CANCER

Defined lifestyle and germline factors predispose Asian populations to gastric cancer

Akihiro Suzuki[1,2], Hiroto Katoh[3,4], Daisuke Komura[3,4], Miwako Kakiuchi[1], Amane Tagashira[1,5], Shogo Yamamoto[1], Kenji Tatsuno[1], Hiroki Ueda[1], Genta Nagae[1], Shiro Fukuda[1], Takayoshi Umeda[1], Yasushi Totoki[6], Hiroyuki Abe[5], Tetsuo Ushiku[5], Tetsuya Matsuura[2], Eiji Sakai[2], Takashi Ohshima[7], Sachiyo Nomura[8], Yasuyuki Seto[8], Tatsuhiro Shibata[6,9], Yasushi Rino[7], Atsushi Nakajima[2], Masashi Fukayama[5], Shumpei Ishikawa[3,4]*, Hiroyuki Aburatani[1]*

Germline and environmental effects on the development of gastric cancers (GC) and their ethnic differences have been poorly understood. Here, we performed genomic-scale trans-ethnic analysis of 531 GCs (319 Asian and 212 non-Asians). There was one distinct GC subclass with clear alcohol-associated mutation signature and strong Asian specificity, almost all of which were attributable to alcohol intake behavior, smoking habit, and Asian-specific defective *ALDH2* allele. Alcohol-related GCs have low mutation burden and characteristic immunological profiles. In addition, we found frequent (7.4%) germline *CDH1* variants among Japanese GCs, most of which were attributed to a few recurrent single-nucleotide variants shared by Japanese and Koreans, suggesting the existence of common ancestral events among East Asians. Specifically, approximately one-fifth of diffuse-type GCs were attributable to the combination of alcohol intake and defective *ALDH2* allele or to *CDH1* variants. These results revealed uncharacterized impacts of germline variants and lifestyles in the high incidence areas.

INTRODUCTION

Gastric cancer (GC) is the third leading cause of cancer mortality worldwide (*1*). The GC incidence shows strong geographic variations, with the highest incidence in East Asia, especially Japan and Korea (*2*). Although there are several reports that suggest hereditary and lifestyle factors, including strains of *Helicobacter pylori* (*H. pylori*), as contributors to these geographic variations (*3*, *4*), no reports so far have analyzed these factors with genomic resolution in a trans-ethnic manner. Although recent whole-genome/exome sequencing has contributed to the global characterization of somatic genetics and driver genes of GCs (*5–8*), the precise interplays among lifestyles, germline variations, and somatic mutations, especially in the contexts of ethnic variations, are not fully clarified to date. To elucidate this point, we conducted whole-exome sequencing for a large cohort of GCs with detailed etiological information among Japanese, a population with one of the highest incidences of GCs in the world. These results were integrated with The Cancer Genome Atlas (TCGA) dataset to conduct trans-ethnic analysis of somatic and germline genetics of GCs.

[1]Genome Science Division, Research Center for Advanced Science and Technology (RCAST), The University of Tokyo, Tokyo, Japan. [2]Department of Gastroenterology and Hepatology, Yokohama City University Graduate School of Medicine, Kanagawa, Japan. [3]Department of Preventive Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [4]Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan. [5]Department of Pathology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [6]Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan. [7]Department of Surgery, Yokohama City University Graduate School of Medicine, Kanagawa, Japan. [8]Department of Gastrointestinal Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [9]Laboratory of Molecular Medicine, Institute of Medical Science, The University of Tokyo, Tokyo, Japan.
*Corresponding author. Email: haburata-tky@umin.ac.jp (H.A.); ishum-prm@m.u-tokyo.ac.jp (S.I.)

RESULTS

Trans-ethnic mutational signature of GC

The overall profile of the whole-exome sequencing for 531 GC cases (288-case TCGA dataset and our 243-case Japanese dataset) exhibited substantial diversity in the somatic mutation profiles among GCs (Fig. 1A). Comparing the distributions of GC subtypes from the perspective of the TCGA subtypes (*9*), Epstein-Barr virus (EBV)–associated GC and microsatellite instability (MSI)–GC were more frequent in the TCGA dataset, while genomically stable (GS)–GC was more frequent in the Japanese cohort (table S1). Cancer genomes are known to have acquired various, but specific, somatic mutational patterns during carcinogenesis depending on the carcinogenic etiologies they experienced, recently called as mutational "Signatures" (*10*). We used a pan-cancer catalog of 30 Signatures from the COSMIC (Catalogue Of Somatic Mutations In Cancer) database (*11*) for this analysis (see Materials and Methods). According to the hierarchical clustering based on the mutational signature contributions in each case (see Materials and Methods), GCs were subdivided into six groups (marked by color bars at the top of the heat map), each of which exhibited dominant contribution of Signatures 1, 3, 6, 15, 16, 17, and others (Fig. 1A). Single-nucleotide variants (SNVs) attributable to Signatures 1, 3, 6, and 15 are known to be associated with the aging of the patients, *BRCA* deficiency, and mismatch repair (MMR) deficiency, respectively (*10*). To the contrary, theoretical etiologies of Signature 17 have not been determined to date.

To identify any unknown interplays between SNV Signatures and other factors, we examined possible clinicopathological factors in each of the clusters, including patient races and well-known germline variants. We then found that a subgroup (Fig. 1A, orange bar) was strongly contributed by Signature 16, whose contributions consisted of approximately 28% of somatic SNVs, and almost all the
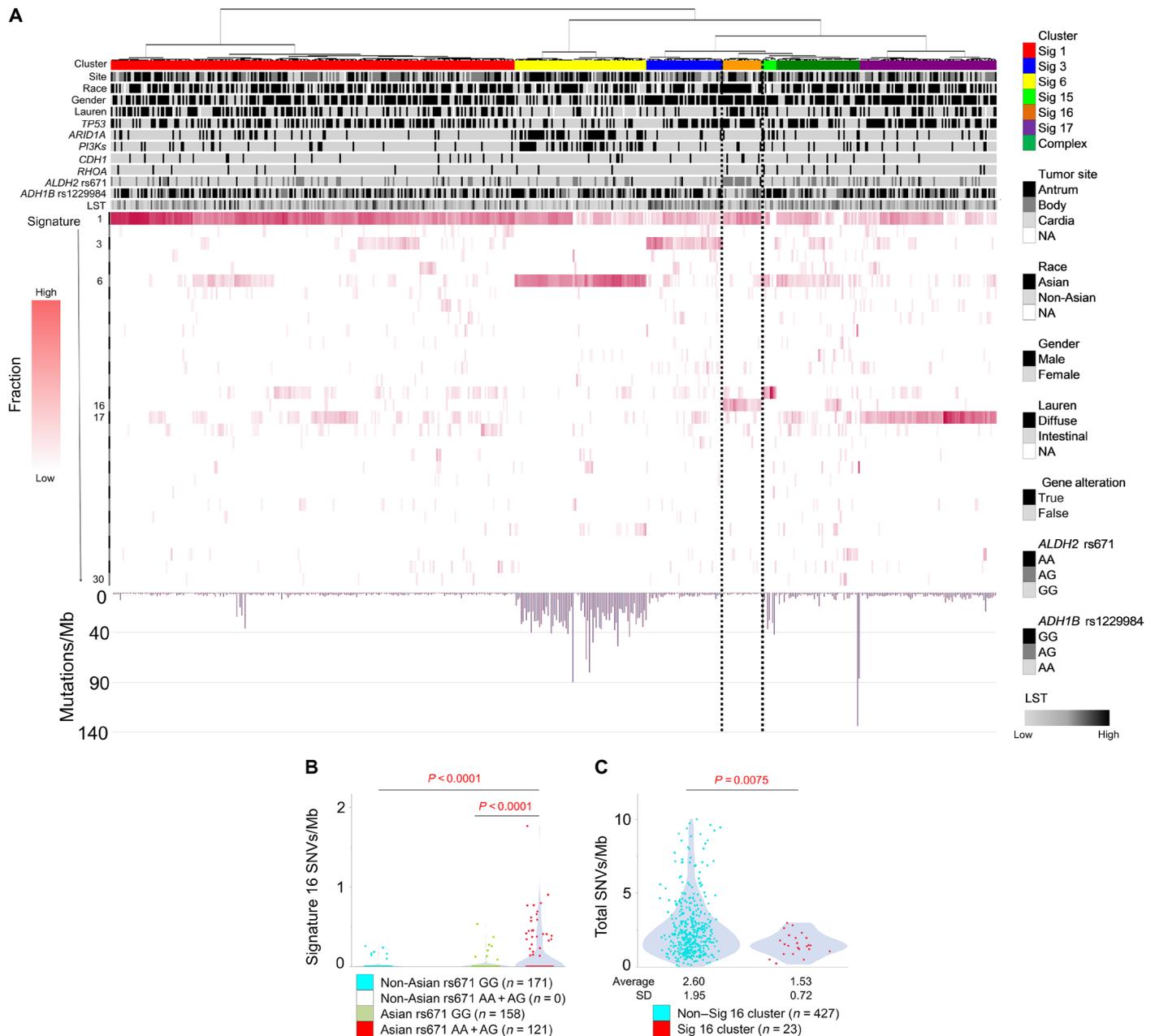
Fig. 1. Trans-ethnic mutational signature analysis of GC. (**A**) The overall genetic profiles of 531 GCs, including Japanese GC cohort and TCGA GC datasets, are shown, on the basis of the hierarchical clustering of mutational signatures. GCs were genetically subdivided into six subgroups, each of which is characterized by Signatures 1 (red), 6 (yellow), 3 (blue), 16 (orange, surrounded by the dotted line), 15 (green), and 17 (purple). Clinicopathological information, representative somatic driver gene alterations, and germline variations of *ALDH2* rs671/*ADH1B* rs1229984 are indicated as black/white columns at the top. White-to-red colored columns in the hierarchical clustering map represent the contribution rates (red bar on the left side) of the mutational signatures in each case. Mutation frequencies per megabase (Mb) are indicated at the bottom as bar graphs. *PI3Ks*: *PIK3CA* and *PTEN* mutations. (**B**) The numbers of the Signature 16 SNVs/Mb in GC cases are shown in relation to the races and *ALDH2* genotypes of the patients. *P* values were calculated using unpaired Wilcoxon rank sum test. (**C**) The numbers of total SNVs/Mb in cases of Signature 16 cluster and those of other clusters are shown. Cases with hypermutator signature (yellow and green bars) were excluded. *P* values were calculated using the unpaired Wilcoxon rank sum test. NA, not applicable.

patients in this cluster were of Asian ethnic background (90.5%), and a high proportion of these patients (16 of 23 = 69.6%) harbored a well-known inactive *ALDH2* allele (rs671 AA or AG) (table S2). We further confirmed this observation of the correlation between Signature 16 and the *ALDH2* allele by analyzing the Signature 16

contributions and *ALDH2* genotypes in all the 531 cases, finding their positive correlation in an Asian-specific manner (Fig. 1B; *P* < 0.0001, unpaired Wilcoxon rank sum test). As is reported (*12*), inactive *ALDH2* rs671-A allele is specific to Asian populations; thus, the correlation of the *ALDH2* genotype with Signature 16

contribution is considered attributable specifically in the Asian populations.

