites that potentially harbor small insertions or deletions in either the tumor or the matched normal, to decrease the number of false positive single nucleotide variations caused by misaligned reads.

*Identification of Somatic Single Nucleotide Variations (SSNVs)*. This step detects candidate SSNVs using a statistical analysis of the bases and qualities in the tumor and normal BAMs, using Mutect (Cibulskis et al., 2013).

*Identification of Somatic Small Insertions and Deletions*. In this step, putative somatic events were first identified within the tumor BAM file and then filtered out using the corresponding normal data, using Indellocator (Ratan et al., 2015).

## Mutation Data

A series of quality-control filters according to the MC3 MAF were applied to mutations: (1) A filter for artificial CC>CA mutations caused by sample oxidation (8-oxoguanine) was applied to remove potential CC>CA artifacts (Costello et al., 2013); (2) Variants that were frequently observed in the Exome Aggregation Consortium (http://exac.broadinstitute.org) were excluded; (3) mutations with evidence of strand bias were excluded; (4) mutations with "ndp" labels were excluded; (5) duplicated mutations due to redundant tumor or normal samples were excluded. Somatic mutation calling was focused on coding mutations spanning missense and nonsense mutations, in-frame and frame-shift indels, and mutations that occurred on splice site, start codon, or stop codon.

The MutSig2CV (Cancer Genome Atlas Research Network, 2011) was applied to the quality-controlled mutation data to evaluate significance of mutated genes and estimate mutation densities of samples. MutSig2CV combines evidence from background mutation rate, clustering of mutations on hotspots and conservation of mutated sites to calculate the false discovery rates (q values). Genes of q value < 0.1 were declared significant.

## Microsatellite Instability

DNA samples were evaluated for Microsatellite Instability using the MSI-Mono-Dinucleotide assay, which examines four mononucleotide repeat loci (polyadenine tracts BAT25, BAT26, BAT40 and transforming growth factor receptor type II) and three dinucleotide repeat loci (CA repeats in D2S123, D5S346 and D17S250).

## Somatic Copy Number Alterations

DNA from each tumor or germline sample was hybridized to Affymetrix SNP 6.0 arrays using protocols from the Genome Analysis Platform of the Broad Institute as previously described (McCarroll et al., 2008). From raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus (Korn et al., 2008). For each tumor, genome-wide copy-number estimates were refined

using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor. This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then underwent segmentation using Circular Binary Segmentation (Olshen et al., 2004). Segmented copy-number profiles for tumor and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy-number changes underlying each segmented copy number profile, and the analysis of broad copy-number alterations was then conducted as previously described (Mermel et al., 2011). Significant focal copy-number alterations were identified from segmented data using GISTIC 2.0 (Mermel et al., 2011). Allelic copy number, regions of homozygous deletions, whole genome doubling and purity and ploidy estimates were calculated using the ABSOLUTE algorithm (Carter et al., 2012).

Copy ratios of the genomic segments were adjusted by purity and ploidy using the In Silico Admixture Removal (ISAR) method (Carter et al., 2012). The tumor purity and ploidy were estimated with ABSOLUTE (Absolute quantification of somatic DNA alterations in human cancer (Carter et al., 2012). GISTIC 2.0 (Mermel et al., 2011) was used to identify significant genomic regions, and q values that were smaller than 0.1 were considered significant. The gene under selective pressure in each significant amplification/deletion peak was manually curated with consideration of the common fragile sites (CFS). The gene-level copy numbers were obtained from GISTIC, and the gene was considered as amplified or deleted if the gene-level copy number change (ploidy-adjusted) was larger than 2 or smaller than -1.3, respectively. Whole-genome doubling (WGD) calls, absolute allelic copy numbers, and clonal statuses of the SCNAs were all obtained from ABSOLUTE.

