Supplementary Methods and Results

Patients and Tissue Processing

GC and germline (adjacent normal tissue or blood) samples were collected from young (≤45 years) patients with DGC in Korea. Fresh-frozen tumors were obtained from surgical resection performed between 2003 and 2014. Samples were collected from members of National Biobank of Korea (Asan Bio-Resource Center, Keimyung Human Bio-Resource Bank, Biobank of Pusan National University, Biobank of Chonnam National University Hwasun Hospital, Biobank of Chungnam National University Hospital, and Ajou Human Bio-Resource Bank), which is supported by the Ministry of Health and Welfare; Resource Banks at Dong-A University Medical Center and Kosin University Gospel Hospital; and the National Cancer Center of Korea. All tumor samples were collected from patients who had no prior chemotherapy or radiotherapy. This study was approved by the National Cancer Center Institutional Review Board (NCCNCS-120581) and all patients signed Institutional Review Board–approved consent forms. Tumors were staged according to the 7th edition of the American Joint Committee on Cancer.

A total of 80 resected DGCs with adequate yield of DNA/RNA for genomic analysis were subjected to WES, RNA sequencing, and SNP6.0 array analysis in Korea (EODGC-WES, Supplementary Table 1). WES was performed in these 80 resected tumors and matched germline DNA samples (buffy coats [n = 78] or adjacent normal tissues [n = 2]) from the same patients. The WES and RNA sequencing data were then reprocessed and analyzed by the same centers at which the stomach cancer samples for TCGA study[1] were analyzed, thereby facilitating the cross-comparisons of these tumor sets.

To achieve a statistical power adequate for comparison analyses, we expanded EODGC-WES cohort by recruiting 29 additional EO Korean DGC samples, which included biopsy samples that had been too small for inclusion in the WES analyses (EODGC [n = 109], Figure 1*A*). Assuming proportions of *TP53* mutations of 0.44 and 0.28 for LODGCs and EODGCs, respectively, a sample size required to detect the difference of *TP53* mutations rates between the 2 cohorts was estimated at 109, with 2-sided α- and β-errors of .05 and .2, respectively. Fresh-frozen EODGC tumors were obtained from either surgical resection or endoscopic biopsy. Targeted sequencing and SNP6.0 array analyses were performed for these cases. Supplementary Table 1 shows the clinicopathologic characteristics of the 109 patients in the EODGC cohort, which includes 80 patients in the EODGC-WES cohort and 29 additional Korean EODGC cases. Fourteen (12.8%) patients had family history of GC in a first- or second-degree relatives (Supplementary Table 4). At a median follow-up of 33.7 months, median overall survival of the 109 patients in the EODGC cohort was 46.6 months (95% CI, 33.3–not reached).

Overall, a total of 216 cases (tumor and matched germline samples) were collected for this study from Korean DGC patients ≤45 years old. During the sample qualification process, a total of 107 cases were disqualified for the inadequate tumor (n = 65) or normal samples (n = 7), inadequate RNA/DNA quality (n = 32) or quantity (n = 3), or intestinal/mixed histology (n = 7). After sequencing analyses, 4 patients with pathogenic germline mutations in *CDH1* (n = 1), *TP53* (n = 2), or *ATM* (n = 1) were excluded (Supplementary Table 2).

EODGCs were compared with a control set of diffuse-type stomach cancers were collected from 115 Korean patients who were ≥46 years old at the time of diagnosis in the aforementioned hospitals in Korea between 2002 and 2014 (LODGCs) (Supplementary Table 1). LODGCs and EODGCs were similar with regard to the proportion of metastatic disease (American Joint Committee on Cancer stage IV) and tumor purity as estimated by the H&E evaluation of top slides. These 2 cohorts were different in sex distribution ($P = .0002$, χ^2) (Supplementary Table 1).

EODGC-WES tumors were also compared with LO (≥46 years) diffuse tumors that originated from TCGA study (LODGCs from TCGA [LODGC-TCGA]).[1] In LODGC-TCGA, 70% of patients were Caucasian, including 26% female, and median age of 63 years (range, 46–82 years). We compared the WES data between LODGC-TCGA and EODGC-WES (n = 80). The proportion of metastatic disease at presentation (American Joint Committee on Cancer stage IV) was 14.8%, and did not differ from 16.3% of EODGC-WES (P = NS, χ^2).

Sample collection and processing were basically performed according to the protocol of a central Biospecimen Core Resource of TCGA.[1] After a gross examination, the non-necrotic portions were excised from the tumor specimens by pathologists, and adjacent normal tissues were collected from the luminal side of the gastrectomy specimen at least 2 cm from the tumor border. To isolate genomic DNA and RNA from frozen tumor tissue samples, light microscopic examinations were performed on top slides by pathologists to analyze the tumor-rich area. Then, genomic DNA was extracted from macrodissected frozen tumor tissues using a DNeasy blood and tissue kit (Qiagen, Hilden, Germany). Genomic DNA was also extracted from buffy coats using the same kit. Total RNA was isolated from the same piece of the tumor that was used for DNA isolation and from adjacent normal tissue samples using a mirVana kit (Thermo Fisher Scientific, Waltham, MA). RNA sequencing was performed in 80 tumors and 65 adjacent normal tissue samples that were collected from patients in the EODGC-WES cohort. SNP6.0 array analyses were performed using 109 tumors that were collected from the EODGC cohort, and 88 tumors that were collected from the LODGC cohort. Targeted sequencing was performed for all tumors collected from the EODGC (n = 109) and LODGC cohorts (n = 115) for 10 mutations that were recurrent among 80 EODGC-WES tumors.

Whole Exome Sequencing and Somatic Mutation Calling

For EODGC-WES tumors (n = 80), a total of 1–3 μg of fragmented DNA was prepared to construct libraries using a

SureSelect Human All Exon Kit V4+UTR (Agilent Technologies, Inc, Santa Clara, CA) according to the manufacturer's protocol. Briefly, the qualified genomic DNA sample was randomly fragmented using an ultrasonicator (Covaris, Woburn, MA) followed by adapter ligation, purification, hybridization, and polymerase chain reaction (PCR). Captured libraries were subjected to an Agilent 2100 Bioanalyzer to estimate the quality and then were paired-end sequenced using HiSeq platforms (Illumina, San Diego, CA). Raw image files were processed using HCS1.4.8 for base-calling with the default parameters, and the sequences for each individual sample were generated as 101-bp paired-end reads. We then re-processed and analyzed the WES data at the Broad Institute.

We used MuTect and Indelocator to identify somatic SNVs and indels, respectively.[2,3] SNVs/indels were annotated using Oncotator.[4] A filter for artificial CC>CA mutations caused by sample oxidation (8-oxoguanine) was applied to remove potential CC>CA artifact. ReCapSeg and AllelicCapSeg (http://www.broadinstitute.org/cancer/cga/acsbeta) were used to obtain allele-specific copy number profiles, and germline samples were aggregated as a reference panel for ReCapSeg to calculate normalized copy ratios. Given allele-specific copy ratios, the ABSOLUTE algorithm was used to estimate absolute copy numbers and tumor purity.[5] We used MutSig2CV to evaluate the significance of mutated genes, which combines evidence from background mutation rate, clustering of mutations on hotspots, and conservativeness of genes.[6] Mutations with q value <.1 were declared significant. Hypermutated samples (defined as a mutation rate >11.4/Mb) were removed from MutSig2CV analysis. None of the 80 EODGC-WES tumors were identified as hypermutated. We evaluated the degree of aneuploidy by determining the fraction of aneuploid genome in tumors. The fraction of aneuploid genome was defined as the relative length of the genome (to the length of human genome) that has integer copy number estimates not equal to 2 (or 4 for genome-doubled and 6 for genome-tripled cases), or was identified as loss-of-heterozygosity (LOH) regions. The same procedure was performed for TCGA data.

Comparison of the Whole Exome Sequencing Data Between Intestinal and Diffuse Tumors

We created a WES set from all of the non-hypermutated TCGA diffuse tumors and EODGC-WES (n = 147: diffuse tumors). Then we performed MuTect and Indelocator analyses on diffuse tumors, as described. According to the MutSig2CV analysis, *TP53, CDH1, ARID1A, KRAS, PIK3CA, ERBB3, TGFBR1*, and *RHOA* were significant mutations in 147 diffuse tumors. When the rates of these recurrent mutations were compared between 147 diffuse tumors and a set of non-hypermutated intestinal tumors originating from TCGA (n = 157), *TP53, CDH1,* and *TGFBR1* were significantly different in frequency between diffuse and intestinal tumors (Supplementary Figure 11*A* and Supplementary Table 7).

Mutation Signature Analysis

Mutation signatures were analyzed from SNVs identified in the WES data of EODGC-WES and LODGC-TCGA, using a Bayesian version of the non-negative matrix factorization method.[7] The mutation signature analysis is processed using the common classification of SNVs, which is based on 6 base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) with 16 possible combinations of neighboring bases result in 96 possible mutation types. We constructed input data for the mutation signature discovery, which is given as 96 by M mutation matrix (M = number of sample). The signature fraction, which is the estimated proportion of mutations that are contributed by each signature, was used as a metric for the intensity of a signature. Transcriptional strand bias of a signature was tested using logistic regression between the transcriptional strand of mutations and the corresponding signature fractions. We used the cosine similarity to compare our 3 signatures with 30 reported COSMIC signatures (http://cancer.sanger.ac.uk/cosmic/signatures).

In a combined WES dataset of EODGC-WES and LODGC-TCGA, we found 6 distinct mutation signatures present. As shown in Figure 7*A*, these 6 mutation signatures include 2 associated MSI (MSI signature and an alternative type of MSI signature [MSI-2 signature]), a BRCA/homologous recombination deficiency signature (BRCA signature), a signature of C>T transitions at CpG's sites and signatures similar COSMIC signature 17 (deamination (DA) signature), AA>AC-predominant signature seen previously in gastric and esophageal adenocarcinoma (G/E signature), and 1 resembling COSMIC signature 18 (http://cancer.sanger.ac.uk/cosmic/signatures; YF signature; Supplementary Figure 9*A*).

In EODGC-WES, the BRCA mutation signature was stronger in CIN (vs non-CIN, P = .008, Wilcoxon; Supplementary Figure 10*A*). In EODGC-WES, the fraction of BRCA mutation signature was not different between wild-type tumors and tumors with *CDH1* alteration (P = NS, Wilcoxon; Supplementary Figure 10*B*).

We could correlate mutation signatures with clinical response of 15 relapsed EODGC-WES cases to platinum (either cisplatin or oxaliplatin)-containing first-line chemotherapy. This analysis included patients who underwent palliative gastrectomy at metastatic stage and subsequently received chemotherapy without gross residual disease. No mutation signatures were significantly associated with progression-free survival after platinum-containing chemotherapy after relapse (Supplementary Table 10 and Supplementary Figure 10*C*).

We also asked whether other signatures may be preferentially present in patients with EODGC-WES, most especially in female patients, as such a signature may provide insight into the etiology of more frequent DGC diagnoses in this population. We noted that the signature resembling COSMIC signature 18 with prevalent C>A/T was significantly enriched in EODGC-WES (P = .0002, compared to LODGC-TCGA, Wilcoxon), especially in EODGC-WES females (P = .007, compared to EO males, Wilcoxon, Figure 7*D*). This mutation signature was therefore referred to as the YF signature. Although the etiology of signature 18 is not known, it is most strongly associated with the pediatric cancer, neuroblastoma. Further inquiry into YF signature demonstrated a significant strand bias of C>A/T mutations toward the transcribed strand (P = .001; Supplementary Figure 9*C*). Future studies into the etiology of this signature are highly warranted and may identify factors that contribute to formation of EODGC.

ReCapSeg and GISTIC Analyses

To evaluate SCNA significance, copy numbers were first adjusted for tumor purity and ploidy with the ISAR method,[8] and then GISTIC 2.0[9] was used to evaluate significances. A q value of .25 was set for the significance threshold. To eliminate spurious germline CNVs, we also performed the above ISAR-GISTIC on the matched normal samples and, for the next GISTIC run, we excluded those significant regions obtained from the normal samples. The procedures were repeated until there was no significant CNV in normal samples. Significant deletions and amplifications in EODGC-WES are shown in Figure 1*B* and Supplementary Figure 1*A*, respectively. Significant deletions and amplifications in EODGC-WES are shown in Figure 1*B* and Supplementary Figure 1*A*, respectively. ReCapSeg and GISTIC analyses of the WES data of diffuse tumors (n = 147) and non-hypermutated TCGA intestinal tumors (n = 157) are shown in Supplementary Figure 11*B*.

Targeted DNA Sequencing

Targeted sequencing was performed on samples from EODGC (n = 109) and LODGC (n = 115) cohorts for 10 mutations that were recurrent in the EODGC-WES cohort (Figure 1*E*).[6,10] Using the Ion AmpliISeq Designer software, PCR primers were designed for all exons (with 5 bp of padding at the ends of ends) with 98.9% coverage. PCR amplicons ranged from 125 to 175 bp in length. A total of 20 ng of genomic DNA with A260/A280 and A260/A230 ratios >1.6 was used for library generation. Fragment libraries were constructed using DNA fragmentation, barcode and adaptor ligation, and library amplification using an Ion DNA Barcoding kit (Thermo Fisher Scientific, Carlsbad, CA), according to the manufacturer's instructions. The size distribution of the DNA fragments was analyzed using a 2100 Bioanalyzer High Sensitivity Kit (Agilent). Template preparation, emulsion PCR, and ion sphere particle enrichment were performed using an Ion Xpress Template OT2 200 kit (v3: 4488318; Thermo Fisher Scientific), according to the manufacturer's instructions. Ion sphere particles were loaded onto a P1 chip (version 2) and sequenced using an Ion P1 sequencing 200 kit (v3: 4488315; Thermo Fisher Scientific). Ion Torrent platform-specific pipeline software (Torrent Suite, version 4.4) was used to separate the barcoded reads, generate sequence alignments with the hg19 human genome reference, perform target-region coverage analysis, and filter and remove poor signal reads. The single-nucleotide variants and small insertions/deletions (indels) were compared with those in the germline genomic DNA. Initial variant calling was generated using Torrent Suite with a plug-in program (Variant Caller, version 4.4). The alignment file from the Torrent Suite was then transferred to Ion Reporter version 4.4 to generate a somatic variant file using default parameters. The somatic calls generated from Ion Reporter version 4.4 were further filtered using the following criteria: (1) >50 reads in tumor samples; (2) >5 somatic variant reads; (3) somatic variant allele frequencies >0.05 and >0.1 for SNV and indels, respectively; and (4) minor allele frequency <0.02 in germline samples. A total of 346 somatic mutations (316 SNVs and 30 indels) were identified using targeted sequencing. EODGC (n = 109) and LODGC (n = 115) tumors were not different in coverage means (960× and 959x, respectively; P = .91, t test). Targeted sequencing and RNA sequencing validated 84.5% of somatic mutations that were identified in the WES data of EODGC-WES (Supplementary Table 3).

Germline Mutation Calling

We searched for germline mutations using the WES data for germline samples obtained from EODGC-WES patients. Germline variations in the normal samples were identified using GATK.[11] Information regarding the allele frequency in the Korean population was obtained from the Korean Reference Genome DB (http://152.99.75.168/KRGDB/menuPages/firstInfo.jsp). All pathogenic germline mutations were validated using capillary sequencing.

As listed in Supplementary Table 2, four germline mutations in *ATM, CDH1*, and *TP53* were identified as pathogenic germline mutations based on the Clinvar database,[12] the evidence as a somatic alteration,[13] and the data in the literature. One of these pathogenic germline mutations, an R335X *CDH1* nonsense mutation, was identified as a somatic mutation in an EOGC patient. These 4 cases with pathogenic germline mutations were excluded from this analysis.