Although the GC cluster with Signature 16 had no characteristic patterns of major driver gene mutations, this cluster was relatively enriched for diffuse-type histology (12 of 23) (Fig. 1A and table S2). The overall mutation burdens in this cluster were significantly smaller than those of the other GCs ($P = 0.0075$; Fig. 1, A and C), while the age at onset and Signature 1 mutations of the GC cases in the Signature 16 cluster were comparable to the other GCs (fig. S1). In view of chromosomal aberrations, we compared large-scale state transition (LST) scores (13) between all the Signature clusters (fig. S2), revealing that the LST scores of Signature 16 cluster were among the lowest, except for hypermutators (Signature 6/15 clusters) that are already known to harbor fewer chromosomal abnormalities (5, 9) (fig. S2). Together, the Signature 16 cluster is suggested not only as an Asian-specific subclass but also as a distinct biological entity.

Etiological factors of Signature 16 GC

With these findings and recent reports showing that Signature 16 is suggestive of an alcohol-associated signature in liver, esophageal, and several other cancers (14–16), we then performed an in-depth analysis of our Japanese cohort with more information of their lifestyles and etiologies to investigate the relationship among Signature 16, patient ALDH2 genotypes, and alcohol intake. We also analyzed how strongly the combination of germline genetics and lifestyles affects the high-GC incidences in the Japanese population. The existence of the Signature 16 GC cluster was clearly reconfirmed (Fig. 2A), and the cluster contained 6.6% (16 of 243) of Japanese GCs. A high portion of patients (11 of 16, 68.8%) in the Signature 16 GC cluster were characterized as alcohol consumers with inactive ALDH2 allele (table S3). Therefore, the Signature 16 mutations in the GC context were most probably attributable to the combined effects of alcohol intake and loss-of-function allele of ALDH2 (table S3). To quantitatively confirm this phenomenon, we investigated the correlation between the number of Signature 16 SNVs and the combination of the alcohol intake habit and ALDH2 allele among all the Japanese GCs (Fig. 2B). While the effects of either alcohol intake or the ALDH2 allele alone were minimal on the number of Signature 16 mutations, the combination of both the alcohol intake behavior and inactive ALDH2 allele resulted in an 11.1-fold increase of the Signature 16 burdens ($P < 0.0001$, unpaired Wilcoxon rank sum test; Fig. 2B). ALDH2 is an enzyme that metabolizes acetaldehyde to acetic acid, the former of which exhibits significant genotoxic effects (17). As reported, the gastric mucosa expresses ALDH2, and the acetaldehyde levels in the gastric contents of ALDH2-inactive individuals were 5.6 times higher than those in the ALDH2-active individuals, after intragastric alcohol infusion (18, 19). Therefore, the accumulated acetaldehyde due to the low ALDH2 activity most likely induced the Signature 16 mutation patterns in gastric epithelial cells. The Signature 16 SNV occurred at the transcribed strands significantly more frequently than at the untranscribed strands (fig. S3; $P < 0.0001$, Fisher's exact test), suggesting its association with transcription-coupled nucleotide excision repair (20). In summary, we conclude that the mutational Signature 16 significantly contributes to GCs in Asian populations and is attributable to the combined effects of the germline factor and patients' lifestyle, which are alcohol intake behavior and loss-of-function allele of ALDH2. It was revealed that patients with GC

with high numbers of Signature 16 SNVs did not always consume large amounts of alcohol; therefore, it can be concluded that a large amount of alcohol is not necessarily required to induce Signature 16 SNVs in ALDH2-inactive individuals. Mutational Signature 16 is also observed in esophageal squamous cell carcinoma and is related to alcohol intake with inactive ADH1B genotypes (rs1229984-GG) (16). In our Japanese GC cohort, however, such a correlation was not statistically confirmed (fig. S4, A and B), partly because of the limited statistical power due to the low minor allele frequency (MAF) of the ADH1B rs1229984 locus.

It is well known that alcohol consumption together with smoking epidemiologically increases the risk of esophageal squamous cell carcinoma (21). Conversely, it is not fully evident whether or not smoking synergizes with alcohol consumption regarding the GC risk. Here, detailed etiological analysis using Japanese GC cohort revealed that smoking habit has synergistic effects on Signature 16 mutations, specifically among GCs of alcohol consumers with ALDH2 defective allele (rs671 AA/AG) ($P = 0.0339$; Fig. 2C). This synergic effect was not observed among patients with ALDH2-proficient genotypes (rs671 GG) nor was there an increase in the numbers of Signature 16 mutations among smokers without drinking habit. Therefore, it can be considered that combination of alcohol consumption and smoking habit has synergic effects on the generation of Signature 16 SNVs, specifically among people with ALDH2 defective allele. To the contrary, the so-called smoking signature (Signature 4) was not apparently dominant among patients in the Signature 16 cluster (Figs. 1A and 2A), and it was also revealed that Signature 4 SNVs were not synergized by alcohol consumption (fig. S4C), indicating a possible one-side effect of smoking habit on the generation of alcohol-related signatures during GC carcinogenesis.

Immunological features of alcohol-related GC

Because of the distinct nature of the Signature 16 GCs, especially their smaller mutation burdens, we were then motivated to investigate whether this GC subgroup has characteristic biology including immunological features (22). Bioinformatics approach using gene expression profiles of 184 (of 243 cases of exome-sequenced) Japanese GCs pointed out that Signature 16 GCs exhibited characteristic immunological microenvironments (Fig. 3A), where we estimated compositions of 22 types of tumor-infiltrating immune cells in each GC using CIBERSORT (cell-type identification by estimating relative subsets of RNA transcripts) algorism (see Materials and Methods) (23). It was revealed that Signature 16 GCs showed skewed distributions in principal components analysis in view of the compositions of infiltrating immune cells (Fig. 3A), and higher B cell infiltration was the most prominent feature of the Signature 16 GCs (Fig. 3B). This trend of B cell infiltration was commonly observed among Signature 16 GCs regardless of the tumor subtypes, either diffuse or intestinal types. It was also noted that expression of CXCL13, a well-established B cell–recruiting chemokine, was significantly higher among Signature 16 GCs compared to the other groups of GCs (Fig. 3C). To confirm these observations, we histopathologically evaluated the B cell infiltration and CXCL13 expression in Signature 16 GCs. It was revealed that cancer cells themselves, in addition to well-characterized follicular dendritic cells, expressed CXCL13, and in total, 68.8% (11 of 16) of Japanese Signature 16 GCs showed immunohistochemically recognizable positivity for CXCL13 in cancer cells, either in a focal or wide-spread manner (Fig. 3D). In these
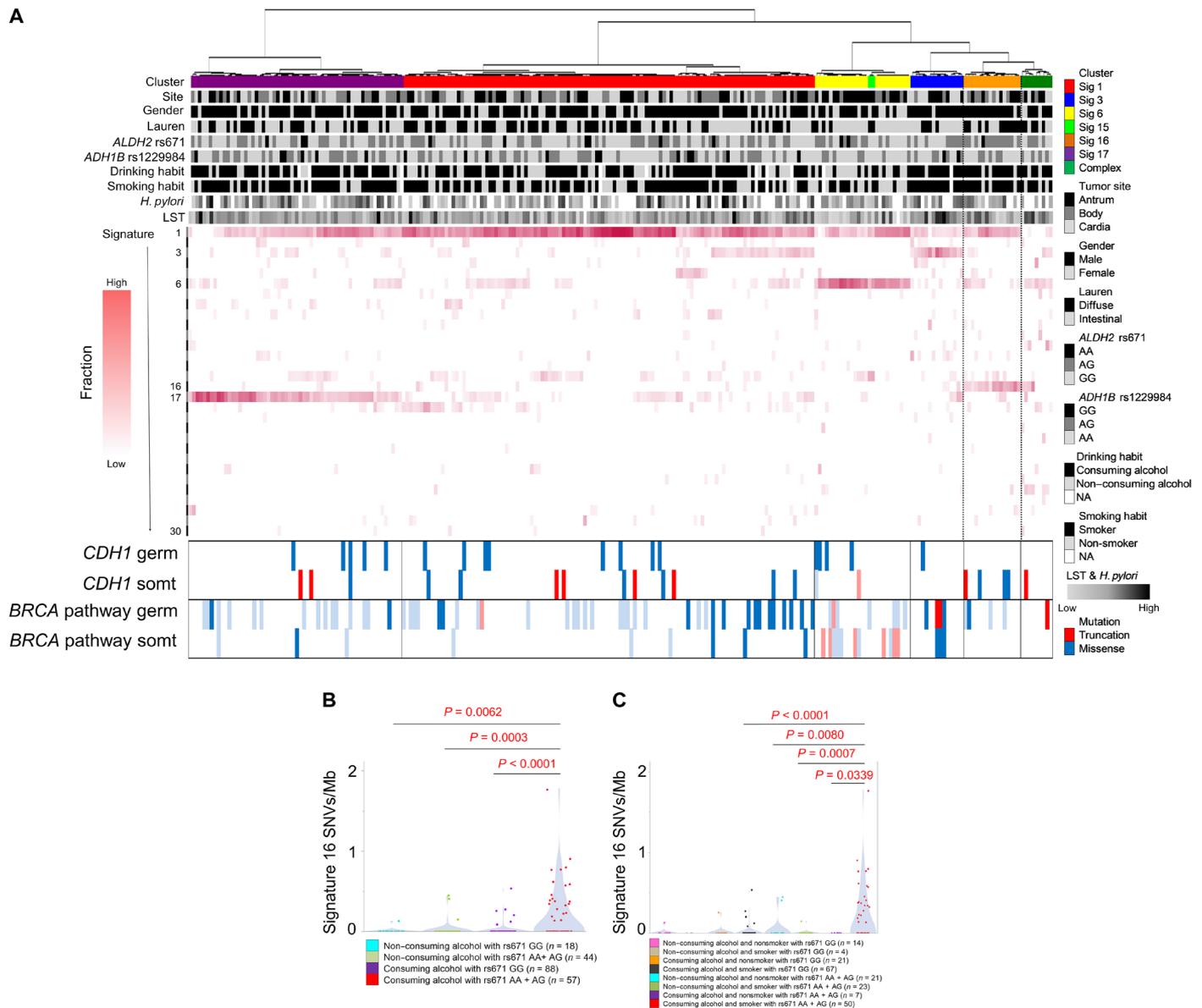
Fig. 2. Mutational signature analysis of Japanese GCs with lifestyle and germline information. (**A**) The overall genetic profiles of 243 Japanese GCs are shown as a hierarchical clustering heat map as in Fig. 1A. Clinicopathological information, germline variations of *ALDH2* rs671 and *ADH1B* rs1229984, and alcohol and smoking habits are indicated as black/white columns at the top. The contribution rate of the mutational signatures in each case is shown as in Fig. 1A. The germline and somatic variations of the *CDH1* and *BRCA* pathway genes are indicated at the bottom of the figure. Red and blue columns indicate truncation and missense variants, respectively, with the proviso that somatic mutations in hypermutators (GCs with Signature 6 contribution, yellow column) and *BRCA* pathway variations without *BRCA* signatures (both somatic and germline) are shown as transparent columns. (**B**) The numbers of Signature 16 SNVs are plotted according to the patient subgroups defined by *ALDH2* genotype and alcohol consumption habit. *P* values were calculated using unpaired Wilcoxon rank sum test. (**C**) The numbers of Signature 16 SNVs are plotted according to the patient subgroups defined by the combinations of *ALDH2* genotype, alcohol consumption, and smoking habit. *P* values were calculated using the unpaired Wilcoxon rank sum test.