### Aneuploidy Scores

The aneuploidy scores were calculated to quantify various kinds of aneuploidy in terms of length and magnitude of the copy-number events including segment gains and losses. The aneuploidy scores in this study were obtained as follows: (1) the original copy ratios of the genomic segments were adjusted by purity and ploidy using the ISAR method as noted above; (2) GISTIC 2.0 was used to deconstruct the ISAR-adjusted copy-number profile into SCNA events (discrete copy-number alterations), and each SCNA event could be categorized based on its length and magnitude (with details below); (3) for each category of SCNA events, e.g., focal amplifications, the corresponding aneuploidy score was calculated as $\log_{10}(1 + n)$, where $n$ is the number of events in that category. Similar approaches to the aneuploidy scores in principle were applied in a recent study (Davoli et al., 2017) as well as in our previous study (Dulak et al., 2012). The categories of SCNA events were defined as (1) Arm-level events: the relative SCNA length (as a proportion to the arm length) $l_{arm} \geq 0.5$, and the absolute value of amplitude $|m|>0.3$, and the threshold of 0.3 was applied to remove low copy ratio changes that were likely noise; (2) Focal events: $l_{arm}<0.5$, $|m|>0.3$; (3) Focal amplifications: $l_{arm}<0.5$, $m>0.3$; (4) Focal deletions: $l_{arm}<0.5$, $m<-0.3$; (5) High-level focal amplifications: $l_{arm}<0.5$, $m>1$; (6) Deep-level focal deletions: $l_{arm}<0.5$, $m<-1$. This method serves as a quantification of different types of genomic aneuploidy, and it is different from the gene-level amplification and deletion mentioned above, where conservative thresholds (2 and -1.3) for the gene-level copy number (not SCNA events) were applied to define functional alterations of the genes.

### CIN-Focal Score

We developed a CIN-Focal (CIN-F) score to capture the most focal high-level amplicons (MFAs), which are likely to be functional gains of specific genomic regions that were subject to positive selection during cancer evolution. Based on the deconstructed copy-number events from GISTIC 2.0, we defined those MFAs as $l<3$ Mb and $m>2$, where $l$ is the length of the amplicon in megabases, and $m$ is the event amplitude as mentioned above. Given each of those amplicons, the CIN-F score of a tumor was first calculated as the weighted sum of the magnitude $m_a$ of each amplicon $a$ (weighted by its length $l_a$), and then log-transformation was applied:

$$S_{\text{CIN-F}} = \log_{10}\left(1 + \sum_a l_a \cdot m_a\right)$$

Because $m_a$ is the ploidy-adjusted amplitude of copy-number gain, $l_a \cdot m_a$ is theoretically proportional to the relative amount of DNA (compared to the total cancer DNA) of the amplicon $a$, so that the CIN-F score corresponds to the amount of additional DNA within the MFAs. An alternative metric to CIN-F score is simply the total number of MFAs in a genome regardless of the lengths and amplitudes of the MFAs. The CIN-F score showed a binomial distribution in the upper GI cancers. We used kernel density estimation of Gaussian kernels (R statistical software, the ''density'' function) to set the threshold for dichotomization at the local minimum of estimated density of the CIN-F score, and this analysis yielded a threshold of $S_{\text{CIN-F}} = 0.438$. The CIN tumors was then dichotomized into CIN-F and CIN-B as shown in Figure 4C.

### Clonal Deletion Score (CDS)

To identify tumors with chromosomal instability, we developed a score, termed the Clonal Deletion Score, or CDS, which quantifies the number of clonally deleted genomic regions in each tumor's genome. The CDS of each tumor was calculated using absolute allelic copy numbers of genomic segments of the tumor. For each genomic segment, the absolute allelic copy numbers are denoted

as $q_1$ and $q_2$ for the two alleles with lower and higher copy number, respectively. If (1) the segment is a deletion, i.e., $q_1<q_2$, and $q_1 + q_2<\tau$, where $\tau$ is the average tumor ploidy; and (2) the deletion is clonal, i.e., $q_1$ is a clonal copy number according to ABSOLUTE; then the clonal deletion effect (CDE) of the segment is calculated as:

$$\mathrm{CDE} = 2\left(1 - \frac{q_1 + q_2}{\tau}\right)$$

If a segment does not satisfy the above criteria, the CDE of that segment is zero. The copy number of the higher allele $q_2$ was incorporated so as to diminish the CDE when there was a gain of the higher allele, e.g., copy-neutral loss of heterozygosity (LOH). Given the CDE of each segment $s$, the CDS of a tumor is the average of CDE weighted by the lengths of the segments:

$$\mathrm{CDS} = \sum_s w_s \cdot \mathrm{CDE}_s,\ w_s = \frac{l_s}{\sum_s l_s}$$

where $l_s$ is the length of a segment. The CDSs from the GI adenocarcinomas showed a clear bimodal separation. The kernel density estimation approach as mentioned above was used to set the threshold for dichotomization of CDS. A threshold of CDS = 0.0249 was then applied for the binary CIN/GS classification (Figures 2B and S2B), which corresponds to distinct copy-number profiles as shown in Figure S2C.