Our targeted DNA sequencing analysis of EODGC and LODGC revealed no pathogenic germline mutations, consistent with the lack of strong familial clustering of GC among our EODGC patients. WES and targeted sequencing revealed several other *CDH1* germline mutations of unknown significance among EODGC and LODGC cases (Supplementary Table 2). We performed functional analyses of these *CDH1* variants to evaluate their oncogenic potentials. CHO aggregation assays indicated that K182N, T340A, and E880K were likely to be benign germline variants (Supplementary Figure 6*C*).

During our mutation signature analyses, we additionally searched for pathogenic germline mutations involved in DNA repair, which might have been missed by the methods described. We extracted mutations in a list of candidate genes that are involved in DNA repair according to the MSigDB gene sets.[14] Then, we filtered out mutations that are not likely to have significant functional impact on the genes, using a combination of prediction methods (SIFT, Polyphen2, LRT, MutationTaster, and MutationAssessor) that were included in the dbNSFP.[15] Mutations were filtered out when <3 of these methods predict them as functional.

Three pathogenic germline mutations involved in DNA repair were identified from this approach (Supplementary Table 2).

RNA Sequencing

Transcriptome libraries were prepared following Illumina's TruSeq mRNA kit protocol using 1–2 μg of total RNA obtained from 80 tumors and 65 adjacent normal tissue samples obtained from patients in the EODGC-WES cohort. Poly(A)+ RNA was isolated using AMPure XP beads (Beckman Coulter, Brea, CA) and fragmented using an Ambion Fragmentation Reagents kit (Thermo Fisher Scientific). Complementary DNA (cDNA) synthesis, end repair, A-base addition, and ligation of the Illumina-indexed adapters were performed according to Illumina protocols. Libraries were size-selected for 250–300 bp cDNA fragments on a 3% Nusieve 3:1 (Lonza, Basel, Switzerland) agarose gel, recovered using QIAEX II gel extraction reagents (Qiagen), and PCR-amplified using Phusion DNA polymerase (New England Biolabs, Ipswich, MA) for 14 PCR cycles. The amplified libraries were purified using AMPure XP beads. Library quality was measured on an Agilent 2100 Bioanalyzer to determine the product size and concentration. Paired-end libraries were sequenced using an Illumina HiSeq 2000 instrument (2×100 nucleotide read length). Reads that passed the chastity filter of the Illumina BaseCall software were used for subsequent analyses.

RNAseq data were aligned at the British Columbia Genome Science Center. Using BWA aln & sampe (version 0.5.7),[16] we aligned all reads to a human reference genome consisting of hg19/GRCh37-lite and exon–exon junction sequences that were constructed from known transcript models in EnsEMBL, RefSeq, and UCSC genes. Default BWA parameters were used for alignments, with the exception of the -s sampe option, which was included to disable Smith-Waterman rescue of unmapped mates, as this feature was not designed to handle the insert size distribution that occurs in paired-end RNA sequencing data. After BWA alignment, a post-alignment process was performed to reposition the read alignments that spanned across the exon–exon junctions and transform them into large-gapped genomic alignments.[17] The reads per kilobase of exon per million mapped reads values of exons and genes were calculated. GENCODE V3 was used in the quantification process.

In 80 EODGC-WES tumors and 65 adjacent normal tissue samples, median 5′ to 3′ coverage ratio was 0.78 (interquartile range [IQR], 0.68–0.85). Median coverage in exons was 97 (IQR, 84–123). Median percentage of total coverage in exons was 91.8% (IQR, 90.2%–93.2%). Median number of genes with at least 10× coverage was 15,510 (IQR, 14,757–16,300).

Chromosomal locus 3p21.1 (g.chr3: 48,369,660–55,002,466) was the most significantly deleted in EODGC-WES (q = .0002, Figure 1*B*). Among the 161 genes in this locus, 6 genes including *BAP1* were significantly ($q < .1$ and fold-change < -0.5) underexpressed in EODGC-WES tumors with 3p21.1 deletion (n = 6), when compared with EODGC-WES tumors without 3p21.1 deletion (n = 74) (Supplementary Table 8 and Figure 1*C*).

The deubiquitinase *BAP1* may be a novel tumor suppressor gene candidate for DGC, because *BAP1* mutation was the 16th most recurrent mutation in EODGC-WES (Supplementary Table 9), and all 4 *BAP1* mutations found in EODGC-WES were associated with LOH, suggesting a tumor suppressive role.

SNP6.0 Array Analysis

We performed a SNP6.0 array analysis on all of 109 tumors in the EODGC cohort and 88 of 115 tumors in the LODGC cohort, according to the Affymetrix Genome-Wide Human SNP Nsp/Sty 6.0 User Guide (Santa Clara, CA). There were no significant differences in clinicopathologic factors between LODGC cases with SNP6.0 data (n = 88) and those without SNP6.0 data (n = 27; Supplementary Table 4). Contrast QC medians for .cel files were 1.5 (IQR, 1.2–1.9) and 1.5 (IQR, 1.2–1.8), for EODGC and LODGC, respectively (P = NS, t test). Contrast QC (Nsp) medians were 1.5 (IQR, 1.2–1.9) and 1.5 (IQR, 1.2–1.8), for EODGC and LODGC, respectively (P = NS, t test). Contrast QC (Sty) medians were 1.3 (IQR, 0.8–1.8) and 1.7 (IQR, 1.4–2.0), for EODGC and LODGC, respectively (P = NS, t test). Genome-wide copy number estimates were refined using tangent normalization. Recurrent focal copy number alterations were identified using GISTIC 2.0.[9]

To determine TCGA subgroup of each sample, CIN status was determined based on Affymetrix SNP6.0 data, using a method described in a TCGA marker paper.[1] We performed the segmentation using the Circular Binary Segmentation method.[18] For copy number-based clustering, EODGC and LODGC tumors were subjected to hierarchical clustering along with TCGA tumors, CIN statuses of which were defined previously.[1] R-based hierarchical clustering was performed for thresholded copy number at reoccurring alteration peaks from GISTIC 2.0 analysis (all_lesions.conf_99.txt), using Euclidean distance and Wards method. The CIN status of new EODGC and LODGC tumors was determined based on the cluster membership (Supplementary Figure 5*B*).

All tumors were categorized into 1 of 4 subtypes (EBV-positive [EBV], MSI-high [MSI-H], GS, and CIN). CIN-positive tumors without EBV or MSI were assigned to the CIN subgroup, and CIN-negative tumors without EBV or MSI were assigned to the GS subgroup. If a tumor was both EBV-positive and CIN, that sample was assigned to the EBV group (Figure 6*A*).

Comparison of SNP6.0-Based Copy Number Profiles Between Early-Onset and the Late-Onset Diffuse Gastric Cancers

Copy number profiles based on SNP6.0 data were compared between EODGCs and LODGCs. Cross-validated misclassification rates were determined using arm-level copy number profiles of EODGC and the LODGC by class prediction algorithms built in the BRB-ArrayTools (version 4.4.1; National Cancer Institute, Bethesda, MD).

Permutation P values were then determined for cross-validated misclassification rates. Two age groups were considered different for arm-level copy number profiles if permutation P values were $<.05$. According to prediction analyses built in the BRB-ArrayTools, permutation P values were $>.05$ (*red dashed line*) with all of the class prediction algorithms at feature selection P values $<.05$. Permutation P values were .14, .15, .51, .14, .13, and .08 for the compound covariate predictor, the diagonal linear discriminant analysis classifier, the 1-nearest neighbor classifier, the 3-nearest neighbor classifier, the nearest centroid classifier, and the support vector machines classifier, respectively (Supplementary Figure 5*D*). Thus, EODGC and LODGC did not differ in arm-level copy number profiles.

In EODGC, a total of 21 chromosomal locations with peaks displayed significantly recurrent focal amplifications. The recurrent focal amplification at 10q26.13 at the locus gene encoding *FGFR2* was the dominant peak ($q = 1.76 \times 10^{-10}$). Significant amplifications were observed at 3q26.2, 6q21, 6q23.3, 6p24.1, 7q22.1, 8q24.21, 9p21.2, 10q26.12, 11p13, 12p12.1, 16p13.3, 17q12, 17q25.3, 18q11.2, 19q12, and 21q21.1. *CD44, ERBB2, GATA6, KRAS, MYC*, and *POU5F1B* were included in these loci. In the LODGC cohort, we found 23 chromosomal locations with peaks that were significantly recurrent focal amplifications. Similar to EODGC, 10q26.13 was a clearly dominant peak ($q = 1.75 \times 10^{-7}$). Significant amplifications were observed at 3q26.31, 7q11.21, 7q31.2, 7p11.2, 8q24.21, 8p23.1, 10q11.21, 11q13.3, 11p13, 12p12.1, 17q12, 18q11.2, 9p24.1, 19q12, and 20q13.2. *CD44, EGFR, ERBB2, FGFR2, GATA4*, and *GATA6* were present in these loci.

Six peaks, including 8q24.21 (*MYC*), 10q26.13 (*FGFR2*), and 17q12 (*ERBB2*), were amplified in both the EODGC and the LODGC cohorts. *FGFR* gene amplification was identified in 8.2% and 7.9% of EODGC and LODGC, respectively (P = NS, χ^2), suggesting ethnic difference. 3q26.2, 6q21, 6q23.3, 6p24.1, 7q22.1, 9p21.2, 10q26.12, 16p13.3, 17q25.3, 19q12, and 21q21.1 were recurrently amplified in EODGC only. *ERBB2* amplifications tended to be more frequent in EODGC than in LODGC (9.1% vs 3.4%, respectively), but the difference was not significant ($P = .1$, χ^2). The following loci were recurrently amplified only in the LODGC cohort: 3q26.31, 7q11.21, 7q31.2, 7p11.2, 8p23.1, 9p24.1, 10q11.21, 11q13.3, and 20q13.2. *EGFR* and *GATA4* were contained in these loci.

Using a silver in situ hybridization (ISH) analysis, we validated the gene amplification of *ERBB2*, which represents the only clinically actionable gene with Food and Drug Administration–approved treatment for GCs, in an independent dataset of DGCs collected from patients who were aged ≤45 years and treated at Asan Medical Center and the National Cancer Center of Korea. The slides were processed using the automated system following the manufacturer's protocols for INFORM *HER2* DNA and chromosome 17 (*CEP17*) probes (Ventana Medical Systems, Tucson, AZ).[19] Both probes were sequentially hybridized in 1 slide. *HER2* gene was visualized as a black dot and CEP17 as a red dot. The specimen was then counterstained with Harris hematoxylin. *HER2* gene amplification status was evaluated by counting *HER2* and *CEP17* signals in nuclei of 20 non-overlapping consecutive tumor cells with the ASCO/CAP guidelines; negative for *HER2* gene amplification if the *HER2/CEP17* ratio was <1.8, equivocal if the *HER2/CEP17* ratio was 1.8 to 2.2, and positive if the*HER2/CEP17* ratio was >2.2.[20,21] In equivocal cases, we counted 20 additional tumor cells and considered the case positive for *HER2* gene amplification if the *HER2/CEP17* ratio was ≥2.0.[20,21] Seventeen of 275 tumors in this validation dataset from patients aged ≤45 years (6.2%) demonstrated silver ISH evidences for *ERBB2* amplifications. The silver ISH positivity in this dataset (6.2%) was similar to the *ERBB2* amplification rate determined based on WES in the EODGC cohort (5.9%) and slightly lower than the rate determined based on SNP6.0 data in EODGC (8.3%).

Using immunohistochemistry, we validated *FGFR2* amplifications. *FGFR2*-amplified tumors (n = 9) showed significantly higher FGFR2 immunohistochemistry grades than did tumors without *FGFR2* amplification (n = 10) ($P = .036$, t test).

Microsatellite Instability Assay

MSI was evaluated using a panel of 2 mononucleotide repeat loci (the polyadenine tracts BAT25 and BAT26) and 3 dinucleotide repeat loci (CA repeats in D2S123, D5S246, and D17S250), as described previously.[22,23] Briefly, fluorescently labeled PCR and capillary electrophoresis were performed using tumor and matched normal tissue DNA samples. The primer sequences for the PCRs were as follows: F (FAM): 5′ -TCG CCT CCA AGA ATG TAA GT-3′ and R: 5′ -TCT GCA TTT TAA CTA TGG CTC-3′ for BAT-25; F (HEX): 5′ -TGA CTA CTT TTG ACT TCA GCC-3′ and R: 5′ -AAC CAT TCA ACA TTT TTA ACC C-3′ for BAT-26; F (FAM): 5′ -AAA CAG GAT GCC TGC CTT TA-3′ and R: 5′ -GGA CTT TCC ACC TAT GGG AC-3′ for D2S123; F (FAM): 5′ -GGA AGA ATC AAA TAG ACA AT-3′ and R: 5′ -GCT GGC CAT ATA TAT ATT TAA ACC-3′ for D17S250; and F (HEX): 5′ -ACT CAC TCT AGT GAT AAA TCG GG-3′ and R: 5′ -AGC AGA TAA GAC AGT ATT ACT AGT T-3′ for D5S246. The touch-up PCR was performed essentially as described previously[23] using HotStarTaq DNA Polymerase (Qiagen). After reaction mixtures with 1 μL of the 1/10 to 1/50-diluted PCR products, 8.5 μL of Hi-Di Formamide, and 0.5 μL of GeneScan 500 ROX Size Standard (Thermo Fisher Scientific) were prepared, denatured at 95°C for 2 minutes, and snap-cooled on ice, the microsatellite patterns were detected after automated capillary electrophoresis using an ABI3730 Genetic Analyzer (Thermo Fisher Scientific) with a POP-7 polymer (Thermo Fisher Scientific). Tumor was classified as high-level MSI (MSI-H) if ≥2 markers demonstrated unequivocal alterations of peak pattern compared with normal DNA.

EODGC (n = 109) and LODGC (n = 115) were significantly different in the frequency of MSI-H (4.6% vs 13.0% for EODGC and LODGC, respectively; $P = .027$, χ^2) (Supplementary Table 1).

Detection of the Epstein-Barr Virus

In EODGC, EBV positivity was determined using ISH and PCR. The presence of EBV in the cancer cells was assessed using EBV chromogenic ISH on an automatic staining device (Benchmark XT; Ventana, Tucson, AZ) according to the manufacturer's guidelines.[24] Briefly, 4-μm-thick sections were cut from representative blocks obtained from each patient, mounted onto coated slides, and dried at 74°C for 30 minutes. After pretreatment with ISH protease 2 (Ventana) for 8 minutes at 37°C, the slides were denatured at 85°C for 12 minutes and hybridization was conducted at 57°C for 1 hour using EBV-encoded small RNA probes (INFORM, Ventana) that had been labeled with fluorescein. Detection was sequentially performed by applying mouse antifluorescein antibody and biotinylated goat anti-mouse antibody (iView blue; Ventana); counterstaining was performed using Nuclear Fast Red. Obvious, strong staining in the nucleus was considered to indicate positivity, and the proportion of EBV-positive tumor cells was thereby evaluated. A representative sample with EBV-positive nuclei in cancer cells is shown in Supplementary Figure 8*G*.

EBV PCR was performed as described previously,[25,26] with some modifications. The PCR mixture (10 μL) contained 0.25 U HotStarTaq DNA Polymerase (Qiagen), 1 μL of 10× PCR Buffer, 200 μM of each dNTP, 0.3 μM of each primer, and 5 ng of genomic DNA from the tumor sample. The primer sequences were as follows[25]: 5′-CCA TGT AAG CCT GCC TCG AG-3′ and 5′-GCC TTA GAT CTG GCT CTT TG-3′. The cycling conditions were as follows: 95°C for 15 minutes followed by 30 cycles of 94°C for 40 seconds, 57°C for 30 seconds, and 72°C for 30 seconds, and a final extension of 72°C for 3 minutes. An unequivocal PCR band was considered to indicate positivity for EBV. In the EODGC cohort, the ISH results and PCR results were 100% concordant. Therefore, in the LODGC cohort, PCR was used to determine EBV status. EBV positivity was almost equivalent between 109 EODGCs and 115 LODGCs (6.2% vs 6.1%, respectively; Supplementary Table 1).