cases, cancer cell nests with CXCL13 positivity were occasionally accompanied with B cell infiltrations around the nests. Representatively, one case showed massive infiltration of B cells within the cancer area as well as high-level expression of cancer cell–intrinsic CXCL13 (Fig. 3D). Pathway enrichment analysis using RNA sequencing (RNA-seq) data (see Materials and Methods) revealed significant enrichment of B cell receptor signaling, cytokine/chemokine signal, and Fc γ receptor signaling pathways in the Signature 16 GCs compared to others (table S4), which further sup-

ports the characteristic B cell infiltrations in Signature 16 GCs. In summary, it is suggested that Signature 16 GCs frequently express CXCL13 in a cancer cell–intrinsic manner and thus recruit the infiltration of B cells within cancer area.

Impact of germline variations in Asian GC
Germline variations of E-cadherin (*CDH1*) are well known to be responsible for hereditary diffuse-type GC (HDGC) (*4*), and recent exome sequencing identified germline mutations of the *BRCA1/2*
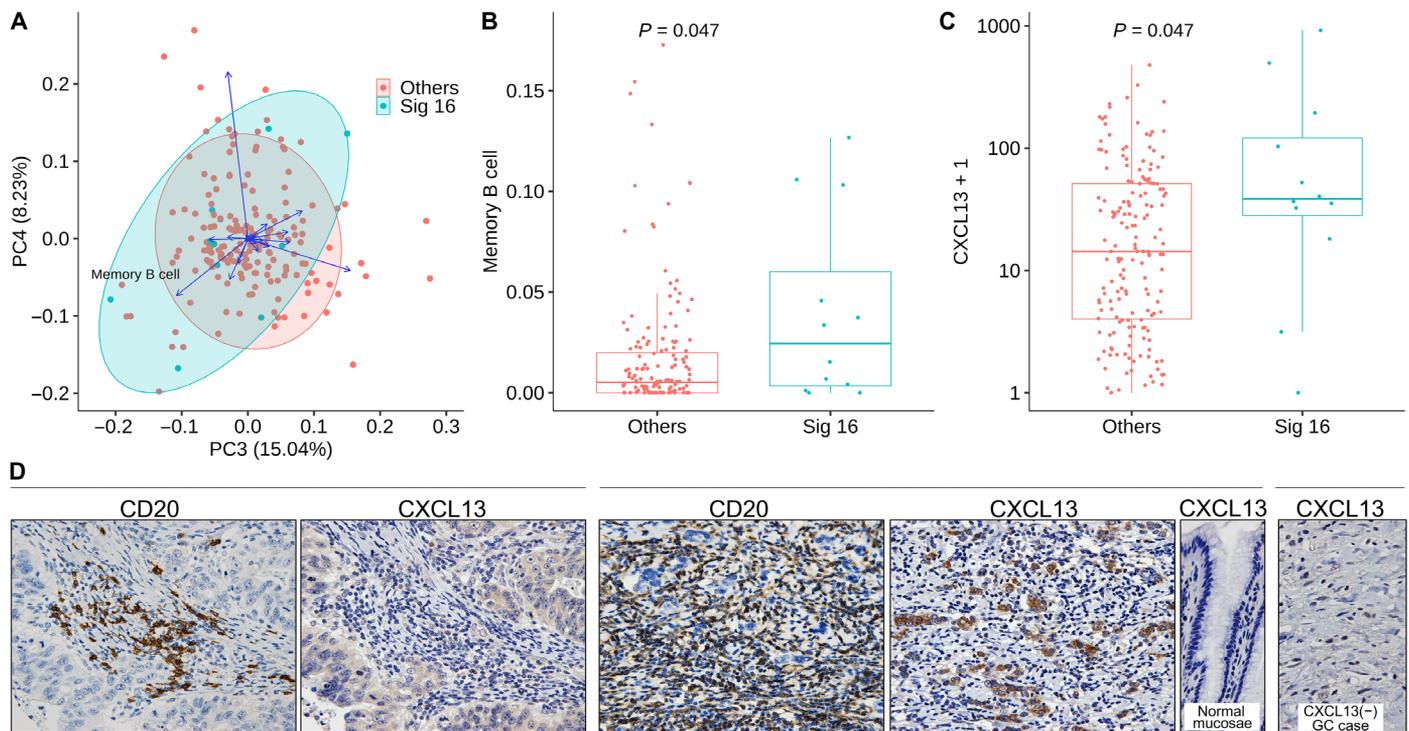
Fig. 3. Cancer-recruited B cell infiltration in alcohol-related GCs. (**A**) Principal components analysis of 184 Japanese GC cases based on their profiles of the proportions of tumor-infiltrating immune cells defined by CIBERSORT algorism (see Materials and Methods). PC3 and PC4 components are shown. Green and red circles indicate Signature 16 GC and other GCs, respectively. Green and red circular areas indicate 95% confidence intervals of Signature 16 GC and other GCs, respectively. Arrows represent the correlations between the principal components (PC) and the variables (compositions of immune cells). (**B**) The proportion of B cells infiltrating in the GC tissues determined by the CIBERSORT algorism (see Materials and Methods) is shown. Green and red dots indicate Signature 16 GCs and other GCs, respectively. *P* value was calculated by unpaired Wilcoxon rank sum test. (**C**) Gene expression levels of B cell–attracting chemokine CXCL13 determined by RNA sequencing of bulk GC tissues (see Materials and Methods) are shown. *Y* axes are shown as log scale. Green and red dots indicate Signature 16 GCs and other GCs, respectively. *P* value was calculated by unpaired Wilcoxon rank sum test. (**D**) Representative cases of Signature 16 GCs with immunohistochemical staining of CD20 and CXCL13 are shown. In a GC case on the left side, as in other Signature 16 GCs, cancer cell–intrinsic CXCL13 expression and B cell infiltration around the tumor nest are observed. The GC case shown in the middle panels represents the prominent expression of tumor cell–intrinsic CXCL13 expression and massive infiltrations of B cells, followed by a picture at the right side, indicating negative CXCL13 staining in the normal gastric mucosae of the same specimen, as a control. A GC case at the rightest side indicates negative staining of CXCL13. Black bars under pictures, 50 μm.

pathway genes as causative factors for GCs especially with familial histories (*24*). However, most of these data were obtained in non-Asian cases. Meanwhile, in Asian countries, the global burdens of GC-causing germline variants and their effects on familial aggregations have remained unclear to date. We summarized possible pathogenic germline variants of 624 cancer-related genes (*25*) in our Japanese GC cohort after filtering out the common and/or presumably nonpathogenic variants in the Japanese population (see Materials and Methods). While several truncating variants were observed in the *BRCA* pathway genes, we unexpectedly found that the *CDH1* gene has the highest variant frequency and density (Table 1, bold), suggesting that the germline variants of the *CDH1* gene significantly contribute to the genetic backgrounds of GCs in the Japanese population. Among Genome-Wide Association Study (GWAS)–reported GC-susceptive genes (*26–30*), we found that *PLCE1* variants were mildly enriched among Japanese GC patients compared to the general Japanese population (table S5; 2.0-fold, *P* = 0.039, Fisher's exact test).

In our detailed analysis of the BRCA pathway genes, we found 20 (in 22 of 243 cases, 9.1%) germline variants among cases with *BRCA*-related mutational signature (Signature 3) [Figs. 2A and 4A

(blue bars) and table S6]. This frequency was not statistically different from that in the TCGA non-Asian GC population (19 of 212 cases, 9.0%; *P* = 1.0000, Fisher's exact test). According to reported studies (*31*, *32*), it has also been concluded that frequencies of BRCA1/2 pathogenic variants among general Japanese and West European populations are almost comparable. Several cancers with the BRCA signature are reported to be good responders to platinum chemotherapeutics and poly(adenosine diphosphate–ribose) polymerase inhibitors (*33*). The Signature 3 cluster (blue bars in Figs. 1A and 2A) consists of a roughly equal fraction of GCs in both Japanese and TCGA non-Asians (25 of 243, 10.3% and 19 of 212, 9.4%, respectively; *P* = 0.8735, Fisher's exact test), which should be appropriate targets of such therapeutics.

As for the *CDH1* germline variants, after filtering out common variations (see Materials and Methods), we summarized all nonsilent germline variants in our 243-case Japanese GC cohort (Fig. 4B, Table 2, and table S7), discovering 18 variants in total. This observed frequency (18 of 243, 7.4%) was higher than the reported consensus in the Japanese population (*34–36*). In contrast to the variants found in TCGA non-Asians (Fig. 4B, green colored circle), most (14 of 18 cases) of the variants in the Japanese population were

Table 1. Summary of germline variants that are presumably pathogenic. Genes with presumably pathogenic germline variants are shown in the order of their densities of nonsilent variants (left, only genes with 5% or higher variant frequency among our cohort are shown) and numbers of truncation variants (right). AA, number of amino acid residues.