### Mutational Signatures

Mutational signatures were identified from SNVs using a Bayesian version of the non-negative matrix factorization method as described previously (Kim et al., 2016). The mutations were deconvoluted into distinct mutational signatures based on the number of mutations partitioned by 6 base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) and 16 possible combinations of neighboring bases that resulted in 96 possible types of mutations. A 96-by-M matrix of mutation counts (M is the number of samples) was constructed as the input data for signature discovery. Cosine similarity was used to evaluate the resemblance of the identified signatures with the COSMIC signatures (http://cancer.sanger.ac.uk/cosmic/signatures). For each sample, the estimated number of mutations from a signature was used as the intensity of that signature. A two-stage strategy of mutational signature discovery was performed in this study to achieve more accuracy in the identification of signatures. In the first stage, all samples were used to identify the signatures. In the second stage, the analysis was performed only for the non-hypermutated cases with the MSI and POLE signatures removed from the mutation counts to facilitate identification of signatures in the non-HM population.

### Stemness Index

We used one-class logistic regression (Sokolov et al., 2016) to derive a stemness index based on a gene expression signature derived from embryonic and differentiated cells from the PCBC dataset (Daily et al., 2017; Salomonis et al., 2016) and applied this to GIAC samples using Spearman correlations between the model's weight vector and the GIAC sample's expression profile (Malta et al., 2018).

### Differential Gene Expression Analysis between GIAC and Non-GI AC

To identify genes differentially abundant in GIAC versus non-GI AC, excluding genes that are differentially expressed between normal GI tissue compared to normal non-GI tissue, we needed to use external gene expression data from normal tissues. We selected 4 gastrointestinal (esophagus, stomach, colon-transverse, and colon-sigmoid) and 5 non-gastrointestinal (breast, lung, ovary, prostate, and uterine) normal tissue types through GTEx repository of normal tissues (GTEx Consortium et al., 2017) (https://www.gtexportal.org/home/datasets, GTEx Version 7), and utilized their RNA-sequencing expression dataset. Normalized expression values for both TCGA tumor and GTEx normal tissue cases were calculated by robust scaling (on values between 2.5 and 97.5 percentile) and winsorizing of each gene's expression (mean ± 3 standard deviations) in the respective case population of tumoral or normal cases. Gastrointestinal and non-gastrointestinal normal tissues were selected based on the matching with composition of available GI and non-GI adenocarcinomas in TCGA PanCancer project. Orthogonal partial least squares-discriminant analysis (OPLS-DA) was used to discover a subgroup of genes (n=671) that were not differentially expressed in GI and non-GI normal tissues, but were members of our list of differentially expressed genes between GI adenocarcinomas (GIAC) and non-GI adenocarcinomas (Non-GI AC). Significance was determined by absolute loading in the OPLS discriminant analysis of higher than 0.05. The genes for which expression was highly associated with the stromal class of GI tumors identified by the method described in Isella et al. (Isella et al., 2015) (n=118) were excluded from further analysis (absolute loading higher than 0.05). By utilizing an OPLS-DA model comparing GIAC and non-GI AC cases, the remaining 553 genes were ranked by their loadings toward overexpression in GIAC. Results were depicted in two heatmaps illustrating the normalized expression values for the selected genes in both GIAC and non-GI AC tissues, and normal GI and non-GI tissues, respectively (Figure S1G).

### Selection of Transcription Factors for Gain- and Loss-of-function Studies

We used multiple sources to select 139 transcription factors (TFs) that are important in GI development. We first identified 40 TFs in the Gene Ontology (GO) database based on the intersection of two GO terms, *RNA polymerase II transcription factor activity, sequence-specific DNA binding* (GO: 0000981) and *digestive tract development* (GO: 0048565), in *Homo sapiens*. Further, we

collected 24 TFs from the review by Noah et al. on human intestinal development and differentiation (Noah et al., 2011). Additionally, 93 genes were identified in the study in which Sherwood and colleagues used microarray and dynamic immunofluorescence technologies to profile gene expression during mouse endodermal organ formation (Sherwood et al., 2009). Finally, we also included nine other TFs that were significantly mutated in GIAC. In all, we examined 139 genes (taking the union of the four gene lists and removing genes with missing platform data).