Detection of Helicobacter pylori Using Real-Time Polymerase Chain Reaction

Real-time PCR was performed using a LightCycler 480 (Roche Diagnostics) with 8 μL of reaction mixture containing 4 μL of 2× QuantiTect SYBR Green PCR Master Mix (Qiagen), 250 nM of each primer, and 5 ng of genomic DNA that was isolated from adjacent normal tissue samples. The cycling conditions were as follows: 95°C for 15 minutes, followed by 45 cycles of 95°C for 20 seconds, 58°C for 20 seconds, and 72°C for 20 s. The primers used were Hp23S 1835F (5′-GGT CTC AGC AAA GAG TCC CT-3′) and Hp23S 2327R (5′-CCC ACC AAG CAT TGT CCT-3′). Cp value <40 was interpreted as a positive result. In the EODGC cohort, 53 of 109 samples (48.6%) were positive for *H pylori*. No mutations or mutation signatures significantly correlated with *H pylori* positivity.

Reverse Transcription Polymerase Chain Reaction Analysis of CDH1 Splice Variants

cDNA was synthesized from 500 ng of total RNA using a SuperScript III First-Strand Synthesis System (Thermo Fisher Scientific), and was PCR-amplified for *CDH1*. The PCR mixture (20 μL) contained 0.5 U of HotStarTaq DNA Polymerase (Qiagen), 2 μL of 10× PCR buffer, 200 μM of each dNTP, 0.3 μM of each primer, and 0.3 μL of the synthesized cDNA. The cycling conditions were as follows: 95°C for 15 minutes followed by 40–45 cycles consisting of 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds. All *CDH1* splice site mutations were validated using capillary sequencing. The primer sequences were as follows: *CDH1* for exons 4 and 5, 5′-CCA TCA GGC CTC CGT TTC TGG-3′ and 5′-TGT CAT TCT GAT CGG TTA CCG TGA TC-3′; *CDH1* for exons 8 and 9, 5′-CTC AGC CAA GAT CCT GAG CTC-3′ and 5′-CTT CAT TCA CAT CCA GCA CAT CCA C-3′. All *CDH1* splice site mutations were validated using capillary sequencing and are summarized in Supplementary Table 4.

Bioinformatic Prediction of CDH1 Cryptic Splice Donor Sites and 3-Dimensional Computation Modeling for CDH1 Mutations

According to the Human Splicing Finder (http://www.umd.be/HSF/) public database,[27] chr16:g.68842471G>A and chr16:g.68842472T>C mutations were predicted to inactivate the endogenous splice donor site in the exon 4-intron 4 junction and to generate a novel cryptic splice donor site at c.468 (Supplementary Table 4). Reverse transcription PCR and Sanger sequencing validated the presence of a truncated transcript that lacked 63 bp (21 amino acids) exclusively in tumors with the corresponding mutations. As summarized in Supplementary Table 4, the same bioinformatics analysis predicted that the endogenous splice donor site in the exon 5–intron 5 junction was inactivated by mutations in chr16:g.68842752G>A, chr16:68842749 ACTG/A, and chr16:68842753del and that a cryptic splice donor site was created at c.645. As predicted, a novel splice variant lacking 42 bp (14 amino acids) was identified in reverse transcription PCR sequencing in tumors with the corresponding mutations. Homology modeling was performed using SWISS-MODEL.[27,28]

CDH1 Loss of Heterozygosity Analysis

LOH analyses were performed using 3 microsatellite markers (D16S4025, D16S396, and D16S4067) that flank the *CDH1* locus and 3 intragenic *CDH1* single-nucleotide polymorphism (SNP) markers (rs16260, rs1801552, and rs1801026), as described previously,[29,30] with modifications.

Microsatellite markers were evaluated as follows: a total of 100 ng of tumor and matched normal tissue DNA were used to perform a fluorescently labeled PCR with the following PCR conditions: an initial activation step of 95°C for 15 minutes, 28 cycles of 94°C for 30 seconds, 58°C for 30 seconds, and 72°C for 30 seconds, and a final extension of 72°C for 3 minutes; the following reaction composition: 0.25 U of HotStarTaq DNA Polymerase (Qiagen), 1 μL of 10× PCR buffer, 200 μM of each dNTP, 0.3 μM of each primer, and 5 ng of genomic DNA in 10 μL of mixture; and the following primer sets: D16S4025, 5′-TCC ATT GGA CTT ATA ACC ATG-3′ and 5′-AGC TGA GAG ACA TCT GGG-3′; D16S396, 5′-GAA AGG CTA CTT CAT AGA TGG CAA T-3′ and 5′-ATA AGC CAC TGC GCC CAT-3′; and D16S4067, 5′-GCC ACC TCA CAC TAG CCT G-3′ and 5′-TTC CAG CTC TCA CTC AAA ATG-3′. The PCR products were visualized using an ABI3730 Genetic Analyzer (Thermo Fisher Scientific). Data were analyzed using Peak Scanner software (version 1.0, Thermo Fisher Scientific), and the samples were determined to display LOH if the LOH index, calculated as follows, was >2 or <0.5.

LOH index = (N1/N2)/(T1/T2),

where N1 is peak areas of germline allele 1; N2 is peak areas of germline allele 2; T1 is peak areas of tumor allele 1; and T2 is peak areas of tumor allele 2.

All experiments involving microsatellite LOH-positive samples were repeated twice for validation.

SNP loci were evaluated as follows: an SNP locus in each tumor sample was considered to be informative when the corresponding germline sequence was heterozygous in capillary sequencing. Targeted sequencing was performed on each tumor genomic DNA sample as described here. The allele frequency of a given SNP locus was determined for each tumor sample. A SNP locus was identified to display LOH if the minor allele frequency was <0.4 in the tumor sample and if the locus was heterozygous in the matched germline sample.

At least 1 LOH marker was informative in 89 and 96 samples in the EODGC and LODGC cohorts, respectively. A total of 31 (34.8%) and 21 (21.9%) samples were determined to display LOH for *CDH1* in the EODGC and LODGC cohorts, respectively ($P = .05$, χ^2; Supplementary Table 4).

Real-Time Polymerase Chain Reaction for Genomic Copy of CDH1

A homozygous deletion of *CDH1* was determined using real-time PCR. The real-time PCR was performed using a LightCycler 480 (Roche Diagnostics, Mannheim, Germany) with 10 μL of reaction mixture containing 5 μL of 2× QuantiTect Probe PCR Master Mix (Qiagen), 200 nM of each primer, 100 nM of TaqMan probe, and 10 ng of genomic DNA. The cycling conditions were as follows: 95°C for 15 minutes, followed by 99 cycles of 95°C for 20 seconds, 58°C for 20 seconds, and 72°C for 20 seconds. Each sample was run in duplicate. The primer sequences used were as follows: *CDH1*, 5′-TTC TAC AGC ATC ACT GGC CAA G-3′ and 5′-AGT GTA TGT GGC AAT GCG TTC TC-3′; and *LINE1*, 5′-AAA GCC GCT CAA CTA CAT GG-3′ and 5′-TGC TTT GAA TGC GTC CCA GAG-3′. The probe sequences used were as follows: *CDH1*, 5′-FAM-AGC TGA CAC ACC CCC TGT TGG TGT C-BHQ-1-3′; and *LINE1*, 5′-FAM-CTG AAC AAC CTG CTC CTG AAT GAC TAC TG-BHQ-1-3′. A germline DNA sample (from buffy coat) was serially diluted with SK-BR3 (a cell line with *CDH1* homozygous deletion) DNA to generate a standard curve. A homozygous deletion was defined as a mean *LINE*-normalized $\log_2$ copy number ratio < -1.0. *CDH1* homozygous deletion was identified in 2 EODGC samples (Supplementary Table 4).

CDH1 *Promoter DNA Methylation Analysis*

CDH1 promoter DNA methylation analysis was performed to analyze 32 CpG islands within the 160-bp upstream and 153-bp downstream of the *CDH1* transcription start site. EpiTect Bisulfite Kit (Qiagen) and 100 ng of genomic DNA were used to convert unmethylated cytosines to uracil, whereas methylated residues remained unmodified. The bisulfite conversion reaction was performed as recommended by the manufacturer (Qiagen). Buffy coat DNA that was in vitro methylated using the M.SssI (New England Biolabs) was used as a positive control. The genomic region was PCR-amplified with the following composition: 0.5 U of i-StarTaq GH DNA Polymerase (iNtRON Biotechnology, Seongnam, Korea), 1 μL of 10× PCR buffer, 250 μM of each dNTP, 0.3 μM of each primer, and 1.5 μL of bisulfite-converted genomic DNA from 20 μL of the eluate. The cycling conditions were as follows: 95°C for 5 minutes followed by 60 cycles of 95°C for 30 seconds, 58°C for 30 seconds, and 72°C for 40 seconds, and a final extension of 72°C for 3 minutes. The following primer set was used: 5′-GAT TTT AGT AAT TTT AGG TTA GAG GGT TAT-3′ and 5′-AAA TAC CTA CAA CAA CAA CAA CAA C-3′.

Cloning of the PCR products and transformation of *Escherichia coli* were performed using TOPO-TA Cloning Kit (Thermo Fisher Scientific) as recommended by the manufacturer. The next day, colony PCRs were performed using the i-StarTaq GH DNA Polymerase with M13 primers. Methylation status of the CpG sites was analyzed using capillary sequencing with the colony PCR products. Tumors that showed methylation at ≥25% of the CpG sites (≥8 of 32 CpG sites) were defined as hypermethylated tumors.[30]

A total of 5 and 16 tumors were identified as hypermethylated tumors in EODGC and LODGC, respectively (4.6% vs 13.9%; $P = .017$, χ^2; Supplementary Table 4).

Immunohistochemistry

The GC tissues were fixed in 10% phosphate-buffered formalin, processed in a routine manner, and embedded in paraffin. IHC was performed on 4-μm-thick serial sections from whole tissue paraffin blocks. Immunohistochemical staining for E-cadherin was performed using an automated Discovery XT (Ventana Medical System, Tucson, AZ) instrument. Briefly, the paraffin sections were deparaffinized, rehydrated using EZ prep (Ventana Medical System), and then washed with reaction buffer (Ventana Medical System). The antigens were retrieved using heat in pH 8.0 Tris−EDTA buffer (CC1; Ventana Medical System) at 90°C for 30 minutes. The slides were incubated for 20 minutes at room temperature with a 4A2C7/monoclonal mouse anti−E-cadherin antibody (1:4000 dilution, 18-0223; Zymed/Thermo Fisher Scientific).[31] For immunohistochemistry of β-catenin, MYC, and FGFR2, an ImmPRESS Peroxidase Polymer kit (Vector Laboratories, Burlingame, CA) was used according to the manufacturer's protocol. Briefly, dewaxed and rehydrated paraffin sections were subjected to antigen retrieval by heating the sections to 100°C for 20 minutes in 0.01 M citrate buffer (pH 6.0). The slides were incubated with 2.5% horse serum during blocking and then incubated for 30 minutes at room temperature with 14/monoclonal mouse anti−β-catenin antibody[32,33] (1:400 dilution; 610154; BD Biosciences, San Jose, CA), Y69/monoclonal rabbit anti-MYC antibody[20,34] (1:200 dilution; ab32072; Abcam, Cambridge, MA), and polyclonal rabbit anti-FGFR2 antibody[35] (1:200 dilution; ab52246; Abcam). After the sections were washed, the slides were incubated for 30 minutes with the peroxidase polymer-linked secondary antibody that was contained in the kit. The slides were subjected to colorimetric detection using the ImmPact DAB substrate (SK-4105; Vector Laboratories). All of the stained slides were scanned using Aperio ScanScope AT (Aperio Technologies, Inc., Vista, CA). IHC grading was performed under high-power microscopic magnification (×400). The positive rates were depicted as the mean value of at least 6 high power fields (×200). Care was taken to select the fields equally from invasive and mucosal layers so that the grading may not be biased by the proportion of invasive fronts included across tumors. Specimens with <6 high-power fields or 2000 cancer cells, mostly small biopsy samples, were excluded from grading for immunohistochemistry. A total of 197 and 194 tumors in the EODGC and LODGC cohorts were eligible for IHC grading for E-cadherin and β-catenin, respectively.

A team of 5 pathologists (K.S.S., J.L., H.K.C., S.Y.K., and J.W.P.) performed blind reviews of E-cadherin and β-catenin immunohistochemistry. E-cadherin immunohistochemistry was independently reviewed and graded by these 5 pathologists. When the grades differed, the pathologists conferred and came to a consensus. Without prior knowledge of genomic data, a pathologist (J.W.P.) visually scored the percentage of nuclear β-catenin immunostaining, the score was then verified by the other 4 pathologists (K.S.S., J.L., H.K.C., and S.Y.K.). The pathologists conferred for consensus in case they disagreed.

For E-cadherin grading, the presence of strong complete circumferential membrane stain in non-cancerous epithelial cells was defined as positive internal and external controls. Staining was considered aberrant if the cancer cells showed (1) an absence of circumferential membrane staining; (2) reduced-intensity, partial or weak linear membrane staining; or (3) a focal or punctate (dot-like) cytoplasmic or membrane reaction (Figure 4*A*).[36] E-cadherin was graded into 3 groups according to the percentages of tumor cells with aberrant staining: grade 2, >80% of the cancer cells showed homogenous membranous E-cadherin staining: grade 1, 20%−80% of the cancer cells were positively stained: and grade 0, <20% of the cancer cells showed positive staining. E-cadherin immunostaining was considered abnormal if aberrant or reduced staining was identified in >20% of tumor cells (ie, grade 0 or 1).[10,37] The nuclear expression of β-catenin was scored according to the percentage of cancer cells that exhibited unequivocal and strong nuclear staining.

RHOA-mutated tumors (n = 30) exhibited more prominent nuclear β-catenin immunostaining than did wild-type tumors for *RHOA* (n = 164) (*P* = .011, Wilcoxon). Independent of the *CDH1* alteration status, the *RHOA* mutation status was significantly associated with nuclear β-catenin immunostaining, according to a multiple regression analysis performed using the SAS GLM procedure (adjusted *P* = .01, linear regression; Supplementary Table 11).

To evaluate the correlation of MYC expression with nuclear β-catenin expression, 64 GC cases were randomly selected for MYC immunohistochemistry. MYC expression was scored according to the percentage of cancer cells exhibiting unequivocal and strong nuclear staining (Supplementary Figure 6*E*).

To validate FGFR2 overexpression in *FGFR2*-amplified samples, a total of 19 cases including 9 *FGFR2*-amplified and randomly selected 10 tumors without FGRF2 amplification were subjected to FGFR2 immunohistochemistry. Cytoplasmic FGFR2 expression was scored according to the intensity of staining: 0, negative staining; 1, weakly positive staining; 2, moderately positive staining; 3, strongly positive staining.[38] The cytoplasmic immunoreactivity in adjacent normal gastric epithelial cells was defined as internal controls showing negative to weak positivity. *FGFR2*-amplified tumors (n = 9) showed significantly higher FGFR2 immunohistochemistry grades than did tumors without *FGFR2* amplification (n = 10) (*P* = .036, *t* test).

Cell Lines

NUGC-4 (JCRB0834; JCRB) was purchased for this study. The identity of MKN-45 (80103; KCLB, Seoul, Korea) was verified by STR profiling was using AmplFLSTR identifiler PCR Amplification kit (Applied Biosystems, Foster, CA). NCC-S1 cell line was established by our group from a diffuse-type gastric adenocarcinoma formed in a *Villin-cre;Smad4*$^{F/F}$*;Trp53*$^{F/F}$*;Cdh1*$^{F/wt}$ mouse.[32,39] *Pdx1-cre;Smad4*$^{F/F}$*;Trp53*$^{F/F}$*;Cdh1*$^{F/+}$ cells were primary cultured by our group from a diffuse-type gastric adenocarcinoma formed in a *Pdx1-cre;Smad4*$^{F/F}$*;Trp53*$^{F/F}$*;Cdh1*$^{F/wt}$ mouse.[32,39] The Chinese hamster ovary cell line CHO-K1 (10061; KCLB) was purchased from the Korean Cell Line Bank (Seoul, Korea). All cells were grown in RPMI-1640 with 10% fetal bovine serum (FBS) and 0.5% penicillin-streptomycin solution (WELGENE, Gyeongsan, Korea) at 37°C under 5% CO_2 in a humidified incubator.