Nonsilent				Truncation			
Gene	**N**	**Amino acid (AA) size**	**Density (N/AA)**	**Gene**	**N**	**Gene**	**N**
CDH1	16	882	0.018	ZFHX3	11	POLD1	1
PRDM1	13	789	0.016	GJB2	8	POLE	1
ARID1B	27	2231	0.012	CRIPAK	7	POLQ	1
BLM	16	1417	0.011	SLC25A13	7	**RAD51D**	1
ABCC4	13	1325	0.010	**BRCA2**	2	RUNX1T1	1
PIK3C2G	14	1445	0.010	PIK3C2G	2	CYP2D6	1
RICTOR	16	1708	0.009	ATM	1	ERBB2	1
SETD2	18	2061	0.009	B4GALT3	1	ERCC4	1
EPPK1	18	2420	0.007	**BRIP1**	1	EXO1	1
ROS1	17	2347	0.007	CUX1	1	FANCA	1
SLX4	13	1834	0.007	FGFR3	1	SBDS	1
ZFHX3	26	3703	0.007	GSK3B	1	SF1	1
DOCK8	13	2031	0.006	HLA-G	1	SMARCB1	1
POLE	14	2286	0.006	ITGAV	1	TLR4	1
FAT1	25	4588	0.005	**PALB2**	1		
MGA	16	3114	0.005				
COL7A1	15	2944	0.005				
APC	14	2843	0.005				
TRRAP	14	3830	0.004				

linked to diffuse-type histology (Fig. 4B, black outlined red circle), which provides a strong biological support to conclude their pathogenic roles ($P = 0.0060$, Fisher's exact test) (37). The variant frequency of CDH1 in the Japanese DGC patients (14 of 105, 13.3%) was approximately 4.1-fold higher than that in the TCGA non-Asian GC populations (2 of 62, 3.2%; $P = 0.0535$, Fisher's exact test; Table 2). It is intriguing that distributions of the CDH1 germline variants in the Japanese population were distinct from those of the TCGA non-Asians and specifically fell into five SNVs (p.G62V, p.T340A, p.L630V, p.V832M, and p.E880K). Among these five SNVs, four were predicted to be damaging in silico, four were reported in cases that met the HDGC criteria [strong familial aggregation (34, 35, 38) or extremely early onset (6)], and at least two were molecularly confirmed to be causative of malignancies due to the deficits in cell-cell adhesion and/or in the inhibition of invasion (Table 2) (39, 40). When germline variants found in a large Korean sporadic DGC dataset (6) were combined, it was revealed that these five recurrent SNVs were shared by both Japanese and Korean populations at measurable frequencies (Fig. 4B). This trend shows a clear contrast to the germline BRCA pathway variants (Fig. 4A), where no shared variants were found between Japanese and Korean GCs. The overall variant frequency of the 11 SNV loci found in this study was significantly higher among Japanese DGC patients than the general Japanese populations (4.0-fold; $P < 0.0001$, Fisher's exact test) and also higher among the general East Asian population than the general European population (6.0-fold; $P = 0.00124$,

Fisher's exact test) (Table 2), supporting the pathogenic characteristics of these variants and confirming their enrichments in East Asian populations.

Familial history of germline CDH1 variant carriers
Out of 18 Japanese GC patients with germline CDH1 variants, 11 had mild familial cancer histories (table S5), including one case with lobular carcinoma of the breast, a well-known trace of germline CDH1 variants. However, although the information available was limited, only one case (DGC at 39 years old) met the clinical criteria of the HDGC (4). On the other hand, seven individuals had no familial histories of cancers and were clinically considered sporadic cases. Therefore, it should be noted that substantial portions of GCs that are clinically assumed as sporadic incidences could be attributable to germline variations of the patients. In view of molecular carcinogenesis pathways, it is hypothesized that the clinical phenotypes of CDH1 germline variations are significantly modified by other environmental factors such as H. pylori infection and food contents. We investigated H. pylori copy numbers at noncancerous mucosa of our GC patient cohort, finding that GC cases with germline variants of CDH1 exhibited statistically higher H. pylori copy numbers compared to others (fig. S5). This finding is consistent with a previous report showing that ongoing infections of H. pylori are more frequent in DGCs (41), since almost all the CDH1 germline variant patients harbor DGCs (table S7). H. pylori is known to disrupt proper function
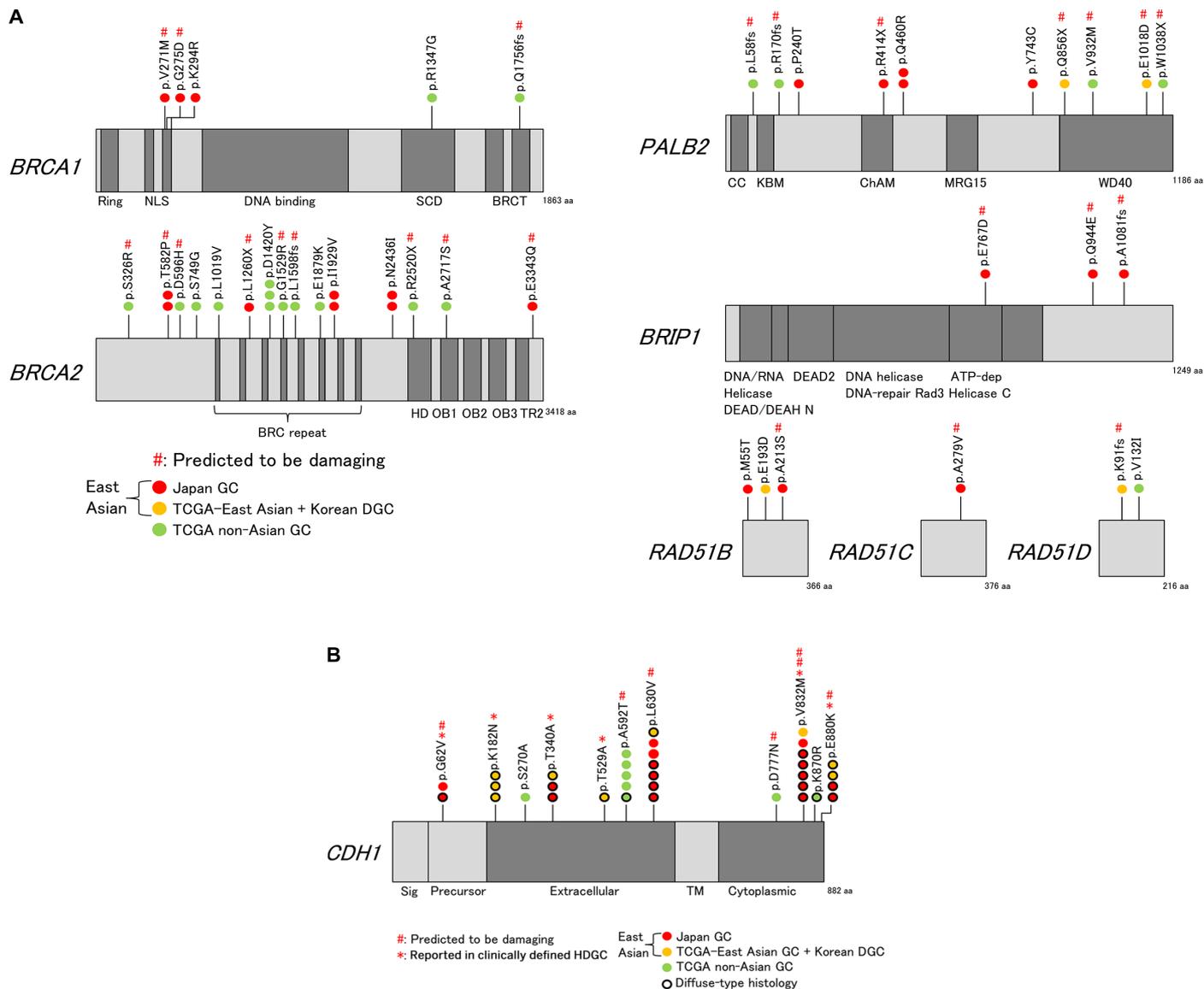
Fig. 4. The landscape of germline variations in East Asian and non-Asian GCs. (**A**) Distributions of germline variations of *BRCA* pathway genes found in combined datasets [Japanese GCs, TCGA GCs (*9*), and Korean DGCs (*6*)]. Only variants found in cases linked to the *BRCA* signature are shown. Red, Japanese; green, TCGA non-Asian. None of the patients were TCGA East Asians/Koreans (orange) who met this criterion. #, predicted to be damaging in silico; NLS, nuclear localization signal; SCD, serine cluster domain; HD, helical domain; OB, oligonucleotide binding; TR2, second RAD51-binding domain; aa, amino acid; ATP, adenosine triphosphate. (**B**) Distributions of germline variations of *CDH1* in the same datasets as in (A). *, reported in clinically defined HDGC (strong familial aggregation or extremely early-onset case). Bold circle, diffuse-type histology. SIG, signal; TM, transmembrane domain.

of CDH1 (*42*), which might suggest that *H. pylori* infection works as an additional second hit for the *CDH1* germline variant population in the GC development. This observation and our speculations, however, are not scientifically conclusive because the copy number of *H. pylori* is influenced by the existence of mucosal metaplasia and the longevity of chronic infection (*43*). Even if the familial histories or clinical phenotypes did not straightforwardly resemble the definitions of familial GCs clinically, at least a portion of GC cases can be attributable to *CDH1* germline variants. The p.V832M (*34*) and p.G62V (*35*) variants, for instance, were reported in families with strong familial aggregations of GCs (7 of 12 and 7 of 13 family members examined,

respectively); however, only four of eight cases with these variants in our Japanese GCs showed, at most, slight familial histories, and none of them fulfilled the clinical criteria of HDGC, although the family information was available only for the first-degree relatives.

DISCUSSION
Large-scale whole-exome sequencing for East Asian GCs in combination with the public dataset has clarified important lifestyle and germline predispositions of GC development among high-incident East Asian populations. Alcohol intake behavior with genetically

Table 2. Summary of germline CDH1 variants in East Asia (Japan and Korea) and TCGA datasets. Germline CDH1 variants in East Asia (Japan and Korea) and TCGA datasets are shown. ToMMo, Tohoku Medical Megabank Organization.