## DNA Methylation Data

Illumina Infinium DNA methylation arrays [including both HumanMethylation27 (HM27) and HumanMethylation450 (HM450)] were used to assay 921 GIAC and 76 adjacent non-malignant tissues. Level 3 data from two generations of Illumina infinium DNA methylation arrays were combined and further normalized between platforms using a probe-by-probe proportional rescaling method as outlined below to yield a final common set of 22,601 probes with comparative methylation levels between platforms. During data generation, a single technical replicate of the same cell line control sample from either of two different DNA extractions (TCGA-07-0227/TCGA-AV-A03D) was included on each plate as a control, and measured 44/198 times and 12/169 times on HM27 and HM450, respectively. These repeated measurements were therefore used for rescaling of the HM27 data to be comparable to HM450. For each probe within each platform, we computed the median β value across all technical replicates of each of the two TCGA IDs. We then combined the two extractions by taking the mean of the two medians obtained for each of the two replicate TCGA IDs, and obtained a single summarized DNA methylation read-out (β value) for the corresponding probe i for each platform, noted as $\overline{Beta}_{hm27,i,}$ and $\overline{Beta}_{hm450,i,}$ respectively. We then applied a constrained (within the range of 0 to 1 for β values) linear rescaling of the HM27 data for each probe and for each patient's sample using $\overline{Beta}_{hm27,i}$ and $\overline{Beta}_{hm450,i}$. When the HM27 β value of a patient's sample *j* for probe *i* was smaller than the mean of median replicate samples on the HM27 for that probe, we linearly rescaled the HM27 β value $Beta_{hm27,i,j}$ in the (0, $Beta_{hm27,i,j}$ ) space; and when $Beta_{hm27,i,j}$ was greater, we linearly rescaled the HM27 beta value $Beta_{hm27,i,j}$ in the ($Beta_{hm27,i,j}$ ,1) space; This translates into the following mathematical computation: $Beta_{hm450,i,j} = Beta_{hm27,i,j} * (\overline{Beta}_{hm450,i}/\overline{Beta}_{hm27,i})$, if $Beta_{hm27,i,j} < \overline{Beta}_{hm27,i}$; and $Beta_{hm450,i,j} = 1 - (1 - Beta_{hm27,i,j}) * ((1 - \overline{Beta}_{hm450,i})/(1 - \overline{Beta}_{hm27,i}))$, if $Beta_{hm27,i,j} > \overline{Beta}_{hm27,i}$.

After the between-platform normalization, we further excluded 779 probes that still showed a consistent platform difference (mean β value difference greater than or equal to 0.1) in six or more tumor types.

## Unsupervised Clustering Analysis of DNA Methylation Data

Unsupervised clustering analyses of DNA methylation data were performed based on promoter CpG sites that did not exhibit tissue-specific DNA methylation in normal tissues and blood cells (mean β value < 0.2 for each tissue type), but acquired methylated in tumors.

### *GIAC and Non-GI AC*

We analyzed DNA methylation profiles of 3,759 adenocarcinomas including 921 GI adenocarcinomas and 2,828 non-GI adenocarcinomas representing 12 disease types (four GIAC and eight non-GI AC) (Figure S1A). We also included data from 333 histologically normal tumor-adjacent tissue specimens corresponding each disease type (BRCA n=101, PRAD n=39, OV n=12, CEAD n=1, UCEC n=43, EAC/GAC n=33, COAD n=37, READ n=6, CHOL n=9, PAAD n=10, LUAD n=42). We first used the data from the normal tissues and leukocytes to select CpG sites that lacked tissue-specific DNA methylation (mean β value < 0.2 in any tissue type and β value >0.3 in no more than five samples across the entire set). We then performed clustering analysis of the adenocarcinomas using 2,783 CpG sites that were hypermethylated (β value ≥0.3) in more than 10% within any of the 12 disease types. To minimize the influence of tumor purity on a clustering result, we dichotomized the data using a β value of ≥0.3 to define positive DNA methylation and < 0.3 to specify lack of methylation. We applied hierarchical clustering with Ward's method to cluster the distance matrix computed with the Jaccard index. Heatmap was generated based on the original β values for 1,000 loci (a subset of 2,783 loci) with the highest standard deviation in DNA methylation measurements among all adenocarcinomas.