BAP1 Expression Lentiviral Vectors

For cDNA cloning of *BAP1*, full length CDS of *BAP1* was PCR-amplified with the primers 5′-ATG AAT TCG CCA CCA TGA ATA AGG GCT GGC TGG AG-3′/5′-TGT GAT TGT CTA GAA AGG CCG-3′ and 5′-TTT CTA GAC AAT CAC AAT TAT GCC AAG-3′/5′-TAG CGG CCG CTC ACT GGC GCT TGG CCT TG-3′. The amplicons were then digested with EcoRI/XbaI and XbaI/NotI, and were cloned into a pCDH-CMV-MCS-EF1-Puro-copGFP vector.

CDH1 Expression Lentiviral Vectors

Using lentivirus vectors, we ectopically expressed *CDH1* using CMV promoters in human GC cells, and expressed *CDH1* under CAG promoters CHO-K1 and mouse DGC cells. We first generated vectors by modifying a selection marker (copGFP) in pCDH-CMV-MCS-EF1-copGFP (System Biosciences, Mountain View, CA). To generate an appropriate enzyme site (KpnI) near the downstream promoter of the selection marker to join the vector and the new selection marker together, the EF1 promoter (pEF1) in the original vector was subjected to PCR reaction with the primers 5′-ATG CGG CCG CAA GGA TCT GCG ATC GCT C-3′/5′-ATG GTA CCG GTA GGC GCC GGT CAC AG-3′. For the dual selection marker puromycin resistance gene (PURO) and copGFP, the full-length PURO (without a stop codon) and copGFP genes were amplified from template vectors (pLKO.1-Puro (Sigma, St Louis, MO) and pCDH-CMV-MCS-EF1-copGFP, respectively) using PCR with the primer pairs: 5′-ATG GTA CCG CCA CCA TGA CCG AGT ACA AGC CC-3′/5′-ATA GAT CTG GCA CCG GGC TTG CGG GT-3′ and 5′-ATA GAT CTG CCA CCA TGG AGA GCG ACG AGA GC-3′/5′-TAG TCG ACT TAG CGA GAT CCG GTG GAG-3′. Restriction enzyme digestion of vectors and amplified DNA fragments was performed as follows: NotI/SalI for the vector backbone (pCDH-CMV-MCS-EF1-copGFP), NotI/KpnI for pEF1, KpnI/BglII for PURO, and BglII/SalI for copGFP. All fragments were then ligated altogether to generate a vector (pCDH-CMV-MCS-EF1-Puro-copGFP) containing a PURO-copGFP fusion selection marker. To obtain a selection marker containing a neomycin resistance gene (NEO), the full-length NEO gene was amplified from a pcDNA3.1 (Thermo Fisher Scientific) template vector using PCR with the primers 5′-ATG GTA CCG CCA CCA TGA TTG AAC AAG ATG GAT TG-3′/5′-TAG TCG ACT CAG AAG AAC TCG TCA AGA AG-3′. Digestion was then performed using NotI/SalI for the vector backbone, NotI/KpnI for pEF1, and KpnI/SalI for NEO. All of the fragments were then ligated together to construct pCDH-CMV-MCS-EF1-Neo. To obtain a selection marker containing PURO, pEF1 was first reconstituted using PCR with the primers 5′-ATG CGG CCG CAA GGA TCT GCG ATC GCT C-3′/5′-ATC TCG AGG TAG GCG CCG GTC ACA GC-3′ to tag an appropriate enzyme site (XhoI) that could be used to join pEF1 and PURO together. Then, the full-length PURO gene was amplified from pLKO.1-Puro using PCR with the primers 5′-ATC TCG AGA TGA CCG AGT ACA AGC CC-3′/5′-TAG TCG ACT CAG GCA CCG GGC TTG CG-3′. Digestion was performed using NotI/SalI for the vector backbone, NotI/XhoI for pEF1, and XhoI/SalI for PURO. All fragments were then ligated together to obtain pCDH-CMV-MCS-EF1-Puro. For easy PCR amplification of pCAG, pCAG was divided and amplified into 2 fragments from the Ai9 (Addgene, Cambridge, MA) template vector using the primers 5′-ATT ACT AGT TAT TAA TAG TAA TCA ATT ACG GG-3′/5′-ACA AAG GGC CCT CCC GGA G-3′ and 5′-GGG AGG GCC CTT TGT GCG-3′/5′-ATA GAA TTC GCT AGC TTT GCC AAA ATG ATG AGA CAG-3′. After digestion with SpeI/EcoRI, SpeI/ApaI for upstream pCAG, and ApaI/EcoRI for downstream pCAG all fragments were ligated altogether to pCDH-CAG-MCS-EF1-Puro-copGFP/Neo/Puro. To clone the CDH1 cDNA, the full-length coding sequence (CDS) of *CDH1* was divided and PCR-amplified into 2 fragments by tagging a KpnI junction with the primers 5′-ATG CTA GCG CCA CCA TGG GCC CTT GGA GCC GC-3′/5′-AAG GTA CCA CAT TCG TCA CTG C-3′ and 5′-GTG GTA CCT TTT GAG GTC TCT C-3′/5′-ATG CGG CCG CCT AGT CGT CCT CGC CGC C-3′. Wild-type and mutant *CDH1* fragments were amplified from the cDNA of the relevant clinical samples. Specifically, for the E880K variant, the downstream fragment of *CDH1* was amplified using the primers 5′-GTG GTA CCT TTT GAG GTC TCT C-3′/5′-ATG CGG CCG CCT AGT CGT CCT TGC CGC CTC-3′. Restriction enzyme digestion of the newly made vectors and amplified DNA fragments was performed using NheI/NotI for the vectors, NheI/KpnI for upstream *CDH1*, and KpnI/NotI for downstream *CDH1*. The cut vector and the upstream/downstream fragments of each *CDH1* gene were ligated to wild-type and mutant *CDH1*-expressing vectors.

RHOA Expression Vectors

To clone the *RHOA* cDNA, full-length CDS of *RHOA* was PCR-amplified from a RHOA-wild-type cDNA sample. Primer pairs (5′-ATG AAT TCG CCA CCA TGG CTG CCA TCC GGA AG-3′/5′-ATG CGG CCG CTC ACA AGA CAA GGC ACC CAG-3′ and 5′-ATG AAT TCG CCA CCA TGG CTG CCA TCT GGA AG-3′/5′-ATG CGG CCG CTC ACA AGA CAA GGC ACC CAG-3′) were used to amplify the wild-type and mutant (R5W) *RHOA* genes, respectively. The wild-type and mutant *RHOA* genes were cloned into pcDNA3.1 (+) (Thermo Fisher Scientific) using EcoRI and NotI.

Generation of Stable Cell Lines

All cloned vector sequences were validated using Sanger sequencing. Three days after 293FT cells (Thermo Fisher Scientific) were transfected with one of the cloned gene expression vectors, pMD2.G (Addgene), and psPAX2 (Addgene), the growth medium was harvested and concentrated using ultracentrifugation. Cells were transduced at a multiplicity of infection of 0.1. For antibiotic selection, NUGC-4 cells and MKN-45 cells were treated with 2 μg/mL of puromycin for 3–4 passages. For CHO-K1 cells, 20 μg/mL of puromycin or 500 μg/mL of G418 was used for 3–4 passages. For validation of transgene expression, after cDNA was synthesized from 5 μg of total RNA using amfiRivert cDNA Synthesis Platinum Master Mix (GenDEPOT, Barker, TX), PCR was performed using 0.4 μL of the synthesized cDNA per 20 μL of the ExTaq mixture (Takara, Shiga, Japan). PCR amplifications were performed using 33 cycles for the transgene and 28 cycles for the internal control (*Actb* or *GAPDH*). The primer sequences used for these experiments were as follows: *CDH1* for exons 4 and 5, 5′-CCA TCA GGC CTC CGT TTC TGG-3′ and 5′-TGT CAT TCT GAT CGG TTA CCG TGA TC-3′; *GAPDH*, 5′-GAG TCA ACG GAT

TTG GTC G-3′ and 5′-TGG AAT CAT ATT GGA ACA TGT AAA C-3′; and *Actb*, 5′-GAA CAT GGC ATT GTT ACC AAC TG-3′ and 5′-GTG TTG AAG GTC TCA AAC ATG ATC-3′.

Small Interfering RNA-Mediated RHOA Knockdown and Exoenzyme C3 Transferase Treatment

Small interfering RNA (siRNA) SMARTpool against *RHOA* (J-003860-10 and J-003860-13; Thermo Fisher Scientific) and *Rhoa* (J-042634-05; Thermo Fisher Scientific) were used to knock down RHOA in human and mouse DGC cells, respectively. Cells suspended in OPTI-MEM medium containing 5% FBS were co-transfected with the reporter plasmids along with siRNAs (20 nM). After 48 hours of the co-transfection, luciferase activity was measured as described.

To evaluate the effect of exoenzyme C3 transferase (*C3*; CT04; Cytoskeleton, Denver, CO) on the reporter activities, GC cells transfected with the reporter plasmids for 24 hours were incubated with 1 μg/mL of C3 in serum-free RPMI-1640 for 24 hours.

BAP1 Proliferation Assays

To evaluate the effect of *BAP1* on cell proliferation, we used MTT assays. The 2 × 10^5 cells stably expressing empty vector or *BAP1* were seeded to each well of 12-well plate. After the cells were incubated in a mixture of 0.1 mL of MTT solution (Sigma; 5 mg/mL in phosphate-buffered saline [PBS]) and 0.5 mL of growth medium, the mixture was completely removed, and then insoluble reactants of the cells were dissolved with 0.5 mL of dimethyl sulfoxide. Absorbance of 0.1 mL of the solution in a 96-well plate was measured every day for 3–4 days at 570 nm using a microplate reader (Molecular Devices, Sunnyvale, CA). Normalized optical density at 570 nm was calculated by subtracting the optical density of dimethyl sulfoxide blank.

Slow Aggregation Assay

After the cells were detached using trypsin, 2 × 10^4 cells in 0.2 mL of growth medium were seeded into 96-well plates that were pre-coated with 45 μL of 0.8% (w/v) DIFCO Noble Agar (BD Biosciences) in PBS.[40,41] After 24 hours of incubation, images showing cellular aggregation were obtained using a light microscope (50×) (Axio Observer Z1, Carl Zeiss, Jena, Germany).

Measurement of the Serum Responsive Factor and Wnt/β-Catenin Activity

To assess the activity of serum responsive factor and Wnt/β-catenin activity, we conducted luciferase reporter assays using the pGL4.34[*luc2P*/serum responsive factor-RE/Hygro] reporter plasmid (Promega, Madison, WI) and Super 8× TOPFlash and FOPFlash reporter plasmids (Addgene). To evaluate whether an R5W *RHOA* mutation is gain- or loss-of-function, we suspended 2.5 × 10^5 293FT cells in 0.5 mL of OPTI-MEM medium (Thermo Fisher Scientific) containing 5% FBS and cell suspension cells was seeded in 24-well plates. Cells were then transfected with 200 ng of plasmids using Lipofectamine 2000 (Thermo Fisher Scientific), as recommended by the manufacturer. After 24 hours of transfection, cells were transfected with 200 ng of reporter plasmids and 2.5 ng of Renilla luciferase-constitutively expressing vector (Addgene; for internal control) per well using Lipofectamine 2000 (Thermo Fisher Scientific). After 48 hours of the first transfection, luciferase activity was measured using a Dual-Luciferase Reporter Assay System (Promega) and a microplate luminometer (Wallac 1420 Victor 3; PerkinElmer, Waltham, MA).

Immunofluorescence Analysis for Actin Stress Fiber and E-Cadherin

293FT cells suspended in OPTI-MEM media containing 5% FBS were transfected with empty, wild-type, or mutant (R5W) *RHOA* vectors. GC cells were suspended and transfected with siRHOA (20 nM). Suspension cells (1 × 10^5 cells) were then seeded on sterile glass coverslips (18 mm in diameter) in 12-well plates. For C3 treatment, 1 × 10^5 GC cells were seed on glass coverslips and grown to 70%–80% confluency, and cells were treated with 1 μg/mL of C3 in serum-free RPMI-1640. After 48 hours after the transfection (24 hours after C3 treatment), cells on glass coverslips were fixed in 4% paraformaldehyde for 10 minutes and permeabilized in 0.5% Triton X-100 (Sigma) for 5 minutes at room temperature. For immunofluorescence (IF) analysis for actin stress fiber, rhodamine-phalloidin (PHDR1, Cytoskeleton) was used. Stained slides were evaluated using LSM510 confocal microscope (Carl Zeiss). For the IF analysis of E-cadherin in wild-type or mutant *CDH1*-transduced NCC-S3 cells, cells were seeded on sterile glass coverslips and grown to 70%–80% confluence. Cells on the coverslips were fixed and permeabilized as described, and incubated with 5% goat serum for blocking. The primary 24E10/monoclonal rabbit against E-cadherin antibody (1:1000; 3195; Cell Signaling) was incubated overnight at 4°C. After washing, Alexa Fluor 594 goat anti-rabbit IgG (H+L) secondary antibody, (1:250; R37117; Thermo Fisher Scientific) was incubated for 2 hours at room temperature. Slides were mounted with Vectashield mounting media (H-1200; Vector Laboratories) and evaluated using Zeiss Axio Imager HBO 100 (Carl Zeiss).

RhoA Pull-Down Activation Assay

RhoA pull-down activation assays (RhoTekin assay) were performed using a RHOA Activation Assay Biochem Kit (BK036; Cytoskeleton) as recommended by the manufacturer, with some modifications. Cells were harvested and lysed using T-PER reagent with protease inhibitors. After protein concentration was measured using a BCA reagent, pull-down was performed using 0.3 mg of total protein and 60 μg of rhotekin-RBD beads in 0.5 mL of T-PER Reagent at 4°C for 1 hour. The bead pellets were washed twice with T-PER Reagent that was diluted 1:10 in PBS. After the last wash, buffer was removed leaving 90 μL of buffer and the bead mixture, 30 μL of 4× Laemmli sample buffer was added to the remaining mixture, and the samples were boiled for 5 minutes. Twenty microliters of the sample volume was loaded for Western blot analysis. Band intensity was quantified using ImageJ software and normalized to total RHOA protein.

Migration Assay

Cells were plated on 24-well inserts with 8-μm pore (353097; BD) at 5 × 10^4/well in serum-free RPMI-1640 media. Medium containing 10% FBS was added to 24-well insert plate (354578, BD) and cells were cultured at 37°C under 5% CO_2. At 24 hours after plating, remaining cells were removed by gently scrapping the upper chamber with a wet cotton swab. Migrated cells were fixed with 10% formalin for 10 minutes and washed with PBS once. The inserts were soaked in hematoxylin for 1 minute and washed. Membranes were cut from inserts and moved to a glass slide. Mean number of migrated cells was determined by counting 3 high-power fields (200×).