	SNVs	Polyphen2 (HumDiv)	Reported in HDGC*	Reported in Korea†	Japanese GC	TCGA non-Asian GC	Japanese DGC	TCGA non-Asian DGC	Japanese DGC versus non-Asian DGC	Japan ToMMo	Japanese DGC versus Japan ToMMo	1000 Genomes EAS	1000 Genomes EUR	1000 Genomes EAS versus EUR	Note
Case					243	212	105	62		1070		504	503		
1	p.G62V	Probably damaging	Yes		2		1			3					(35)
2	p.K182N	Benign	Yes	Yes						1		1			(6)
3	p.S270A	Benign				1									
4	p.T340A	Benign	Yes	Yes	2		2			6		1			(6, 38, 40)
5	p.T529A	Benign	Yes	Yes											(6)
6	p.A592T	Possibly damaging				5		1						2	(48)
7	p.L630V	Probably damaging	Yes		6		4			9		6			(6)
8	p.D777N	Probably damaging				1									
9	p.V832M	Probably damaging	Yes	Yes	6		5			10		2			(34, 40, 49)
10	p.K870R	Benign				1		1							
11	p.E880K	Possibly damaging	Yes	Yes	2		2			7		2			(6)
Sum					7.4% (18)	3.8% (8)	13.3% (14)	3.2% (2)	4.1-fold ($P = 0.0535$)	3.4% (36)	4.0-fold ($P < 0.0001$)	2.4% (12)	0.4% (2)	6.0-fold ($P = 0.0124$)	

*Reported in cases that met the clinical criteria of HDGC.   †Reported in TCGA Korean and in (6).

inactive *ALDH2* background predisposes individuals to GCs with Signature 16. Besides, synergetic effects of alcohol intake and smoking habit were observed among the *ALDH2*-inactive individuals. Alcohol intake and the incidence of various types of cancers have extensively been studied so far (14–16, 21). Although conflicting reports have indicated the correlations between alcohol intake and GC risk based on indirect epidemiological data (44, 45), no reports have shown direct biological evidence of the effects of alcohol on GC carcinogenesis. Global genetic classifications combined with clear quantification of alcohol contributions defined by mutational signatures, as taken by this study, could succeed in unraveling the ethnicity-specific intercorrelations of etiologies in GC development in the high-incidence areas. Alcohol-related GC is a biologically distinct subgroup with low mutation burdens and characteristic immunological profiles of increased B cell infiltration due to cancer cell–derived CXCL13 chemokine. While there are several reports of alcohol-induced dysregulation of adaptive immunity including B cells (46, 47), its biological mechanism and its relation to alcohol exposure are unclear and remain to be investigated further. Our results show that Signature 16 GCs have distinct immunological features, which attracts further attention in relation to the responsiveness of these GCs to current immune therapies as well as their applicability for alternative immunotherapies.

In addition to the *BRCA* pathways, this study also clarified that germline variations of *CDH1* exist at a higher-than-expected frequency among East Asian GCs. The shared distributions of the recurrent *CDH1* variants between Japanese and Koreans strongly suggest the existence of common ancestral events and widespread distributions of these *CDH1* pathogenic variants in the East Asian areas. Thus, these variants affect the prevalence of GCs in these areas, although their precise geographic distributions are yet to be determined. It is important that many of these cases are clinically recognized as sporadic cases and have been overlooked so far without special attention to the preventive cares for family members, such as periodic endoscopies and/or prophylactic surgeries (4). The incomplete penetrance and environmental dependency of *CDH1* variants are likely to have led to the controversy about their pathogenicity so far. Several groups have tried to experimentally show the evidence of pathogenicity of the variants, but their results were often conflicting (6, 40, 48, 49), partially because of the differences in the assays used and the difficulties in proving the biological function of the variants with mild effects. Our large-scale analysis of genomics and clinical data in this study clearly showed multiple evidences of the pathogenic nature of these *CDH1* variants.

Approximately one-fifth (22 of 105, 21.0%) of the DGCs in our study were attributable either to the alcohol consumption coupled with defective *ALDH2* allele or to the germline loss-of-function variants of *CDH1*. Our findings clarified previously unrecognized impacts of the defined germline and lifestyle factors on the high incidences of GCs in East Asian areas and provided us strong motivations for clinical intervention in lifestyles and familial cares coupled with precision germline genotyping in these areas.

MATERIALS AND METHODS
Informed consent and sample preparation
Fresh-frozen GC and paired normal stomach tissues were obtained from essentially consecutive patients who underwent gastrectomy

at the University of Tokyo Hospital (n = 169) and Yokohama City University Hospital (n = 126). Informed consent was obtained from each individual, and this study was approved by the institutional review boards at the University of Tokyo, Yokohama City University, and Tokyo Medical and Dental University. QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) and RNeasy Mini Kit (Qiagen) were used to extract 295 paired DNA and RNA samples from these tissues, respectively, as per the manufacturer's instructions.

Clinicopathological information

Patient clinicopathological information [gender, age at diagnosis, alcoholic drinking habits, smoking habits, Lauren classification, tumor site, stage (UICC TNM classification), and histopathological features] were collected with informed consents. Patients were classified as "consuming alcohol" if they reported socially associated or more severe alcohol usage or otherwise as "not consuming alcohol." Patients who smoked at least one cigarette/day on average for at least 1 year during their lifetime were defined as "smokers" or otherwise defined as "nonsmokers." Tumor TNM staging including an assessment of the tumor size and lymph node and distant metastases were reviewed by at least two pathologists and defined according to the American Joint Committee on Cancer (eighth edition). Detailed data for the 243 patients are shown in table S9. Mix-type GCs (i.e., coexistence of diffuse and intestinal components) were regarded as diffuse-type GCs for this study, as is suggested by guidelines (50).

H. pylori infection status

H. pylori infection status of the patients at the time of surgeries was determined by quantitative polymerase chain reaction (qPCR), using the extracted DNA from noncancerous gastric tissues of patients with GC. The ratio of qPCR quantities of *H. pylori* per human genomic DNA was considered the *H. pylori* copy number in the subjected specimens. Human genomic DNA was quantified by the TaqMan method targeting ribonuclease P gene locus (Thermo Fisher Scientific, MA, USA), and the amount of *H. pylori* was measured by SYBR Green method targeting *ureC* gene, which was reported to be the most sensitive and specific marker for *H. pylori* (51). The *ureC* gene primers were 5′-GCATGCAATTGAATA-AAGCC-3′ (forward) and 5′-GCCGCTATAACGGATCAAAT-3′ (reverse) as reported (51). TaqMan PCR was performed as per the manufacturer's protocol, and SYBR Green qPCR was performed 50 cycles for denature, 95°C for 15 s, and extension, 60°C for 1 min. In total, we obtained informative data of *H. pylori* copy numbers for 170 of 243 Japanese GC patients. Because of the substantial paucities of genomic DNA, it was not possible to perform qPCRs for the remaining 73 cases.

Exome capture, library construction, and whole-exome sequencing

Whole-exome sequencing was performed for 295 GCs and paired normal gastric tissues, as previously described (8). Each DNA sample (1.1 mg) was sheared using a Covaris SS Ultrasonicator (Covaris, MA, USA) according to the manufacturer's instructions. A Sciclone next-generation sequencing workstation (Caliper Life Sciences, MA, USA) was used to automatically construct the DNA sequence libraries, according to the manufacturer's instruction. Exome capture was performed using the Agilent SureSelect Human All Exon Kit v4 and v5+ LincRNA (Agilent Technologies, CA,

USA). Each sample was sequenced using an Illumina HiSeq 2000 platform (Illumina, CA, USA) and the provided protocol for producing 2 × 100–base pair (bp) pair-end reads. Image analyses and base calling were performed using an Illumina pipeline with default settings. The median depth of tumor/normal coding sequence is 170/84 in Japan and 165/91 in TCGA.

Mutation calling

We obtained FASTQ files of the TCGA dataset [published in 2014 (9)] from the TCGA website (https://tcga-data.nci.nih.gov). Then, we performed mutation calling of our dataset and TCGA dataset using exactly the same pipeline from the beginning of sequence mapping procedure as follows. Paired-end reads were aligned to the human reference genome (GRCh37) using NovoAlign (http://novocraft.com/products/novoalign/) for both tumor and nontumor samples. Probable PCR duplications, for which paired-end reads aligned to the same genomic position, were removed, and pileup files were generated using SAMtools (52) and some in-house programs. To find somatic point mutations (SNVs and short indels), the following cutoff values were used for base selection: (i) a mapping quality score of at least 20 and (ii) a base quality score of at least 10. Somatic mutations were selected using the following filtering conditions: (iii) The numbers of reads supporting a mutation for tumor were at least 4 and 8, with at least one base with a quality score greater than 30, when tumor variant allele frequency (TVAF) $\geq$ 0.15 and 0.15 > TVAF $\geq$ 0.05, respectively; (iv) variant allele frequency (VAF) of matched nontumor sample was less than 0.03 with a read depth of at least 8. Since sequence errors are considered to occur in a sequence-specific manner, the sequence read information of all nontumor samples was combined together to accurately discriminate true positives from false positives. Then, the following filter was applied: (v) VAF of grouped nontumor sample (NVAF) was less than 0.03 and 0.01, when TVAF $\geq$ 0.15 and 0.15 > TVAF $\geq$ 0.05, respectively; (vi) the ratio (TVAF/NVAF) of VAF of tumor (TVAF) and grouped nontumor samples (NVAF) must be more than 20. After mutation calling, (vii) mutations with a strand bias (between forward and reverse reads) greater than 95% were removed.

TCGA dataset

TCGA GC dataset [published in 2014 (9)] was obtained from the TCGA website (https://tcga-data.nci.nih.gov). It includes 288 (212 non-Asians and 76 Asians) GC cases with information regarding their somatic mutations, list of reads per kilobase per million (RPKM) mapped reads (for the 262 GCs and 29 normal stomach tissues), and clinical data. No ethnicity information was provided for 41 GC cases (obtained from 38 German and 3 Canadian patients); therefore, we considered these to be non-Asian cases. BAM files were obtained from the TCGA website archives and used to analyze germline variations for the TCGA dataset.

Classifications of Japanese GC cases by the TCGA subtypes

We classified our Japanese GC cases into four subgroups: EBV-associated GC, MSI-GC, GS-GC, and chromosomal instability (CIN)–GC (9). First, we defined EBV-GC by detecting viral sequences in the RNA-seq data using PathSeq (53). If the sequence reads unique to the EBV were more than 1×10^{-5} of the reads unique to human genome per individual, these GCs were classified as EBV-GC. This cutoff was determined in accordance with the

data of Epstein-Barr virus–encoded small RNA (EBER) in situ hybridization. Second, we defined MSI-GC by mutational signature analysis. Signatures 6 and 15 are well-known resultants of the MMR deficiencies; therefore, we identified MSI-GC as those in Signature 6 or 15 clusters. Last, we divided the rest of GCs into GS- or CIN-GCs. These two subtypes were distinguished by the presence or absence of extensive somatic copy number aberrations based on the hierarchical clustering. To analyze copy number aberrations, the read depth was compared between normal and tumor for each capture target region. The tumor/normal depth ratios were calculated, and the values were smoothed using a moving average. Using these values, clustering was done in R based on Euclidean distance using Ward's method as was performed in the TCGA (9).