### *GIAC (Figure 2D)*

We analyzed DNA methylation profiles of 921 GIAC and 77 (33 gastric and 44 colorectal) histologically normal tumor-adjacent tissue specimens. The precise locations within the GI organs from which the normal-adjacent tissue specimens were excised are not available. Unsupervised clustering of GIAC was performed based on 2,845 gene promoter loci unmethylated in normal tissues and leukocytes (mean β value < 0.2 in both normal gastric and colorectal tissues) but methylated (β value > 0.3) more than 5% in at least one of the GIAC tumor types. To minimize the influence of tumor purity, we dichotomized the data into 0's and 1's using a β value threshold of 0.3. The optimal number of clusters was assessed based on 80% probe and tumor resampling over 1,000 iterations of hierarchical clustering for K = 2, 3, 4…20 using the binary distance metric for clustering and Ward's method for linkage as implemented in the R/Bioconductor ConsensusClusterPlus package. The heatmap was generated using the original β values. The probes were displayed based on the order of unsupervised hierarchal clustering of the β values using the Euclidean distance metric and Ward's linkage method.

### *The Union of MSI and CIMP-H GIAC (Figure 3B)*

We used 158 tumors (93 GEA and 65 CRC) that were classified as either CIMP-H or MSI and 44 normals (12 stomach and 32 colorectal), which were assayed on the HM450 platform. Unsupervised and dichotomized clustering was performed using 35,436 sites

lacking DNA methylation in normal tissues (mean β value < 0.2 in both normal gastric and colorectal tissues) and methylated (β value > 0.3) more than 10% in any of the tumor type. Heatmap was generated based on the top 10% of the most variably hypermethylated sites across 158 GIAC.

### GIAC DNA Hypermethylation Subtypes

We chose seven GIAC DNA methylation clusters defined by the consensus clustering. For further integrative analyses, we focused on four prominent clusters showing a high frequency of cancer-associated DNA hypermethylation. We found that the gastroesophageal (GEA) and colorectal adenocarcinomas (CRC) largely clustered separately. Among GEA, $EBV^{+}$ gastric cancers stood out from all the rest by their extensive DNA hypermethylation (cluster 4) and were designated as **EBV-CIMP** as in the previous study (Cancer Genome Atlas Research Network, 2014). Cluster 5 is significantly enriched for MSI tumors originating in both stomach and colon. It included well-known CIMP-High CRC associated with $BRAF^{V600E}$ mutations and MSI-associated Gastric-CIMP described previously (Cancer Genome Atlas Research Network, 2014; Weisenberger et al., 2006). We classified these tumors as **GIAC CIMP-H**, as having a higher prevalence of DNA hypermethylation than all the other clusters with the exception of EBV-CIMP. Further, we named cluster 6 as **CRC CIMP-L** that exhibited features consistent with CIMP-Low subtype previously described (Cancer Genome Atlas Research Network, 2012). It had a significant association with *KRAS* mutations ($p < 2.2 \times 10^{-16}$ [vs. CRC in other groups], Fisher's exact test). Among GEA, cluster 1 was enriched for esophageal tumors ($p = 8.0 \times 10^{-8}$ [vs. GEAs in other groups]), and also had a mean DNA hypermethylation frequency slightly higher than that in CRC CIMP-L and other GEA clusters (cluster 2 and 3). We specified these tumors as **GEA-CIMP-L**. These tumors showed frequent epigenetic silencing of tumor suppressor genes including *CDKN2A* and *MGMT* ($p = 1.5 \times 10^{-10}$ and $p = 1.5 \times 10^{-11}$, respectively, [vs. GEA clusters 2 and 3]).

### Identification of Epigenetically Silenced Genes

We used 775 GIAC and 44 adjacent non-malignant tissues assayed on the HM450 platform. Probes located within potential promoter regions (1500 bp flanking regions upstream and downstream of Transcription Start Sites (TSSs) of all transcripts annotated by UCSC) were examined for evidence of epigenetic silencing. We removed the CpG sites that were methylated in normal tissues and blood cells (mean β value > 0.2 for each tissue type). In order to remove the effect of tissue specificity on gene expression, we z-score-transformed log 2 gene expression data first within each cancer type. The z-scores were derived using the mean and standard deviation calculated with the unmethylated tumors only, defined as those with a β value of (0, 0.2). Samples across all the cancer types were then pooled. For each probe/gene pair, we chose the probes that exhibited epigenetic silencing with the following criteria: 1) at least 8 samples (>1% of all tumors) were observed with a β value of 0.3 or above (defined as the methylated group); 2) mean z-score of the methylated group was lower than -1.65; 3) FDR-corrected p value according to one-side *t*-test on z-scores was lower than 0.001 between the unmethylated and methylated groups. Probes surviving these steps were retained to call epigenetic silencing events based on DNA methylation profiles for each sample. If there were multiple probes associated with the same gene, a sample identified as epigenetically silenced at more than half the probes for the corresponding gene was also labeled as epigenetically silenced at the gene level.