Western Blot Analysis

Cells were lysed using T-PER Tissue Protein Extraction Reagent (Thermo Fisher Scientific) supplemented with protease and phosphatase inhibitors. After cellular debris was removed by centrifugation, and protein quantitation was performed using a BCA Protein Assay Kit (Thermo Fisher Scientific), the protein samples were prepared in Laemmli sample buffer by boiling. The same amount of each protein sample was separated on sodium dodecyl sulfate polyacrylamide gels and transferred onto nitrocellulose membranes using electrophoresis and blotting apparatuses (Bio-Rad Laboratories, Hercules, CA). After blocking using 2.5% bovine serum albumin in PBS with Tween20 (PBS-T; Sigma), we probed membranes using primary antibodies (1:1000) in 3% BSA/PBS-T overnight, followed by a horseradish peroxidase–conjugated secondary antibody (anti-mouse IgG or anti-rabbit IgG; GenDEPOT) (1:5000) for 2 hours. SuperSignal West Pico Chemiluminescent Substrate kit (Thermo Fisher Scientific) was used for detection. C4/monoclonal mouse anti-BAP1 (sc-28383; Santa Cruz Biotechnology, Dallas, TX), 4/monoclonal mouse anti–β-catenin antibody (610154; BD Biosciences),[33,39] 24E10/monoclonal rabbit against E-cadherin antibody (3195; Cell Signaling, Danvers, MA),[39] monoclonal mouse anti-RHOA (26007; NewEast Biosciences, Malvern, PA), and C4/monoclonal mouse anti–β-actin (sc-47778; Santa Cruz Biotechnology) were used.

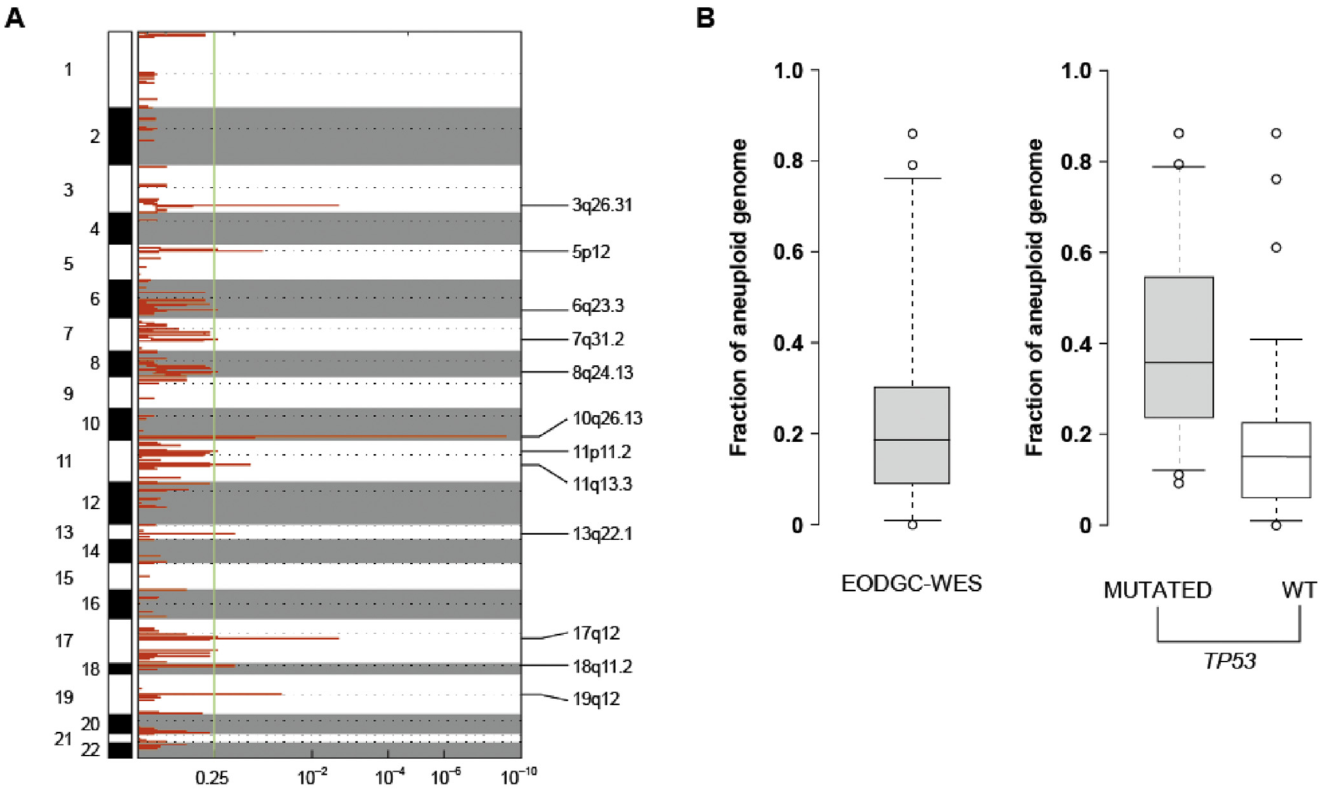
Supplementary References

1. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202–209.
2. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213–219.
3. Chapman MA, Lawrence MS, Keats JJ, et al. Initial genome sequencing and analysis of multiple myeloma. Nature 2011;471:467–472.
4. **Ramos AH, Lichtenstein L**, Gupta M, et al. Oncotator: cancer variant annotation tool. Hum Mutat 2015;36:E2423–E2429.
5. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 2012;30:413–421.
6. **Lawrence MS, Stojanov P, Polak P**, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–218.
7. **Kim J, Mouw KW, Polak P**, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat Genet 2016;48:600–606.
8. **Zack TI, Schumacher SE**, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013;45:1134–1140.
9. Mermel CH, Schumacher SE, Hill B, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;12:R41.
10. **Gerstung M, Pellagatti A**, Malcovati L, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nat Commun 2015;6:5901.
11. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–1303.
12. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42:D980–D985.
13. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43:D805–D811.
14. Mootha VK, Lindgren CM, Eriksson K-F, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003;34:267–273.
15. Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–899.
16. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–595.
17. Morin RD, Bainbridge M, Fejes A, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 2008;45:81–94.
18. Olshen AB, Venkatraman ES, Lucito R, et al. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5:557–572.
19. Nitta H, Hauss-Wegrzyniak B, Lehrkamp M, et al. Development of automated bright field double in situ hybridization (BDISH) application for HER2 gene and chromosome 17 centromere (CEN 17) for breast carcinomas and an assay performance comparison to manual

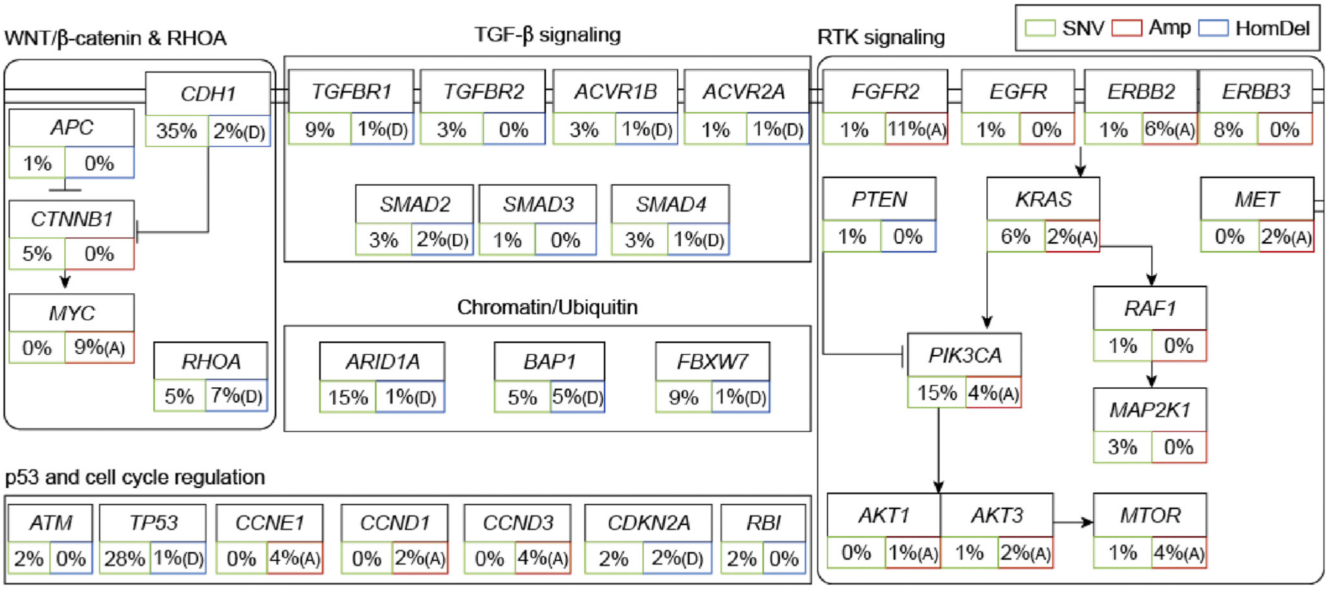
dual color HER2 fluorescence in situ hybridization (FISH). Diagn Pathol 2008;3:41.

20. Park YS, Hwang HS, Park HJ, et al. Comprehensive analysis of HER2 expression and gene amplification in gastric cancers using immunohistochemistry and in situ hybridization: which scoring system should we use? Hum Pathol 2012;43:413–422.
21. Wolff AC, Hammond MEH, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 2007;25:118–145.
22. Boland CR, Thibodeau SN, Hamilton SR, et al. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 1998;58:5248–5257.
23. Drobinskaya I, Gabbert HE, Moeslein G, et al. A new method for optimizing multiplex DNA microsatellite analysis in low quality archival specimens. Anticancer Res 2005;25:3251–3258.
24. **Lim H, Park YS**, Lee JH, et al. Features of gastric carcinoma with lymphoid stroma associated with Epstein-Barr virus. Clin Gastroenterol Hepatol 2015;13:1738–1744.
25. Labrecque LG, Barnes DM, Fentiman IS, et al. Epstein-Barr virus in epithelial cell tumors: a breast cancer study. Cancer Res 1995;55:39–45.
26. Martínez-López JLE, Torres J, Camorlinga-Ponce M, et al. Evidence of Epstein-Barr virus association with gastric cancer and non-atrophic gastritis. Viruses 2014; 6:301–318.
27. Desmet F-O, Hamroun D, Lalande M, et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 2009;37:e67.
28. Brasch J, Harrison OJ, Honig B, et al. Thinking outside the cell: how cadherins drive adhesion. Trends Cell Biol 2012;22:299–310.
29. Corso G, Carvalho J, Marrelli D, et al. Somatic mutations and deletions of the E-cadherin gene predict poor survival of patients with gastric cancer. J Clin Oncol 2013; 31:868–875.
30. Oliveira C, Sousa S, Pinheiro H, et al. Quantification of epigenetic and genetic 2nd hits in CDH1 during hereditary diffuse gastric cancer syndrome progression. Gastroenterology 2009;136:2137–2148.
31. Alami J, Williams BR, Yeger H. Differential expression of E-cadherin and beta catenin in primary and metastatic Wilms's tumours. Mol Pathol 2003;56:218–225.
32. Park JW, Jang SH, Park DM, et al. Cooperativity of E-cadherin and Smad4 loss to promote diffuse-type gastric adenocarcinoma and metastasis. Mol Cancer Res 2014; 12:1088–1099.
33. Rodrigues P, Macaya I, Bazzocco S, et al. RHOA inactivation enhances Wnt signalling and promotes colorectal cancer. Nat Commun 2014;5:5458.
34. Toon CW, Chou A, Clarkson A, et al. Immunohistochemistry for Myc predicts survival in colorectal cancer. PLoS One 2014;9:e87456.
35. Ramsey MR, Wilson C, Ory B, et al. FGFR2 signaling underlies p63 oncogenic function in squamous cell carcinoma. J Clin Invest 2013;123:3525–3538.
36. Choi YJ, Pinto MM, Hao L, et al. Interobserver variability and aberrant E-cadherin immunostaining of lobular neoplasia and infiltrating lobular carcinoma. Mod Pathol 2008;21:1224–1237.
37. Blok P, Craanen ME, Dekker W, et al. Loss of E-cadherin expression in early gastric cancer. Histopathology 1999; 34:410–415.
38. Nagatsuma AK, Aizawa M, Kuwata T, et al. Expression profiles of HER2, EGFR, MET and FGFR2 in a large cohort of patients with gastric adenocarcinoma. Gastric Cancer 2015;18:227–238.
39. Park JW, Park DM, Choi BK, et al. Establishment and characterization of metastatic gastric cancer cell lines from murine gastric adenocarcinoma lacking Smad4, p53, and E-cadherin. Mol Carcinog 2015;54:1521–1527.
40. Suriano G, Oliveira C, Ferreira P, et al. Identification of CDH1 germline missense mutations associated with functional inactivation of the E-cadherin protein in young gastric cancer probands. Hum Mol Genet 2003;12:575–582.
41. Simões-Correia J, Figueiredo J, Lopes R, et al. E-Cadherin destabilization accounts for the pathogenicity of missense mutations in hereditary diffuse gastric cancer. PLoS One 2012;7:e33783.

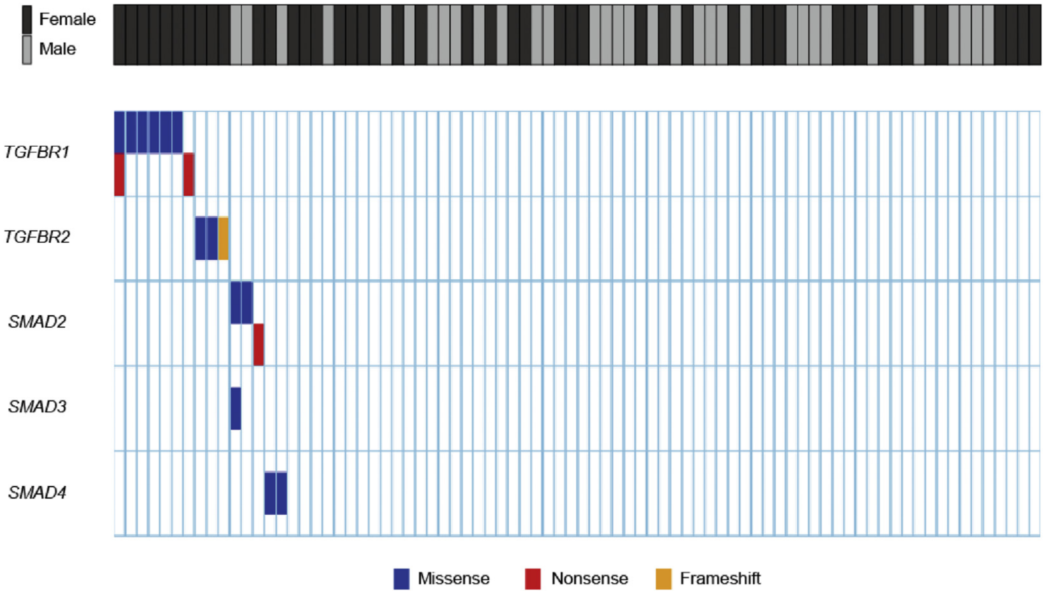
Author names in bold designate shared co-first authorship.