Mutational signature analysis

We obtained somatic SNVs for 531 GC cases (288 TCGA cases and 243 Japanese cases). We used the pan-cancer catalog of 30 Signatures referenced in the COSMIC database (11) (http://cancer.sanger.ac.uk/cosmic), and the contribution of each mutational Signature was calculated by deconstructSigs software (table S10) (54). Unsupervised hierarchical clustering was performed on the Signature contribution scores for each GC case, by using pheatmap v1.0.2 software in R to define the GC subgroups (Fig. 1A). The ward.D was used as the clustering method, and Euclidean distances (Japan and TCGA datasets; Fig. 1A) and correlation distances (Japanese GCs; Fig. 2A), respectively, were used to assemble the GC clusters. The incidence of Signature 4– or Signature 16–associated SNVs/Mb was calculated on the basis of the contribution scores of Signature 16 in each case, as defined above. We excluded hypermutator GCs from the analyses of the frequencies of SNVs between subgroups, since large numbers of SNVs in hypermutators could extremely affect the statistical analyses. We defined the hypermutator GCs as those with mutation burdens over 10/Mb (55).

RNA sequencing

For 184 Japanese GC cases, total RNA was extracted from frozen sections (10 sections × 10 μm thickness) of GC specimens by RNeasy Mini Kit (Qiagen). Quantification and qualification of extracted RNAs were determined by Agilent Bioanalyzer (Agilent Technologies, CA, USA). RNA-seq libraries were constructed by the TruSeq Stranded mRNA Sample Prep Kit (Illumina, CA, USA) according to the manufacturer's protocol. Each RNA-seq library was sequenced on an Illumina HiSeq platform (Illumina) by 2 × 100-bp paired-end sequencing as per the manufacturer's protocol. FASTQ files were processed using Kallisto version 0.43.1 with Ensemble transcriptome (release 79) to create gene expression profiles.

Estimation of LSTs

We estimated the numbers of LSTs by the following three steps: (i) copy number ratio calculation, (ii) detection of large copy number changes, and (iii) LST detection. First, initial copy number ratios were obtained by comparing read depth information for tumor and normal samples, and the copy number estimates were adjusted by tumor purity as described previously (8). Next, copy number ratios within 100-kbp window were averaged and further smoothed by moving average filters with a window of five points. Then, circular binary segmentation implemented in python (https://github.com/jeremy9959/cbs) was applied with 10,000 random permutations. To detect only highly confident copy number variation (CNV) regions, the cutoff value was set to $P = 0$. The detected segments were merged when the difference of mean log ratios between the regions ≤ 0.3. Last, LSTs were detected on the basis of the definition of LST as previously described (13). After filtering and smoothing of all variations less than 3 Mb, a chromosomal break between adjacent regions of at least 10 Mb was defined as LST. The number of LSTs in the tumor genome was estimated for each chromosome arm independently.

Detection of differentially expressed genes among Signature 16 GCs

To detect any differentially expressed gene sets among Signature 16 GCs, we compared the fragments per kilobase of exon per million reads mapped (FPKM) between Signature 16 GCs ($n = 12$) and the other GCs ($n = 172$). Initially, we calculated q values, but there were no genes with statistical significances, probably because of the few numbers of Signature 16 GCs; therefore, we used P values to define differentially expressed genes. We selected genes with P values of <0.05 and, at the same time, the average FPKM of the genes among Signature 16 GCs > the other GCs. Last, we identified 1303 genes. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed using The Database for Annotation, Visualization and Integrated Discovery (DAVID) platform.

CIBERSORT

CIBERSORT (23) was applied to 184 Japanese GC tissues to enumerate the proportions of 22 types of immune cells [B cells naïve, B cells memory, plasma cells, T cells CD8, T cells CD4 naïve, T cells CD4 memory resting, T cells CD4 memory activated, T cells follicular helper, T cells gamma delta, T cells regulatory, natural killer (NK) cells resting, NK cells activated, monocytes, macrophages M0, macrophages M1, macrophages M2, dendritic cells resting, dendritic cells activated, mast cells resting, mast cells activated, eosinophils, and neutrophils] using the LM22 dataset. CIBERSORT was run in "absolute mode," and "disable quantile normalization" option was selected as recommended for RNA-seq data.

Immunohistochemistry

Immunohistochemistry staining was performed onto formalin-fixed and paraffin-embedded (FFPE) specimens of GC cases. FFPE tissues were deparaffinized by xylene (Wako Pure Chemical Industries, Japan), and antigens were retrieved by autoclave treatments in citrate buffer (pH 6.0) (Abcam, Cambridge, UK). Endogenous peroxidases were removed by 3% H_2O_2 (Sigma-Aldrich, MO, USA). Then, 2% bovine serum albumin (Sigma-Aldrich) in phosphate-buffered saline was used for a blocking solution. Antibodies α-CD20 #ab78237 (Abcam) and α-CXCL13 #20927-1-AP (Proteintech Group, IL, USA) were used as primary antibodies. Histostar (MBL, Japan) and Histostar diaminobenzidine (DAB) substrate solution (MBL) were used for detecting signals.

Signature 16 transcriptional strand bias

Transcriptional strand biases of the Signature 16 were determined as follows. All SNVs called by Karkinos were divided into transcribed or untranscribed SNVs. The contribution of each mutational Signature was calculated by deconstructSigs software (54) on transcribed or untranscribed SNVs.

Germline variant analysis

We analyzed germline variants by using GATK software (https://software.broadinstitute.org/gatk/) and identified 85,006 germline

nonsilent (missense and nonframeshift indels) and 7000 germline truncation variants (nonsense, nonstop, and frameshift indels) using the following parameters: sequence depth $\geqq 30$ and <1% of population-level MAF, according to the Tohoku Medical Megabank Organization (ToMMo) (56), 1000 Genomes (57), and Exome Aggregation Consortium (ExAC) (58) databases. We retained 66,438 nonsilent variants and 3952 truncation variants after the removal of variants with VAF <0.2 and/or GATK quality value score < 500. We further screened the 262 gastric tumors and 29 normal stomach tissues sourced from TCGA according to whether they achieved a C score (59) $\geqq 10$ and a RPKM value of >1 and resultantly generated a list of 34,622 nonsilent and 2424 truncation variants. Last, we focused on 624 cancer-associated genes selected using a previously described pan-cancer germline variant analysis method (25) and obtained 1777 nonsilent variants (in 390 genes) and 60 truncation variants (in 28 genes) (fig. S6). Table 1 shows genes with presumed pathogenic germline variants identified in the 243 Japanese GC cases. For the focused analysis of *CDH1*, we included the TCGA GC dataset whose germline variants were called using GATK by a filtering depth of $\geqq 30$, <1% of population-level MAF, VAF $\geqq 0.2$, and a quality score of $\geqq 500$ (as described above) (9) and the germline variants called in Korean DGC dataset (6). We identified 11 SNVs in total (Table 2).

The truncation germline variants identified in the 243 Japanese GC cases are also summarized in Table 1. According to a previous paper of comprehensive breast cancer study (60), *BRCA* pathway genes include *BRCA1*, *BRCA2*, *PALB2*, *RAD51*, *BARD1*, and *BRIP1*; thus, we selected GC cases that exhibited nonsilent and/or truncation variants in these *BRCA* genes and were associated with Signature 3 contributions (the *BRCA* signature) using a filtering depth of $\geqq 30$, <1% population-level MAF, and VAF $\geqq 0.2$, and a quality score of $\geqq 500$ (table S3) from the Japanese cohort and TCGA dataset (9).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/19/eaav9778/DC1