*CDKN2A* epigenetic silencing calls were made using the exon-level RNA-seq data. *CDKN2A* DNA methylation status was assessed in each sample, based on the probe (cg13601799) located in the p16INK4 promoter CpG island. p16INK4 expression was determined by the log2(RPKM+1) level of its first exon (chr9: 21974403-21975132). The epigenetic silencing calls for each sample were made by evaluating a scatter plot showing an inverse association between DNA methylation and expression. For *RAD51C*, there was no common probe between HM27 and HM450 that was located in the promoter region. However, probe cg14837411 from HM27 and probe cg27221688 from HM450 were only 100bp apart, and both correlated with gene expression. Therefore, we combined them in determining the silencing status of this gene. Samples with a β value of 0.2 or above for either probe were designated as cases with epigenetic silencing.

### DNA Hypermethylation Frequency in GIAC and Non-GI AC

We identified a set of 13,809 CpG sites that were unmethylated in normal tissues and blood cells (mean β value < 0.2 for each tissue type). For each CpG locus, tumors with a β value of 0.3 or greater were designated as methylated, and tumors with a β value of lower than 0.3 were designated as unmethylated. We then calculated the percentage of loci that were methylated among the loci investigated in each tumor.

### Methods for Integrative Pathway Analysis

We evaluated somatic mutations and copy-number changes relevant to well-studied signaling pathways curated in previous TCGA publications. Oncogenic relevance was assessed using OncoKB, a knowledge base for the oncogenic effects of cancer genes, that is manually curated by researchers and physicians at Memorial Sloan Kettering (Chakravarty et al., 2017). Specifically, a mutation was counted and included in the diagrams if either (1) it had been reported as a recurrent alteration in COSMIC (Forbes et al., 2011) or (2) it had been labeled as oncogenic or likely oncogenic in OncoKB. Amplifications and deep deletions were based on GISTIC calls and reflect a change of more than half of the baseline gene copies. The actual list of oncogenic and likely oncogenic alterations is regularly updated based on the literature; the most recent version can be retrieved online from the OncoKB public website (www.oncokb.org) or visualized when viewing the data in the cBioPortal (www.cbioportal.org). For known oncogenes, only genetic alterations inferred to be activating were considered; for tumor suppressor genes, only alterations inferred to be inactivating were considered.

## QUANTIFICATION AND STATISTICAL ANALYSIS

We used Fisher's exact test for independence between two categorical variables throughout the analyses. Wilcoxon rank-sum test was performed for any independence test between a continuous variable and a binary categorical variable. For any test between two continuous variables or any association test that needed to be adjusted by covariates, a (multiple) linear model was fitted to evaluate the significance of coefficients, and analysis of variance was used to calculate the proportion of variance explained by each variable. Non-negative variables that were heavily right-skewed, which included the aneuploidy scores, CIN-F score, number of MFAs, and the intensities of mutational signatures, were log-transformed (with a pseudo-count of 1 added) for appropriate fitting of multiple linear models. For the association test between aneuploidy scores and BRCA signature, the arm-level score and focal score were simultenously included as explanatory variables in the multiple linear model. The association test between BRCA signature and PARP1 expression (log-transformed) was adjusted by the copy number of PARP1. The intensity of the CpG>TpG signature was modeled by multiple linear regression with explanatory variables of upper/lower GI, molecular subtype, age, and CIMP status as an ordinal variable. A logistic regression model was fitted when the response variable was binary. The test between the CIN-F score and clinical stage was performed using an ordered logit model as the clinical stage was considered an ordinal variable, and the p values were calculated using normal approximation. The association test between number of MFAs and the CRC stromal subtype was performed using negative-binomial regression that models the sparse number of MFAs, so as to increase statistical power. Cox regression was used for survival analysis to evaluate the significance of the variables. All statistical analyses in this study were performed using the R statistical software (https://www.r-project.org).

## DATA AND SOFTWARE AVAILABILITY

The raw data, processed data and clinical data can be found at the legacy archive of the GDC (https://portal.gdc.cancer.gov/legacy-archive/search/f) and the PancanAtlas publication page (https://gdc.cancer.gov/about-data/publications/pancanatlas). The mutation data can be found here: (https://gdc.cancer.gov/about-data/publications/mc3-2017). TCGA data can also be explored through the Broad Institute FireBrowse portal (http://gdac.broadinstitute.org) and the Memorial Sloan Kettering Cancer Center cBioPortal (http://www.cbioportal.org). Details for software availability are in the Key Resource Table.