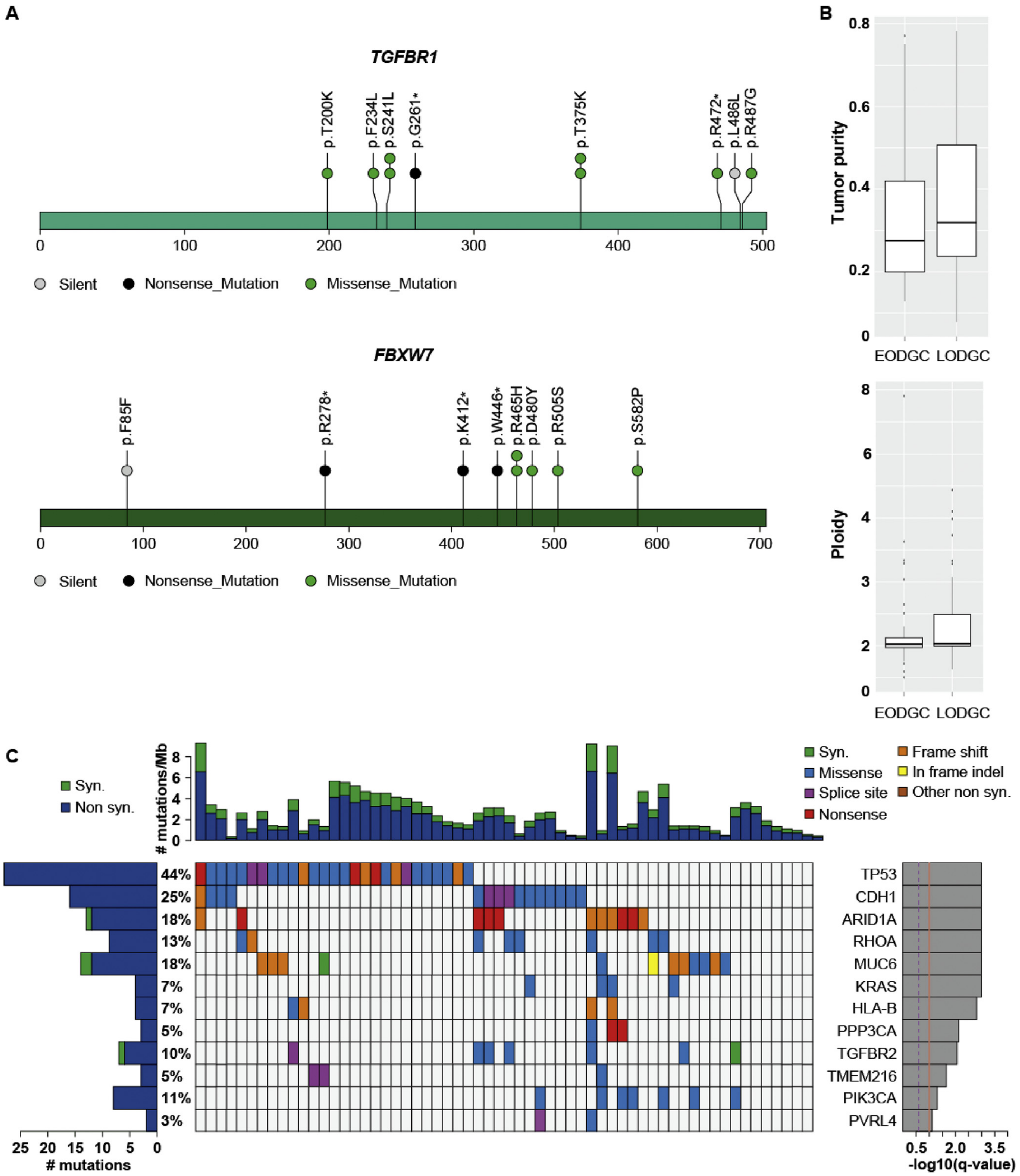
Supplementary Figure 1. Aneuploidy and copy number analyses of the WES data of EODGC-WES. (*A*) Focal amplifications in EODGC-WES, as identified by ReCapSeg and ISAR-GISTIC. (*B*) Fraction of aneuploidy genome in EODGC-WES. *Left*, median fraction of aneuploid genome in EODGC-WES (0.19); *right*, fraction of aneuploidy genome in EODGC-WES according to *TP53* mutation status. The *TP53* mutation was a mutation that most significantly correlated with the aneuploidy. Mean fractions of aneuploid genome were 0.40 and 0.18 in *TP53*-mutated and wild-type (WT) EODGC-WES tumors, respectively ($P = 2.8 \times 10^{-6}$, *t* test).



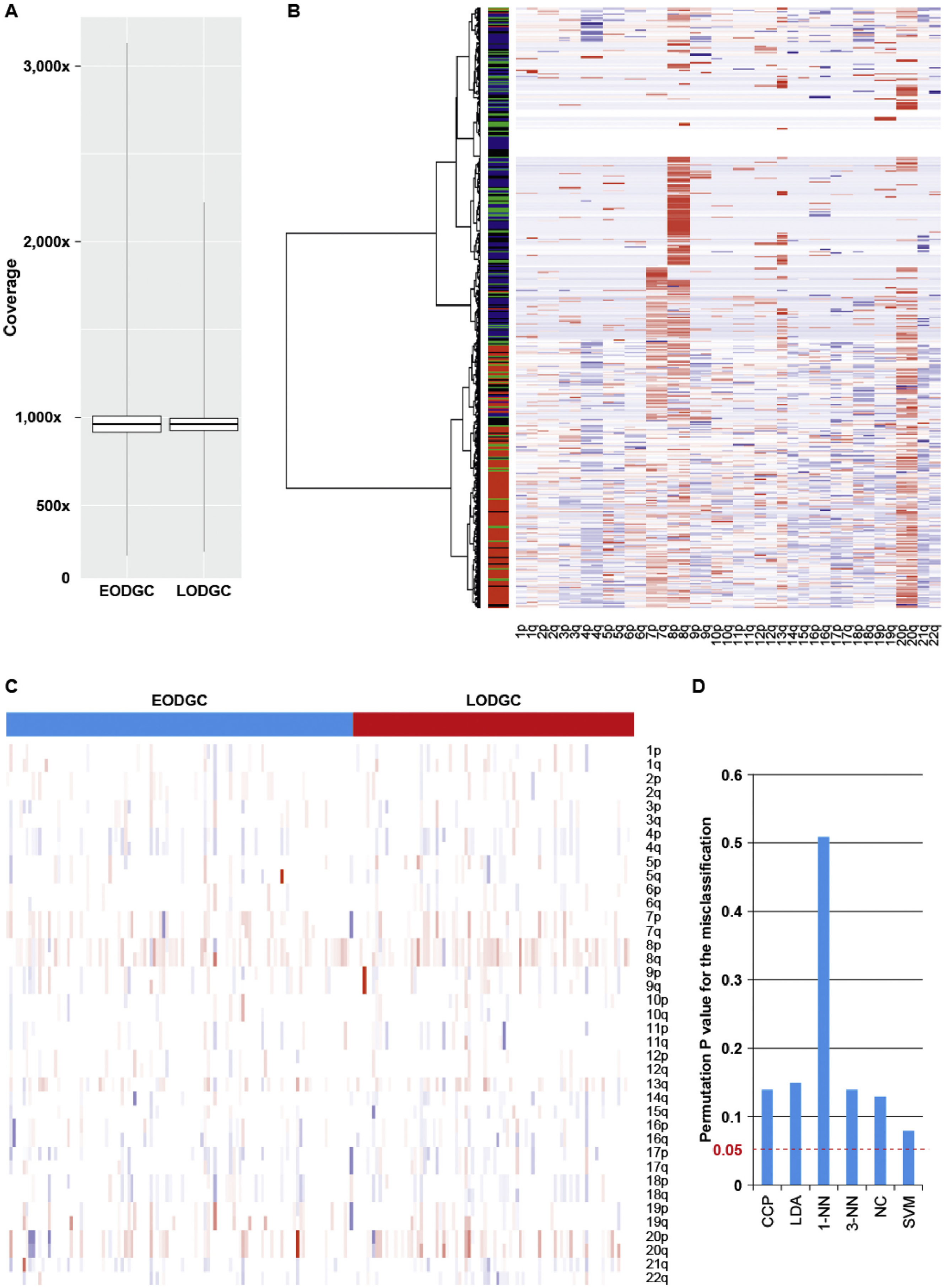
Supplementary Figure 2. Somatic genomic alterations in key pathways of EODGC-WES as identified by ReCapSeg/ISAR-GISTIC and MuTect/Indelocator analyses. A *green box* indicates the number of EODGC-WES tumors with at least 1 non-silent mutation. *Red* and *blue boxes* indicate numbers of EODGC-WES tumors with focal amplifications and homozygous deletions, respectively. Amp, amplification (focal); HomDel, homozygous deletion; TGF, transforming growth factor.



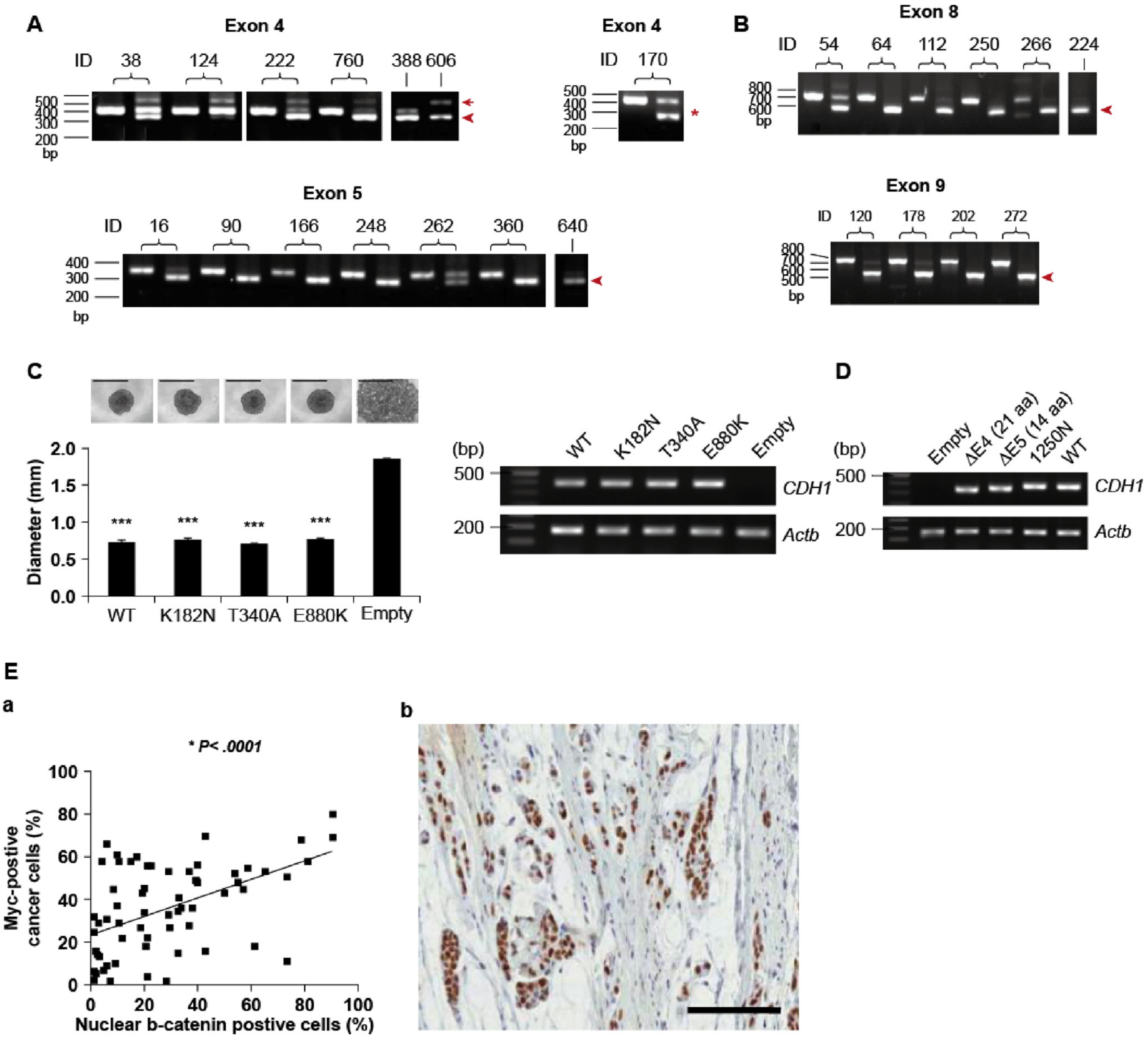
Supplementary Figure 3. Female-specific *TGFBR1* mutations in EODGC-WES. *Black bar*, female; *gray bar*, male. *TGFBR1* mutations were found in 7 tumors from EODGC-WES women, but in no tumors from EODGC-WES men ($P = .014$, χ^2).


A
TGFBR1
p.T200K
p.F234L
p.S241L
p.G261*
p.T375K
p.R472*
p.L486L
p.R487G
0
100
200
300
400
500
Silent
Nonsense_Mutation
Missense_Mutation
FBXW7
p.F85F
p.R278*
p.K412*
p.W446*
p.R465H
p.D480Y
p.R505S
p.S582P
0
100
200
300
400
500
600
700
Silent
Nonsense_Mutation
Missense_Mutation
B
Tumor purity
0.8
0.6
0.4
0.2
0
EODGC
LODGC
Ploidy
8
6
4
3
2
0
EODGC
LODGC
C
Syn.
Non syn.
# mutations/Mb
8
6
4
2
0
Syn.
Missense
Splice site
Nonsense
Frame shift
In frame indel
Other non syn.
44%
25%
18%
13%
18%
7%
7%
5%
10%
5%
11%
3%
TP53
CDH1
ARID1A
RHOA
MUC6
KRAS
HLA-B
PPP3CA
TGFBR2
TMEM216
PIK3CA
PVRL4
25
20
15
10
5
0
# mutations
0.5
2.0
3.5
-log10(q-value)

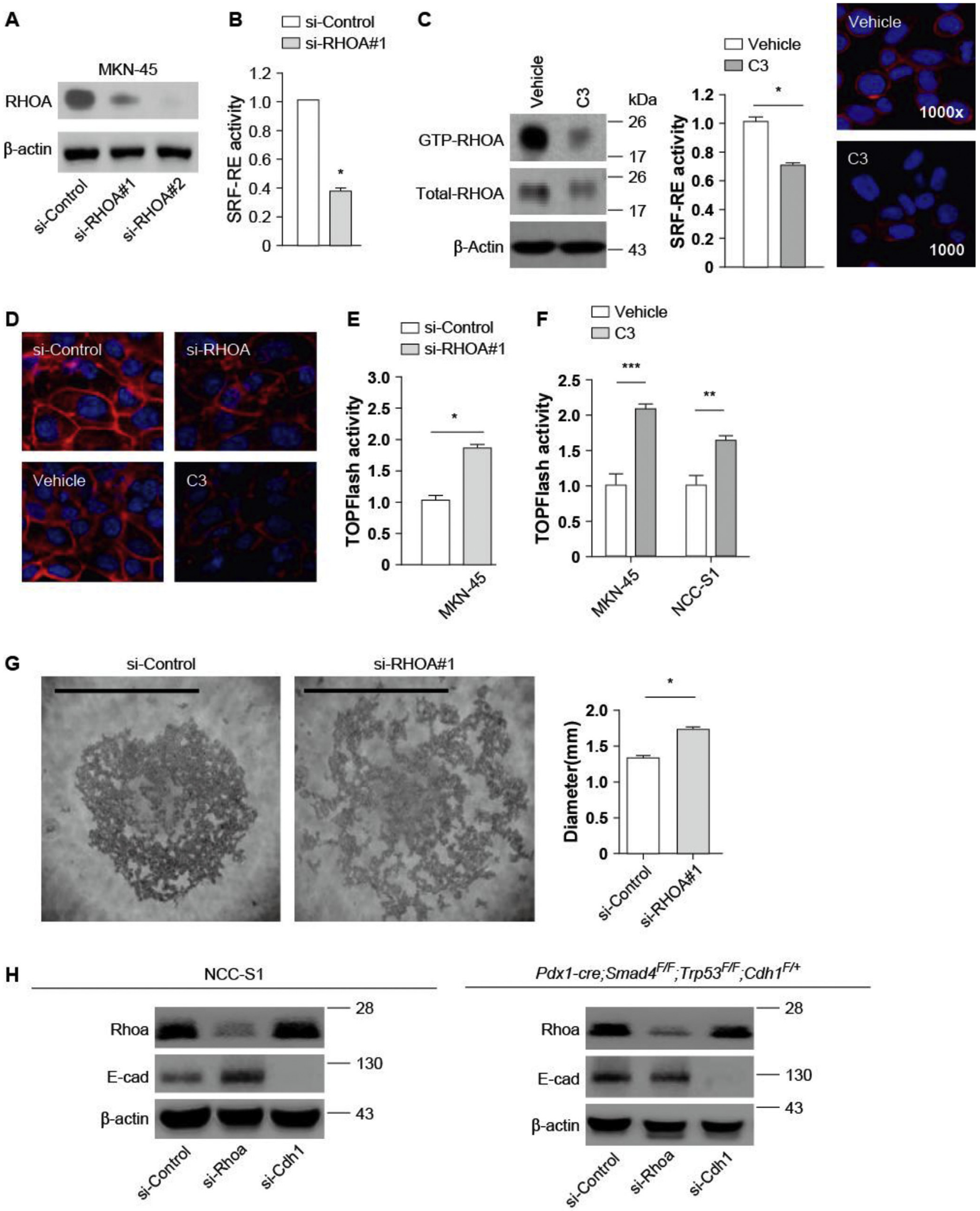
Supplementary Figure 4. WES data of EODGC-WES and LODGC-TCGA. (*A*) *Stick figures* for *TGFBR1* and *FBXW7* mutation sites in EODGC-WES (*B*) ABSOLUTE analysis of tumor purity and ploidy in EODGC-WES. *Upper panel*, tumor purity in EODGC-WES (EODGC) and LODGC-TCGA (LODGC). There was no significant difference in tumor purity between EODGC-WES (n = 80) and LODGC-TCGA (n = 61) (mean purity, 0.32 vs 0.38, respectively. *P* = NS). *Lower panel*, there was no significant difference in the ploidy between EODGC-WES and LODGC-TCGA (mean ploidy, 2.1 vs 2.3, respectively; *P* = NS). Callable sequencing coverage means were 93.6-fold (range, 90.6–95.2) and 87.4-fold (range, 76.2–91.2, for EODGC-WES (n = 80) and LODGC-TCGA (n = 61), respectively. In EODGC-WES, mean read depths were 138× and 135× in tumor and germline DNA, respectively. In LODGC-TCGA, mean read depths were 85× and 87× in tumor and germline DNA, respectively. (*C*) Recurrent mutations in LODGC-TCGA (n = 61). When we compared mutation profiles between EODGC-WES and LODGC-TCGA, *CDH1*, *TP53, ARID1A, PIK3CA, KRAS*, and *RHOA* mutations were recurrent in both cohorts. EODGC-WES had suggestively higher rates of *TGFBR1* (8.8% vs 3.3% in LODGC-TCGA; *P* = .188) and *CDH1* mutations (35.0% vs 24.6% in LODGC-TCGA; *P* = .184) than LODGC-TCGA. In contrast, *RHOA* mutation rate tended to be higher in LODGC-TCGA than in EODGC-WES (13.1% vs 5.0% in EODGC-WES; *P* = .087).