REFERENCES AND NOTES

1. J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
2. L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, Global cancer statistics, 2012. *CA Cancer J. Clin.* **65**, 87–108 (2015).
3. S. Tsugane, S. Sasazuki, Diet and the risk of gastric cancer: Review of epidemiological evidence. *Gastric Cancer* **10**, 75–83 (2007).
4. R. S. van der Post, I. P. Vogelaar, F. Carneiro, P. Guilford, D. Huntsman, N. Hoogerbrugge, C. Caldas, K. E. C. Schreiber, R. H. Hardwick, M. G. E. M. Ausems, L. Bardram, P. R. Benusiglio, T. M. Bisseling, V. Blair, E. Bleiker, A. Boussioutas, A. Cats, D. Coit, L. DeGregorio, J. Figueiredo, J. M. Ford, E. Heijkoop, R. Hermens, B. Humar, P. Kaurah, G. Keller, J. Lai, M. J. L. Ligtenberg, M. O'Donovan, C. Oliveira, H. Pinheiro, K. Ragunath, E. Rasenberg, S. Richardson, F. Roviello, H. Schackert, R. Seruca, A. Taylor, A. ter Huurne, M. Tischkowitz, S. T. A. Joe, B. van Dijck, N. C. T. van Grieken, R. van Hillegersberg, J. W. van Sandick, R. Vehof, J. H. van Krieken, R. C. Fitzgerald, Hereditary diffuse gastric cancer: Updated clinical guidelines with an emphasis on germline *CDH1* mutation carriers. *J. Med. Genet.* **52**, 361–374 (2015).
5. Y. Liu, N. S. Sethi, T. Hinoue, B. G. Schneider, A. D. Cherniack, F. Sanchez-Vega, J. A. Seoane, F. Farshidfar, R. Bowlby, M. Islam, J. Kim, W. Chatila, R. Akbani, R. S. Kanchi, C. S. Rabkin, J. E. Willis, K. K. Wang, S. J. McCall, L. Mishra, A. I. Ojesina, S. Bullman, C. S. Pedamallu, A. J. Lazar, R. Sakai; Cancer Genome Atlas Research Network, V. Thorsson, A. J. Bass, P. W. Laird, Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell* **33**, 721–735.e8 (2018).
6. S. Y. Cho, J. W. Park, Y. Liu, Y. S. Park, J. H. Kim, H. Yang, H. Um, W. R. Ko, B. I. Lee, S. Y. Kwon, S. W. Ryu, C. H. Kwon, D. Y. Park, J.-H. Lee, S. I. Lee, K. S. Song, H. Hur, S.-U. Han, H. Chang, S.-J. Kim, B.-S. Kim, J.-H. Yook, M.-W. Yoo, B.-S. Kim, I.-S. Lee, M.-C. Kook, N. Thiessen, A. He, C. Stewart, A. Dunford, J. Kim, J. Shih, G. Saksena, A. D. Cherniack, S. Schumacher, A.-T. Weiner, M. Rosenberg, G. Getz, E. G. Yang, M.-H. Ryu, A. J. Bass, H. K. Kim, Sporadic early-onset diffuse gastric cancers have high frequency of somatic *CDH1* alterations, but low frequency of somatic *RHOA* mutations compared with late-onset cancers. *Gastroenterology* **153**, 536–549.e26 (2017).
7. K. Wang, S. T. Yuen, J. Xu, S. P. Lee, H. H. N. Yan, S. T. Shi, H. C. Siu, S. Deng, K. M. Chu, S. Law, K. H. Chan, A. S. Y. Chan, W. Y. Tsui, S. L. Ho, A. K. W. Chan, J. L. K. Man, V. Foglizzo, M. K. Ng, A. S. Chan, Y. P. Ching, G. H. W. Cheng, T. Xie, J. Fernandez, V. S. W. Li, H. Clevers, P. A. Rejto, M. Mao, S. Y. Leung, Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* **46**, 573–582 (2014).
8. M. Kakiuchi, T. Nishizawa, H. Ueda, K. Gotoh, A. Tanaka, A. Hayashi, S. Yamamoto, K. Tatsuno, H. Katoh, Y. Watanabe, T. Ichimura, T. Ushiku, S. Funahashi, K. Tateishi, I. Wada, N. Shimizu, S. Nomura, K. Koike, Y. Seto, M. Fukayama, H. Aburatani, S. Ishikawa, Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. *Nat. Genet.* **46**, 583–587 (2014).
9. The Cancer Genome Atlas Research Network, Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
10. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
11. S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, P. A. Futreal, COSMIC: Mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
12. H. Li, S. Borinskaya, K. Yoshimura, N. Kal'ina, A. Marusin, V. A. Stepanov, Z. Qin, S. Khaliq, M.-Y. Lee, Y. Yang, A. Mohyuddin, D. Gurwitz, S. Q. Mehdi, E. Rogaev, L. Jin, N. K. Yankovsky, J. R. Kidd, K. K. Kidd, Refined geographic distribution of the oriental *ALDH2*504Lys* (nee 487Lys) variant. *Ann. Hum. Genet.* **73**, 335–345 (2009).
13. T. Popova, E. Manié, G. Rieunier, V. Caux-Moncoutier, C. Tirapo, T. Dubois, O. Delattre, B. Sigal-Zafrani, M. Bollet, M. Longy, C. Houdayer, X. Sastre-Garau, A. Vincent-Salomon, D. Stoppa-Lyonnet, M.-H. Stern, Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
14. J. I. Garaycoechea, G. P. Crossan, F. Langevin, L. Mulderrig, S. Louzada, F. Yang, G. Guilbaud, N. Park, S. Roerink, S. Nik-Zainal, M. R. Stratton, K. J. Patel, Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* **553**, 171–177 (2018).
15. E. Letouzé, J. Shinde, V. Renault, G. Couchy, J.-F. Blanc, E. Tubacher, Q. Bayard, D. Bacq, V. Meyer, J. Semhoun, P. Bioulac-Sage, S. Prévôt, D. Azoulay, V. Paradis, S. Imbeaud, J.-F. Deleuze, J. Zucman-Rossi, Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315 (2017).
16. J. Chang, W. Tan, Z. Ling, R. Xi, M. Shao, M. Chen, Y. Luo, Y. Zhao, Y. Liu, X. Huang, Y. Xia, J. Hu, J. S. Parker, D. Marron, Q. Cui, L. Peng, J. Chu, H. Li, Z. Du, Y. Han, W. Tan, Z. Liu, Q. Zhan, Y. Li, W. Mao, C. Wu, D. Lin, Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nat. Commun.* **8**, 15290 (2017).
17. M. Wang, E. J. McIntee, G. Cheng, Y. Shi, P. W. Villalta, S. S. Hecht, Identification of DNA adducts of acetaldehyde. *Chem. Res. Toxicol.* **13**, 1149–1157 (2000).
18. S. J. Yin, C. S. Liao, C. W. Wu, T. T. Li, L. L. Chen, C. L. Lai, T. Y. Tsao, Human stomach alcohol and aldehyde dehydrogenases: Comparison of expression pattern and activities in alimentary tract. *Gastroenterology* **112**, 766–775 (1997).
19. R. Maejima, K. Iijima, P. Kaihovaara, W. Hatta, T. Koike, A. Imatani, T. Shimosegawa, M. Salaspuro, Effects of ALDH2 genotype, PPI treatment and L-cysteine on carcinogenic acetaldehyde in gastric juice and saliva after intragastric alcohol administration. *PLOS ONE* **10**, e0120397 (2015).

20. M. Fousteri, L. H. Mullenders, Transcription-coupled nucleotide excision repair in mammalian cells: Molecular mechanisms and biological effects. *Cell Res.* **18**, 73–84 (2008).

21. S. Ishiguro, S. Sasazuki, M. Inoue, N. Kurahashi, M. Iwasaki, S. Tsugane; JPHC Study Group, Effect of alcohol consumption, cigarette smoking and flushing response on esophageal cancer risk: A population-based cohort study (JPHC study). *Cancer Lett.* **275**, 240–246 (2009).

22. H. Katoh, D. Komura, M. Konishi, R. Suzuki, A. Yamamoto, M. Kakiuchi, R. Sato, T. Ushiku, S. Yamamoto, K. Tatsuno, T. Oshima, S. Nomura, Y. Seto, M. Fukayama, H. Aburatani, S. Ishikawa, Immunogenetic profiling for gastric cancers identifies sulfated glycosaminoglycans as major and functional B cell antigens in human malignancies. *Cell Rep.* **20**, 1073–1087 (2017).

23. A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, A. A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

24. R. Sahasrabudhe, P. Lott, M. Bohorquez, T. Toal, A. P. Estrada, J. J. Suarez, A. Brea-Fernández, J. Cameselle-Teijeiro, C. Pinto, I. Ramos, A. Mantilla, R. Prieto, A. Corvalan, E. Norero, C. Alvarez, T. Tapia, P. Carvallo, L. M. Gonzalez, A. Cock-Rada, A. Solano, F. Neffa, A. Della Valle, C. Yau, G. Soares, A. Borowsky, N. Hu, L.-J. He, X.-Y. Han; Latin American Gastric Cancer Genetics Collaborative Group, P. R. Taylor, A. M. Goldstein, J. Torres, M. Echeverry, C. Ruiz-Ponte, M. R. Teixeira, L. G. Carvajal-Carmona, M. Echeverry, M. Bohorquez, R. Prieto, J. Suarez, G. Mateus, M. M. Bravo, F. Bolaños, A. Vélez, A. Corvalan, P. Carvallo, J. Torres, L. Carvajal-Carmona, Germline mutations in PALB2, BRCA1, and RAD51C, which regulate DNA recombination repair, in patients with gastric cancer. *Gastroenterology* **152**, 983–986.e6 (2017).

25. C. Lu, M. Xie, M. C. Wendl, J. Wang, M. D. McLellan, M. D. M. Leiserson, K.-l. Huang, M. A. Wyczalkowski, R. Jayasinghe, T. Banerjee, J. Ning, P. Tripathi, Q. Zhang, B. Niu, K. Ye, H. K. Schmidt, R. S. Fulton, J. F. McMichael, P. Batra, C. Kandoth, M. Bharadwaj, D. C. Koboldt, C. A. Miller, K. L. Kanchi, J. M. Eldred, D. E. Larson, J. S. Welch, M. You, B. A. Ozenberger, R. Govindan, M. J. Walter, M. J. Ellis, E. R. Mardis, T. A. Graubert, J. F. Dipersio, T. J. Ley, R. K. Wilson, P. J. Goodfellow, B. J. Raphael, F. Chen, K. J. Johnson, J. D. Parvin, L. Ding, Patterns and functional implications of rare germline variants across 12 cancer types. *Nat. Commun.* **6**, 10086 (2015).

26. N. Hu, Z. Wang, X. Song, L. Wei, B. S. Kim, N. D. Freedman, J. Baek, L. Burdette, J. Chang, C. Chung, S. M. Dawsey, T. Ding, Y.-T. Gao, C. Giffen, Y. Han, M. Hong, J. Huang, H. S. Kim, W.-P. Koh, L. M. Liao, Y. M. Mao, Y.-L. Qiao, X.-O. Shu, W. Tan, C. Wang, C. Wu, M.-J. Wu, Y.-B. Xiang, M. Yeager, J. H. Yook, J.-M. Yuan, P. Zhang, X.-K. Zhao, W. Zheng, K. Song, L.-D. Wang, D. Lin, S. J. Chanock, A. M. Goldstein, P. R. Taylor, C. C. Abnet, Genome-wide association study of gastric adenocarcinoma in Asia: a comparison of associations between cardia and non-cardia tumours. *Gut* **65**, 1611–1618 (2016).

27. C. C. Abnet, N. D. Freedman, N. Hu, Z. Wang, K. Yu, X.-O. Shu, J.-M. Yuan, W. Zheng, S. M. Dawsey, L. M. Dong, M. P. Lee, T. Ding, Y.-L. Qiao, Y.-T. Gao, W.-P. Koh, Y.-B. Xiang, Z.-Z. Tang, J.-H. Fan, C. Wang, W. Wheeler, M. H. Gail, M. Yeager, J. Yuenger, A. Hutchinson, K. B. Jacobs, C. A. Giffen, L. Burdett, J. F. Fraumeni, M. A. Tucker, W.-H. Chow, A. M. Goldstein, S. J. Chanock, P. R. Taylor, A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.* **42**, 764–767 (2010).

28. Y. Shi, Z. Hu, C. Wu, J. Dai, H. Li, J. Dong, M. Wang, X. Miao, Y. Zhou, F. Lu, H. Zhang, L. Hu, Y. Jiang, Z. Li, M. Chu, H. Ma, J. Chen, G. Jin, W. Tan, T. Wu, Z. Zhang, D. Lin, H. Shen, A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. *Nat. Genet.* **43**, 1215–1218 (2011).

29. Z. Wang, J. Dai, N. Hu, X. Miao, C. C. Abnet, M. Yang, N. D. Freedman, J. Chen, L. Burdette, X. Zhu, C. C. Chung, C. Ren, S. M. Dawsey, M. Wang, T. Ding, J. Du, Y.-T. Gao, R. Zhong, C. Giffen, W. Pan, W.-P. Koh, N. Dai, L. M. Liao, C. Yan, Y.-L. Qiao, Y. Jiang, X.-O. Shu, J. Chen, C. Wang, H. Ma, H. Su, Z. Zhang, L. Wang, C. Wu, Y.-B. Xiang, Z. Hu, J.-M. Yuan, L. Xie, W. Zheng, D. Lin, S. J. Chanock, Y. Shi, A. M. Goldstein, G. Jin, P. R. Taylor, H. Shen, Identification of new susceptibility loci for gastric non-cardia adenocarcinoma: pooled results from two Chinese genome-wide association studies. *Gut* **66**, 581–587 (2017).

30. Study Group of Millennium Genome Project for Cancer, Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat. Genet.* **40**, 730–740 (2008).

31. M. J. Hall, J. E. Reid, L. A. Burbidge, D. Pruss, A. M. Deffenbaugh, C. Frye, R. J. Wenstrup, B. E. Ward, T. A. Scholl, W. W. Noll, BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer. *Cancer* **115**, 2222–2233 (2009).

32. Y. Momozawa, Y. Iwasaki, M. T. Parsons, Y. Kamatani, A. Takahashi, C. Tamura, T. Katagiri, T. Yoshida, S. Nakamura, K. Sugano, Y. Miki, M. Hirata, K. Matsuda, A. B. Spurdle, M. Kubo, Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls. *Nat. Commun.* **9**, 4083 (2018).

33. H. Farmer, N. McCabe, C. J. Lord, A. N. J. Tutt, D. A. Johnson, T. B. Richardson, M. Santarosa, K. J. Dillon, I. Hickson, C. Knights, N. M. B. Martin, S. P. Jackson, G. C. M. Smith, A. Ashworth, Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).

34. T. Yabuta, K. Shinmura, M. Tani, S. Yamaguchi, K. Yoshimura, H. Katai, T. Nakajima, E. Mochiki, T. Tsujinaka, M. Takami, K. Hirose, A. Yamaguchi, S. Takenoshita, J. Yokota,

35. K. Shinmura, T. Kohno, H. Takahashi, A. Sasaki, A. Ochiai, P. Guilford, A. Hunter, A. E. Reeve, H. Sugimura, N. Yamaguchi, J. Yokota, Familial gastric cancer: Clinicopathological characteristics, RER phenotype and germline p53 and *E-cadherin* mutations. *Carcinogenesis* **20**, 1127–1131 (1999).

36. S. Iida, Y. Akiyama, W. Ichikawa, T. Yamashita, T. Nomizu, Z. Nihei, K. Sugihara, Y. Yuasa, Infrequent germ-line mutation of the E-cadherin gene in Japanese familial gastric cancer kindreds. *Clin. Cancer Res.* **5**, 1445–1447 (1999).

37. M. Takeichi, Cadherin cell adhesion receptors as a morphogenetic regulator. *Science* **251**, 1451–1455 (1991).

38. C. A. Moreira-Nunes, M. B. L. Barros, B. do Nascimento Borges, R. C. Montenegro, L. M. Lamarão, H. F. Ribeiro, A. B. Bona, P. P. Assumpção, J. A. Rey, G. R. Pinto, R. R. Burbano, Genetic screening analysis of patients with hereditary diffuse gastric cancer from north and northeastern Brazil. *Hered. Cancer Clin. Pract.* **12**, 18 (2014).

39. A. R. Mateus, J. Simões-Correia, J. Figueiredo, S. Heindl, C. C. Alves, G. Suriano, B. Luber, R. Seruca, E-cadherin mutations and cell motility: A genotype–phenotype correlation. *Exp. Cell Res.* **315**, 1393–1402 (2009).

40. G. Suriano, D. Mulholland, O. de Wever, P. Ferreira, A. R. Mateus, E. Bruyneel, C. C. Nelson, M. M. Mareel, J. Yokota, D. Huntsman, R. Seruca, The intracellular E-cadherin germline mutation V832 M lacks the ability to mediate cell–cell adhesion and to suppress invasion. *Oncogene* **22**, 5716–5719 (2003).

41. H.-W. Kwak, I. J. Choi, S.-J. Cho, J. Y. Lee, C. G. Kim, M.-C. Kook, K. W. Ryu, Y.-W. Kim, Characteristics of gastric cancer according to *Helicobacter pylori* infection status. *J. Gastroenterol. Hepatol.* **29**, 1671–1677 (2014).

42. N. Murata-Kamiya, Y. Kurashima, Y. Teishikata, Y. Yamahashi, Y. Saito, H. Higashi, H. Aburatani, T. Akiyama, R. M. Peek, T. Azuma, M. Hatakeyama, *Helicobacter pylori* CagA interacts with E-cadherin and deregulates the β-catenin signal that promotes intestinal transdifferentiation in gastric epithelial cells. *Oncogene* **26**, 4617–4626 (2007).

43. R. M. Genta, D. Y. Graham, Comparison of biopsy sites for the histopathologic diagnosis of *Helicobacter pylori*: A topographic study of *H. pylori* density and distribution. *Gastrointest. Endosc.* **40**, 342–345 (1994).

44. E. J. Duell, N. Sala, N. Travier, X. Muñoz, M. C. Boutron-Ruault, F. Clavel-Chapelon, A. Barricarte, L. Arriola, C. Navarro, E. Sánchez-Cantalejo, J. R. Quirós, V. Krogh, P. Vineis, A. Mattiello, R. Tumino, K.-T. Khaw, N. Wareham, N. E. Allen, P. H. Peeters, M. E. Numans, H. B. Bueno-de-Mesquita, M. G. H. van Oijen, C. Bamia, V. Benetou, D. Trichopoulos, F. Canzian, R. Kaaks, H. Boeing, M. M. Bergmann, E. Lund, R. Ehrnström, D. Johansen, G. Hallmans, R. Stenling, A. Tjønneland, K. Overvad, J. N. Ostergaard, P. Ferrari, V. Fedirko, M. Jenab, G. Nesi, E. Riboli, C. A. González, Genetic variation in alcohol dehydrogenase (ADH1A, ADH1B, ADH1C, ADH7) and aldehyde dehydrogenase (ALDH2), alcohol consumption and gastric cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. *Carcinogenesis* **33**, 361–367 (2012).

45. T. Shimazu, I. Tsuji, M. Inoue, K. Wakai, C. Nagata, T. Mizoue, K. Tanaka, S. Tsugane; Research Group for the Development and Evaluation of Cancer Prevention Strategies in Japan, Alcohol drinking and gastric cancer risk: An evaluation based on a systematic review of epidemiologic evidence among the japanese population. *Jpn. J. Clin. Oncol.* **38**, 8–25 (2008).

46. P. A. Drew, P. M. Clifton, J. T. LaBrooy, D. J. Shearman, Polyclonal B cell activation in alcoholic patients with no evidence of liver dysfunction. *Clin. Exp. Immunol.* **57**, 479–486 (1984).

47. S. Pasala, T. Barr, I. Messaoudi, Impact of alcohol abuse on the adaptive immune system. *Alcohol Res.* **37**, 185–197 (2015).

48. G. Keller, H. Vogelsang, I. Becker, S. Plaschke, K. Ott, G. Suriano, A. R. Mateus, R. Seruca, K. Biedermann, D. Huntsman, C. Döring, E. Holinski-Feder, A. Neutzling, J. R. Siewert, H. Höfler, Germline mutations of the E-cadherin(CDH1) and TP53 genes, rather than of RUNX3 and HPP1, contribute to genetic predisposition in German gastric cancer patients. *J. Med. Genet.* **41**, e89 (2004).

49. M. W. Curtis, Q. P. Ly, M. J. Wheelock, K. R. Johnson, Evidence that the V832M E-cadherin germ-line missense mutation does not influence the affinity of α-catenin for the cadherin/catenin complex. *Cell Commun. Adhes.* **14**, 45–55 (2007).

50. Y.-C. Chen, W.-L. Fang, R.-F. Wang, C.-A. Liu, M.-H. Yang, S.-S. Lo, C.-W. Wu, A. F.-Y. Li, Y.-M. Shyr, K.-H. Huang, Clinicopathological Variation of Lauren Classification in Gastric Cancer. *Pathol. Oncol. Res.* **22**, 197–202 (2016).

51. S. K. Shukla, K. N. Prasad, A. Tripathi, U. C. Ghoshal, N. Krishnani, H. Nuzhat, Quantitation of *Helicobacter pylori* ureC gene and its comparison with different diagnostic techniques and gastric histopathology. *J. Microbiol. Methods* **86**, 231–237 (2011).

52. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The sequence alignment/ map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

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

E-cadherin gene variants in gastric cancer families whose probands are diagnosed with diffuse gastric cancer. *Int. J. Cancer* **101**, 434–441 (2002).

54. R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, C. Swanton, deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).

55. B. B. Campbell, N. Light, D. Fabrizio, M. Zatzman, F. Fuligni, R. de Borja, S. Davidson, M. Edwards, J. A. Elvin, K. P. Hodel, W. J. Zahurancik, Z. Suo, T. Lipman, K. Wimmer, C. P. Kratz, D. C. Bowers, T. W. Laetsch, G. P. Dunn, T. M. Johanns, M. R. Grimmer, I. V. Smirnov, V. Larouche, D. Samuel, A. Bronsema, M. Osborn, D. Stearns, P. Raman, K. A. Cole, P. B. Storm, M. Yalon, E. Opocher, G. Mason, G. A. Thomas, M. Sabel, B. George, D. S. Ziegler, S. Lindhorst, V. M. Issai, S. Constantini, H. Toledano, R. Elhasid, R. Farah, R. Dvir, P. Dirks, A. Huang, M. A. Galati, J. Chung, V. Ramaswamy, M. S. Irwin, M. Aronson, C. Durno, M. D. Taylor, G. Rechavi, J. M. Maris, E. Bouffet, C. Hawkins, J. F. Costello, M. S. Meyn, Z. F. Pursell, D. Malkin, U. Tabori, A. Shlien, Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056.e10 (2017).

56. M. Nagasaki, J. Yasuda, F. Katsuoka, N. Nariai, K. Kojima, Y. Kawai, Y. Yamaguchi-Kabata, J. Yokozawa, I. Danjoh, S. Saito, Y. Sato, T. Mimori, K. Tsuda, R. Saito, X. Pan, S. Nishikawa, S. Ito, Y. Kuroki, O. Tanabe, N. Fuse, S. Kuriyama, H. Kiyomoto, A. Hozawa, N. Minegishi, J. Douglas Engel, K. Kinoshita, S. Kure, N. Yaegashi; ToMMo Japanese Reference Panel Project, M. Yamamoto, Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).

57. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

58. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur; Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

59. M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

60. P. Polak, J. Kim, L. Z. Braunstein, R. Karlic, N. J. Haradhavala, G. Tiao, D. Rosebrock, D. Livitz, K. Kübler, K. W. Mouw, A. Kamburov, Y. E. Maruvka, I. Leshchiner, E. S. Lander, T. R. Golub, A. Zick, A. Orthwein, M. S. Lawrence, R. N. Batra, C. Caldas, D. A. Haber, P. W. Laird, H. Shen, L. W. Ellisen, A. D. D'Andrea, S. J. Chanock, W. D. Foulkes, G. Getz, A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).

Citation: A. Suzuki, H. Katoh, D. Komura, M. Kakiuchi, A. Tagashira, S. Yamamoto, K. Tatsuno, H. Ueda, G. Nagae, S. Fukuda, T. Umeda, Y. Totoki, H. Abe, T. Ushiku, T. Matsuura, E. Sakai, T. Ohshima, S. Nomura, Y. Seto, Y. Shibata, Y. Rino, A. Nakajima, M. Fukayama, S. Ishikawa, H. Aburatani, Defined lifestyle and germline factors predispose Asian populations to gastric cancer. *Sci. Adv.* **6**, eaav9778 (2020).