A
3,000x
2,000x
1,000x
500x
0
Coverage
EODGC
LODGC
B
1p
1q
2p
2q
3p
3q
4p
4q
5p
5q
6p
6q
7p
7q
8p
8q
9p
9q
10p
10q
11p
11q
12p
12q
13q
14q
15q
16p
16q
17p
17q
18p
18q
19p
19q
20p
20q
21q
22q
C
EODGC
LODGC
1p
1q
2p
2q
3p
3q
4p
4q
5p
5q
6p
6q
7p
7q
8p
8q
9p
9q
10p
10q
11p
11q
12p
12q
13q
14q
15q
16p
16q
17p
17q
18p
18q
19p
19q
20p
20q
21q
22q
D
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0
Permutation P value for the misclassification
CCP
LDA
1-NN
3-NN
NC
SVM

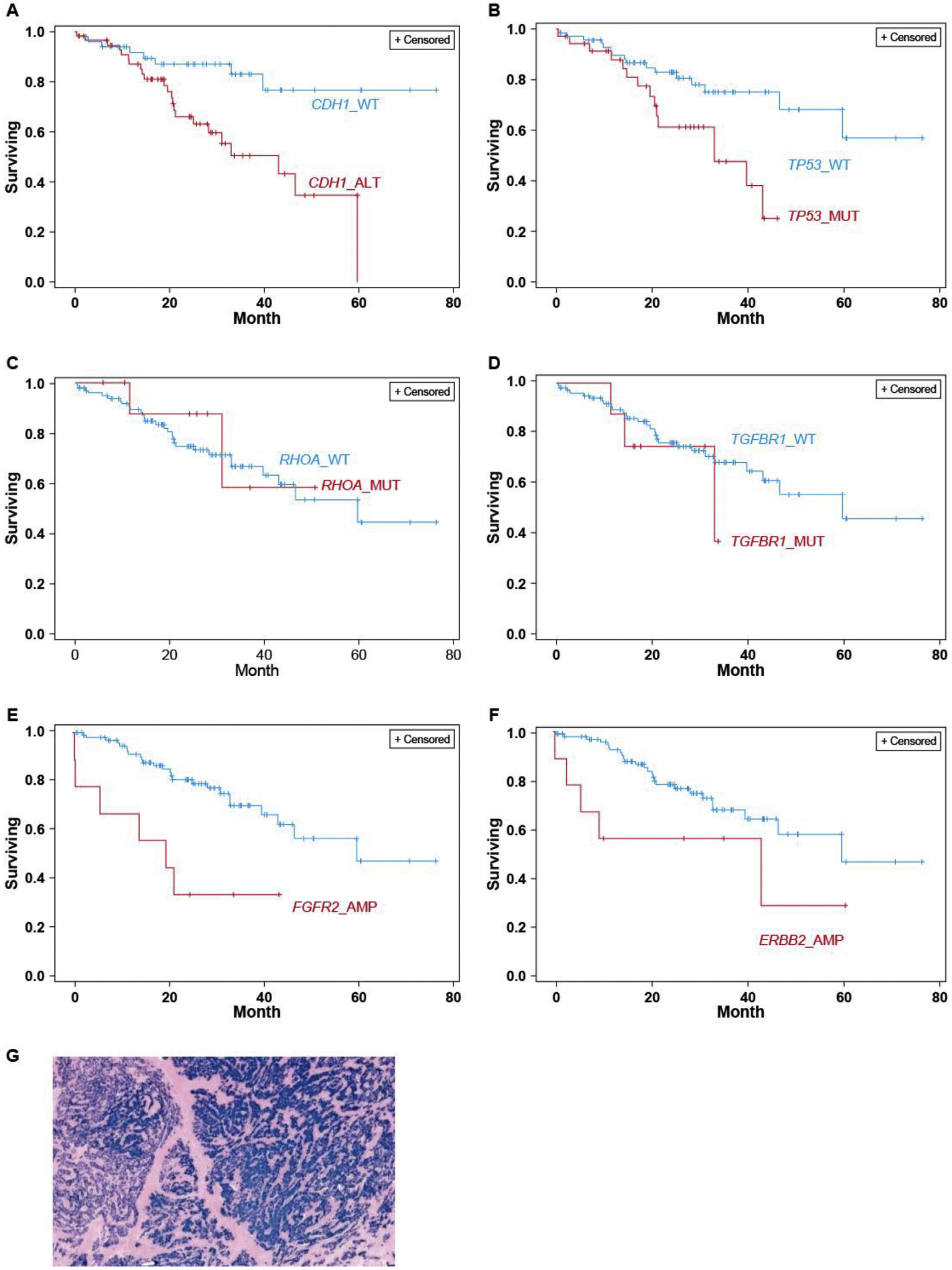
Supplementary Figure 5. Targeted sequencing and SNP6.0 array analyses. (*A*) Targeted sequencing coverages in EODGCs ($n = 109$) and LODGCs ($n = 115$). Targeted sequencing of tumor (mean coverage 960-fold, with 81% of target bases covered $>50\times$) and germline (mean coverage 908-fold, with 77% of target bases covered $>50\times$) samples targeted 38,510 bases in 10 genes. EODGCs ($n = 109$) and LODGCs ($n = 115$) were not different in coverage means (960× and 959x, respectively; $P = .91$, *t* test). Median coverage depths were 854× and 899× for EODGC and LODGC, respectively. A total of 346 somatic mutations (316 SNVs and 30 indels) were identified using targeted sequencing. (*B*) SNP6.0-based copy number clustering for tumors in the EODGC, LODGC, and TCGA cohorts. EODGCs and LODGCs were subjected to hierarchical clustering with 293 TCGA tumors whose CIN statuses were defined previously. *Red*, CIN TCGA tumors; *blue*, non-CIN TCGA tumors; *black*, EODGC; *green*, LODGC. The CIN statuses of new EODGCs and LODGCs were determined based on the cluster membership. (*C*) Arm-level, SNP6.0-based copy number profiles in EODGC (shown in *blue*) and LODGC (shown in *red*). EODGCs and LODGCs were not significantly different in arm-level copy number profiles. (*D*) Permutation *P* values for cross-validated misclassification rates (at feature selection $P < .05$) for the age cohort membership (EODGC vs LODGC), based on arm-level copy number profiles. According to prediction analyses built in the BRB-ArrayTools, permutation *P* values for cross-validated misclassification rates were .14, .15, .51, .14, .13, and .08 for the compound covariate predictor, the diagonal linear discriminant analysis classifier (LDA), the 1-nearest neighbor classifier (1-NN), the 3-nearest neighbor classifier (3-NN), the nearest centroid classifier (NC), and the support vector machines (SVM) classifier, respectively, at feature selection *P* values $<.05$. *Red dashed line*, cutoff for permutation *P* value ($P = .05$). These results indicate that EODGCs and LODGCs do not differ in arm-level copy number profiles.



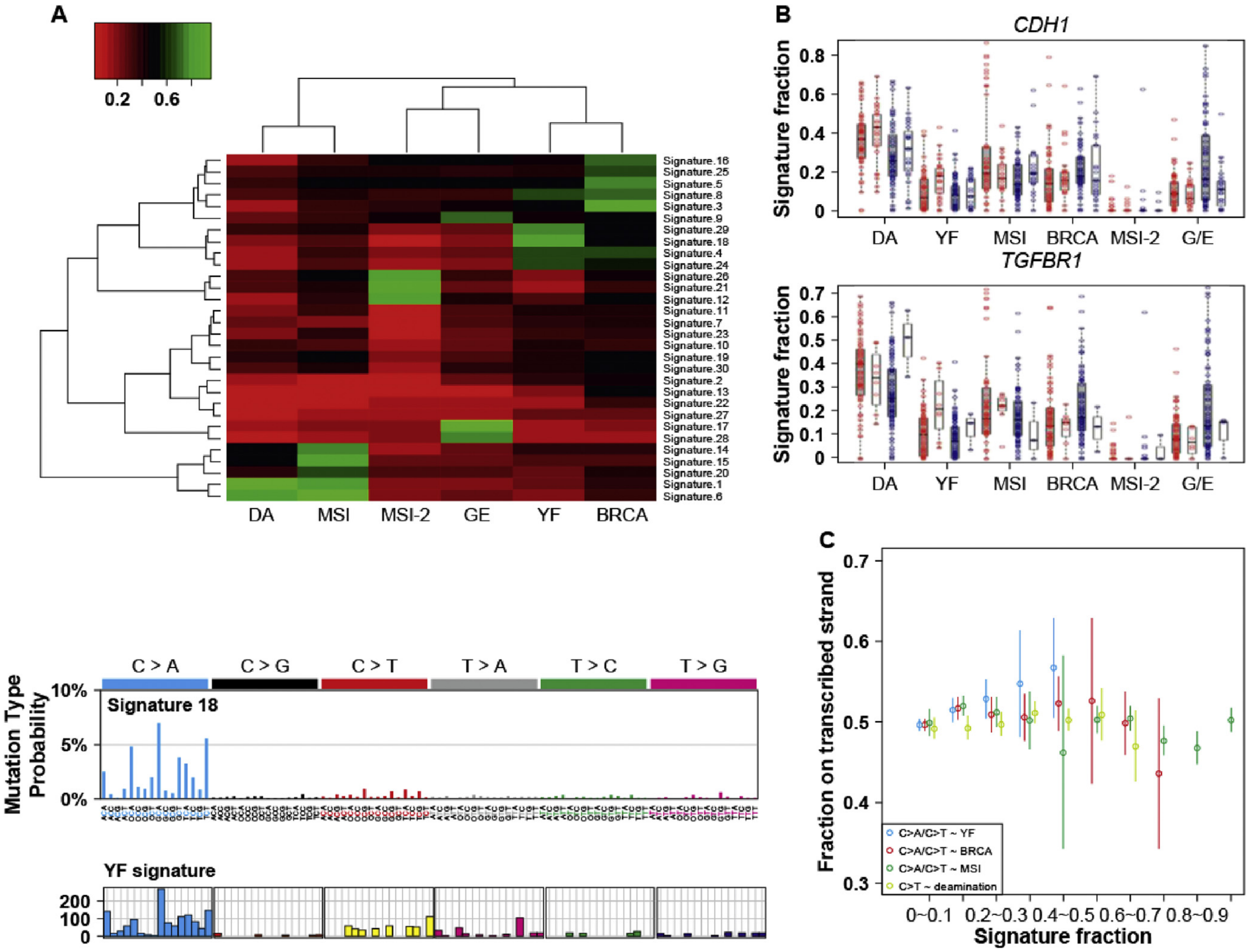
Supplementary Figure 6. The functional effects of *CDH1* variants (*A*) reverse transcription polymerase chain reaction (RT-PCR) for *CDH1* EC1 splice variants in EODGC. *Upper left*, exon 4 splice variants; *lower left*, exon 5 splice variants; upper right, exon 4 skipping (*asterisk*); *arrowheads*, RT-PCR bands for exon 4/5 truncations; *arrow*, an intron retention (*B*) RT-PCR for *CDH1* EC2/3 splice variants in EODGC. *Top*, exon 8 splice variants; *bottom*, exon 9 splice variants. EC2/3 splice variants included a transcript with exon 8/9 skipping (*arrowhead*) and wild-type *CDH1*. (*C*) Slow aggregation assays in mutant *CDH1*-overexpressing CHO-K1 cells. *Left*, *CDH1* germline mutation candidates (K182N, T340A, and K880K) overexpressing CHO-K1 cells aggregated as efficiently as did wild-type *CDH1*-overexpressing CHO-K1 cells. *Left upper panel*, representative images (50× magnification) showing aggregation of cells on soft agar at 24 hours after seeding. Scale bar = 1 mm. *Left lower panel*: mean diameters from triplicate experiments. Error bar = SD; ****P* < .001 (compared with the empty vector). *Right*, *CDH1* RT-PCR. (*D*) RT-PCR confirmation of *CDH1* overexpression in CHO-K1 cells used for cell aggregation assay. (*E*) Correlation between nuclear MYC and β-catenin immunostainings in 64 random GC tissue samples (*R* = 0.5, *P* < .0001, Pearson), suggesting the functional impact.


A
MKN-45
RHOA
β-actin
si-Control
si-RHOA#1
si-RHOA#2
B
si-Control
si-RHOA#1
SRF-RE activity
1.2
1.0
0.8
0.6
0.4
0.2
0
*
C
Vehicle
C3
kDa
GTP-RHOA
Total-RHOA
β-Actin
26
17
26
17
43
Vehicle
C3
SRF-RE activity
1.2
1.0
0.8
0.6
0.4
0.2
0
*
Vehicle
1000x
C3
1000
D
si-Control
si-RHOA
Vehicle
C3
E
si-Control
si-RHOA#1
TOPFlash activity
3.0
2.5
2.0
1.5
1.0
0.5
0
*
MKN-45
F
Vehicle
C3
TOPFlash activity
2.5
2.0
1.5
1.0
0.5
0
***
**
MKN-45
NCC-S1
G
si-Control
si-RHOA#1
Diameter(mm)
2.0
1.5
1.0
0.5
0
*
si-Control
si-RHOA#1
H
NCC-S1
Rhoa
E-cad
β-actin
28
130
43
si-Control
si-Rhoa
si-Cdh1
Pdx1-cre;Smad4F/F;Trp53F/F;Cdh1F/+
Rhoa
E-cad
β-actin
28
130
43
si-Control
si-Rhoa
si-Cdh1

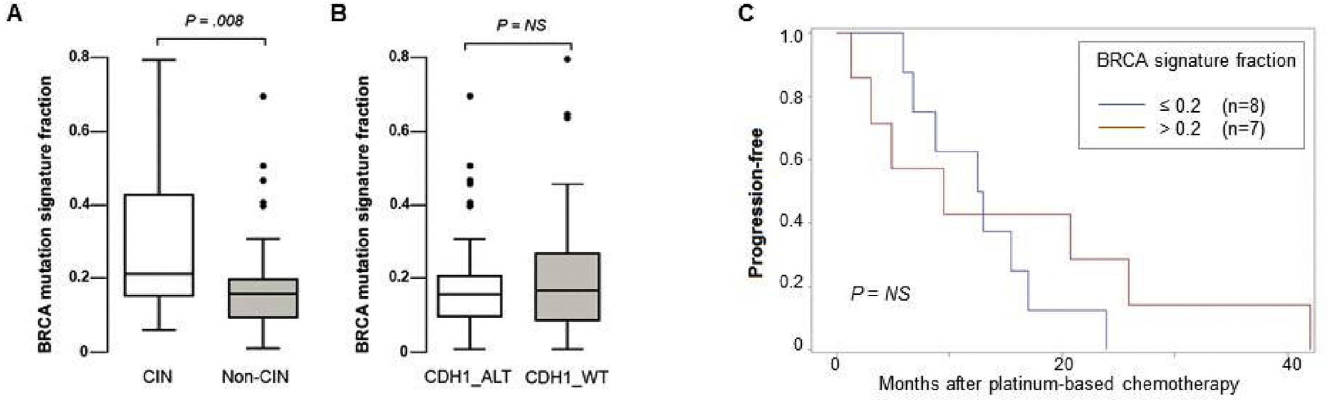
Supplementary Figure 7. RHOA functional assay. (*A*) Western blot validation of small interfering RNA (siRNA) knockdown for *RHOA* in MKN-45 cells. (*B*) The serum responsive factor (SRF)-RE reporter activities in MKN-45 GC cells after treatment with si-RHOA#1. (*C*) RHOA inactivation after C3 treatment. *Left*: RhoTekin assay after C3 treatment. Reduced GTP-bound RHOA fraction in MKN-45 cells after 24 hours of treatment with exoenzyme C3 transferase (1 μg/mL in serum-free RPMI-1640); *middle*: SRF-RE reporter activity in MKN-45 cells after C3 treatment. *$P < .05$, *t* test. *Right*: representative IF staining of actin stress fibers in MKN-45 cells after C3 treatment. Filamentous actin staining (shown in *red*) was reduced after RHOA suppression. The nucleus was visualized by 4′6-diamidino-2-phenylindole stain (*blue*) (magnification, 1000×). (*D*) Representative IF staining of actin stress fibers in NCC-S1 mouse GC cells after *Rhoa* knockdown and exoenzyme C3 transferase treatment. (*E*) The WNT/β-catenin reporter (TOPFlash reporter) activity in MKN-45 cells after treatment with si-RHOA #1. *$P < .05$, *t* test. (*F*) TOPFlash activities in MKN-45 and NCC-S1 cells after C3 treatment. **$P < .01$; ***$P < .001$. (*G*) Slow aggregation assays in MKN-45 after treatment with si-RHOA #1. Representative photo (scale bar = 0.5 mm) and mean diameters are shown. (*H*) Western blot validation of siRNA knockdown for *Rhoa* and *Cdh1* in NCC-S1 and *Pdx1-cre;Trp53*$^{F/F}$;*Cdh1*$^{F/+}$ primary cultured cells.


A
+ Censored
CDH1_WT
CDH1_ALT
Surviving
Month
B
+ Censored
TP53_WT
TP53_MUT
Surviving
Month
C
+ Censored
RHOA_WT
RHOA_MUT
Surviving
Month
D
+ Censored
TGFBR1_WT
TGFBR1_MUT
Surviving
Month
E
+ Censored
FGFR2_AMP
Surviving
Month
F
+ Censored
ERBB2_AMP
Surviving
Month
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
G

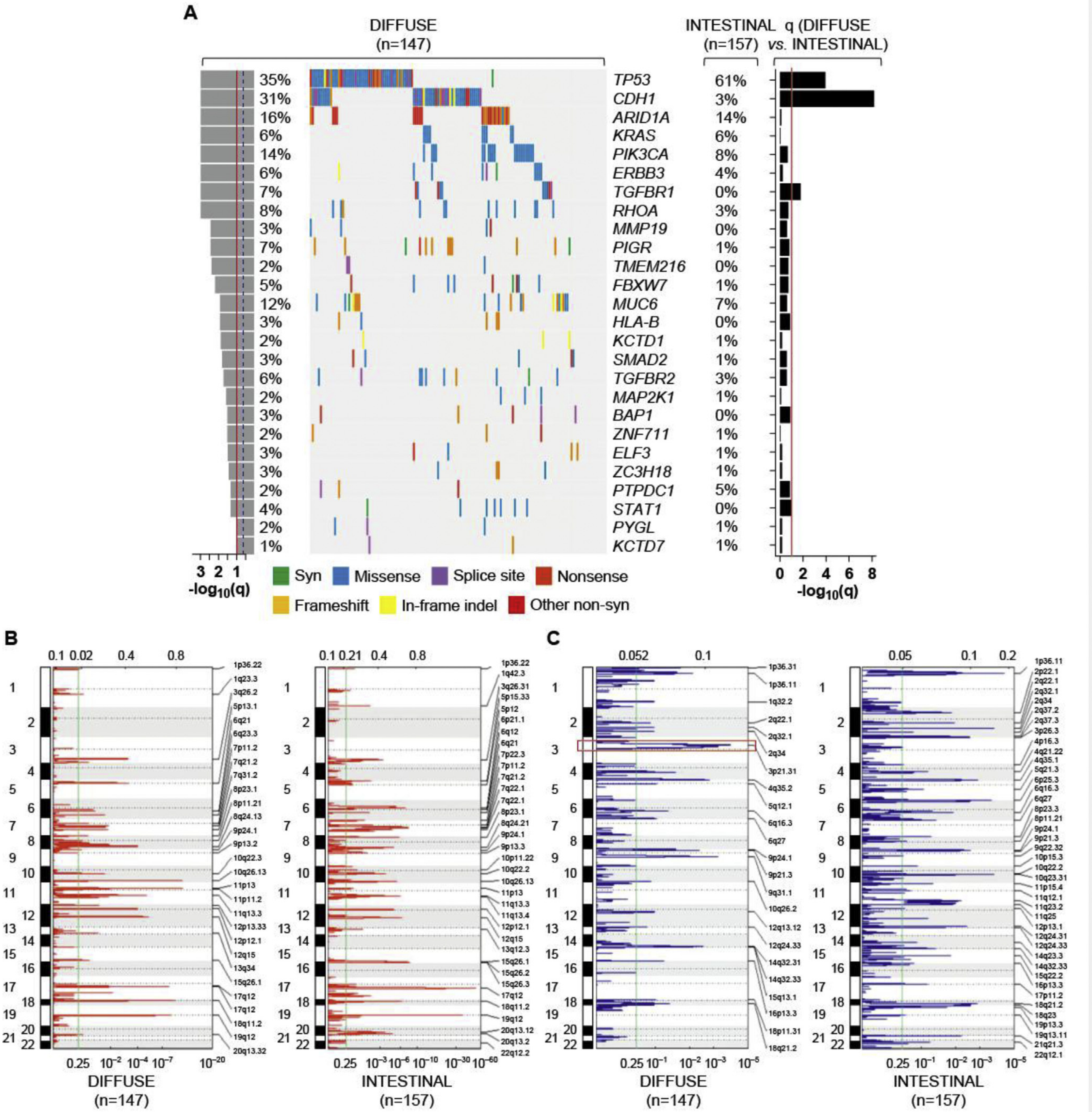
Supplementary Figure 8. Survival analyses. Prognostic values of the *CDH1* alteration (*A*), *TP53* (*B*), *RHOA* (*C*), and *TGBR1* (*D*) mutations, and *FGFR2* (*E*) and *ERBB2* (*F*) amplifications in EODGC. CIN and *CDH1* alterations were the most significant adverse prognostic biomarkers in EODGC, and the *TP53* mutation and *FGFR2* and *ERBB2* amplifications were also associated with poor prognosis. In contrast, the *RHOA* mutation (HR, 0.7; 95% CI, 0.2−3.0) and the *TGFBR1* mutation (HR, 1.7; 95% CI, 0.5−5.7) were not significantly associated with prognosis. The lack of poor prognostic value of RHOA mutations in our study is unlikely to result from overall paucity of RHOA mutations, given that the RHOA mutation was slightly more frequent in patients without distant metastasis (9.4% vs 8.3% in patients with metastasis) and was associated with an HR for mortality of 0.7. AMP, amplified; MUT, mutated; WT, wild-type. (*G*) Representative positive results of EBV ISH. EBV-positive tumor cells are shown in *blue*.



Supplementary Figure 9. Mutation signature analysis (*A*) Similarity of YF signature to signature 18. *Top*, YF signature, which is dominated by C>A mutations, most resembles COSMIC signature 18; *bottom*, comparison between COSMIC signature 18 (*upper panel*) and the YF signature (*lower panel*) (*B*) The fraction of mutation signature according to mutation statuses of *CDH1* and *TGFBR1*. *Gray*, wild-type tumors; *white*, mutated tumors; *red*, female; *blue*, male. (*C*) Transcribed strand bias of the YF signature.



Supplementary Figure 10. BRCA mutation signature. (*A*) The fraction of BRCA mutation signature was different between the CIN subgroup and non-CIN subgroup (including EBV, MSI, and GS subgroups) of EODGC-WES (P = .008, Wilcoxon). (*B*) The fraction of BRCA mutation signature was not different between EODGC-WES tumors with *CDH1* alteration and those wild-type (WT) for *CDH1*. ALT, alteration (P = NS, Wilcoxon). (*C*) *Red*, EODGC-WES patients with tumors having low ($\leq$0.2) BRCA mutation signature fraction; *blue*, EODGC-WES patients with tumors having high (>0.2) BRCA mutation signature fraction. BRCA mutation signature fraction was not associated with progression-free survival after platinum-containing chemotherapy in this small subset of patients (P = .51, log-rank).



Supplementary Figure 11. Analysis of the WES data of diffuse tumors (n = 147) and non-hypermutated TCGA intestinal tumors (n = 157). (A) Co-mutation plot for mutations that were significant within 147 diffuse tumors, which include all the non-hypermutated TCGA diffuse tumors and EODGC-WES. *Continuous red lines* indicate significance cutoffs for the difference between diffuse and intestinal tumors (*q* = .1). Mutation rates for *TP53, CDH1,* and *TGFBR1* were significantly different between diffuse and intestinal tumors. (*B*) Recurrent amplifications in diffuse (*left*) and intestinal tumors (*right*), as identified by ReCapSeg and ISAR-GISTIC. In diffuse tumors, 10q26.13, 11p13, 18q11.2, 19q12, 17q12, 12q15, 11q13.3, 8q24.13, 12p12.1, 5p13.1, 3q26.2, 6q23.3, 10q22.3, 7q21.2, 13q34, 17q12, 12p13.33, and 6q21 regions were significantly amplified. *FGFR2, CD44, GATA6, CCNE1, ERBB2, CCND1, KRAS*, and *CCND2* were included in these loci. In intestinal tumors, 17q12, 19q12, 6p21.1, 12q15, 18q11.2, 8p23.1, 11q13.3, 15q26.1, 12p12.1, 8q24.21, 10q26.13, 7q21.2, 10q22.2, 20q13.2, 3q26.31, 7q22.1, 7p11.2, 1q42.3, 6q21, 7q22.1, 9p13.3, 20q13.12, 7p22.3, 13q12.3, 5p15.33, 10p11.22 and 1p36.22 were significantly amplified. *ERBB2, CCNE1, VEGFA, GATA6, GATA4, CCND1, KRAS, MYC, FGFR2, SMAD9, SKIL* and *EGFR* were included in these loci. *CD44* was significantly amplified in diffuse tumors, but not in intestinal tumors. (*C*) Recurrent deletions in diffuse (*left*) and intestinal tumors (*right*). Fourteen loci, including 3p21.31, 9q31.1, 4q35.2, 1p36.11, 14q32.31, 9p21.3, 18q21.2, 6q16.3, 1p36.31, 16p13.3, 5q12.1, 6q27, 12q13.12, and 18p11.31, were significantly deleted in diffuse tumors. Chromosomal loci 1p36.11, 10q23.31, 6p25.3, 2q32.1, 4q35.1, 9p24.1, 18q21.2, 21q21.3, 6q16.3, 2p22.1, 12q24.31, 15q22.2, 19p13.3, 2q37.3, 6q27, 11q23.2, 14q32.33, 9p21.3, 11p15.4, 8p23.3, 12q24.33, 22q12.1, 2q22.1, 2q37.2, 17p11.2, 5q21.3, and 3p26.3 were significantly deleted in intestinal tumors. *BAP1* locus was significantly deleted only in diffuse tumors (shown in *red box*), whereas *PTEN, ATM, CDKN2A* and *CHEK2* were recurrent deletions only in intestinal tumors.

Supplementary Table 8. Genes That Were Contained in 3p21.1 (g.chr3:48,369,660–55,002,466) Locus and That Were Significantly Underexpressed in Early-Onset Diffuse Gastric Cancer Whole Exome Sequencing Tumors With 3p21.1 Deletion Compared With Early-Onset Diffuse Gastric Cancer Whole Exome Sequencing Tumors Without 3p21.1 Deletion ($q < .1$ and Fold-Change < -0.5)

		RPKM		
Gene	*q* Value	With deletion	Without deletion	FC
RPL29	8×10^{-7}	101.7	161.4	−0.7
TMEM115	5×10^{-5}	13	19.1	−0.5
TWF2	.0001	16.5	23.4	−0.5
BAP1	.0011	9.2	13.8	−0.6
ABHD14A	.0036	9.1	13.6	−0.6
ACY1	.0066	8.7	12.9	−0.6

FC, fold change in gene expression (reads per kilobase of exon per million mapped reads [RPKM]) of EODGC-WES tumors with 3p21.1 (g.chr3: 48,369,660-55,002,466) deletion as compared with EODGC-WES tumors without 3p21.1 (g.chr3:48369832-55002466) deletion.

Supplementary Table 9. Top 20 Significant Mutations in Early-Onset Diffuse Gastric Cancer Whole Exome Sequencing According to MutSig2CV Analysis

Rank	Gene	npat	*q* Value
1	*CDH1*	28	9×10^{-13}
2	*TP53*	23	9×10^{-13}
3	*ARID1A*	12	1×10^{-11}
4	*KRAS*	5	3×10^{-6}
5	*PIK3CA*	12	9×10^{-6}
6	*ERBB3*	7	2×10^{-5}
7	*TGFBR1*	7	9×10^{-5}
8	*FBXW7*	8	4×10^{-4}
9	*RHOA*	4	.02
10	*MAP2K1*	3	.04
11	*ELFN2*	4	.16
12	*SLC25A5*	2	.16
13	*SLC38A9*	3	.16
14	*SHANK3*	4	.16
15	*RFC4*	3	.16
16	*BAP1*	4	.27
17	*CXCR3*	2	.37
18	*PRSS3*	4	.37
19	*KCTD7*	2	.37
20	*SLC2A1*	3	.41

Supplementary Table 10. Cox Proportional Hazards Regression Analysis of Progression-Free Survival After Platinum-Containing Chemotherapy After Relapse

Mutation signature	*P*[a]	HR (95% CI)
Deamination	.67	0.32 (0.002–60.41)
YF	.13	1821.7 (0.11–30,464,974)
MSI	.19	46.1 (0.16–13,760)
BRCA	.45	0.15 (0.001–20.8)
MSI-2	.88	2.9 (0–3,174,752)
Gastroesophageal	.28	0.02 (0–24.9)

[a] $Pr > \chi^2$.

Supplementary Table 11. Multiple Regression Analysis of Nuclear β-Catenin

Source	DF	Type III SS	Mean square	*F* value	$Pr > F$
RHOA mutation	1	3857.468778	3857.468778	6.74	.0102
CDH1 alteration	1	4682.416911	4682.416911	8.18	.0047

DF, degrees of freedom; SS, sum of squares.