Mutational landscape of gastric adenocarcinoma in Chinese: Implications for prognosis and therapy

Kexin Chen1,*, Da Yang1,2,3,4,5, Xiangchun Li4,6,*, Baocun Sun1, Fengju Song5, Wenfeng Cao1, Daniel J. Brat1, Zhibo Gao4, Haixin Li1, Han Liang4, Yanrui Zhao3, Hong Zheng1, Miao Li1, Jan Buckner1, Scott D. Patterson1, Xiang Ye6, Christoph Reinhard1, Anahita Bhathenam1, Deepa Joshi1, Paul S. Mischel1, Carlo M. Croce1, Yi Michael Wang1, Pharamcogenetics, Department of Pharmaceutical Sciences, University of Pittsburgh, Pittsburgh, PA 15261;2Department of Pathology, University of Texas MD Anderson Cancer Center Informatics Center, Houston, TX 77030;3Center of Pharmacogenetics, Department of Pharmaceutical Sciences, University of Pittsburgh, Pittsburgh, PA 15261;4Beijing Genomics Institute-Shenzhen, Shenzhen 518083, Guangdong, People’s Republic of China;5Department of Medicine and Therapeutics, State Key Laboratory of Digestive Disease, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong 999077, People’s Republic of China;6Department of Oncology Discovery, Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA 30322;7Division of Medical Oncology, Mayo Clinic, Rochester, MN 55905;8Medical Sciences, Amgen Inc., Thousand Oaks, CA 91320;9Lilly China Research and Development Center and Translational Sciences Oncology, Eli Lilly and Company, Indianapolis, IN 46285;10Translational Oncology & Personalized Medicine, AbbVie Inc., North Chicago, IL 60064;11Exploratory Clinical & Translational Research, Bristol-Myers Squibb India, Mumbai 400013, India;12Applied Genomics and Bioinformatics, Bristol-Myers Squibb, Princeton, NJ 08540;13Ludwig Institute for Cancer Research, University of California at San Diego, La Jolla, CA 92039;14Human Cancer Genetics Program, Ohio State University, Columbus, OH 43210;15The National Foundation for Cancer Research, Bethesda, MD 20814;16Asian Fund for Cancer Research, Hong Kong, People’s Republic of China; and17Department of Biology, University of Copenhagen, DK-1165 Copenhagen, Denmark

Contributed by Webster K. Cavenee, December 15, 2014 (sent for review November 20, 2014)

Gastric cancer (GC) is a highly heterogeneous disease. To identify potential clinically actionable therapeutic targets that may inform individualized treatment strategies, we performed whole-exome sequencing on 78 GCs of differing histologies and anatomic locations, as well as whole-genome sequencing on two GC cases, each with three primary tumors and two matching lymph node metastases. The data showed two distinct GC subtypes with either high-clonality (HiC) or low-clonality (LoC). The HiC subtype of intratumoral heterogeneity was associated with older age, TP53 (tumor protein P53) mutation, enriched C > G transition, and significantly shorter survival, whereas the LoC subtype was associated with younger age, ARID1A (AT rich interactive domain 1A) mutation, and significantly longer survival. Phylogenetic tree analysis of whole-genome sequencing data from multiple samples of two patients supported the clonal evolution of GC metastasis and revealed the accumulation of genetic defects that necessitate combination therapies. The most recurrently mutated genes, which were validated in a separate cohort of 216 cases by targeted sequencing, were members of the homologous recombination domain DNA repair, Wnt, and PI3K-ERBB pathways. Notably, the drugable NRG1 (neuregulin-1) and ERBB4 (V-Erb-B2 avian erythroblastic leukemia viral oncogene homolog 4) ligand-receptor pair were mutated in 10% of GC cases. Mutations of the BRCA2 (breast cancer 2, early onset) gene, found in 8% of our cohort and validated in The Cancer Genome Atlas GC cohort, were associated with significantly longer survivals. These data define distinct clinicogenetic forms of GC in the Chinese population that are characterized by specific mutation sets that can be investigated for efficacy of single and combination therapies.

Significance

We have identified a lethal subtype of gastric cancer (GC) that is characterized by high levels of clonal heterogeneity and TP53 (tumor protein P53) mutation. We have also uncovered key novel mutations in the targetable NRG1 (neuregulin-1) and ERBB4 (V-Erb-B2 avian erythroblastic leukemia viral oncogene homolog 4) ligand-receptor pair and identified BRCA2 (breast cancer 2, early onset) mutations as new genetic markers to predict better survival for GC. Our study represents a novel approach for GC personalized medicine and identified novel clinical actionable therapies for GC therapy.

clonality | exome sequencing | mutation | ERBB | BRCA2

Gastric cancer (GC) is the fourth most common cancer and the second leading cause of cancer death worldwide, accounting for 8% of all newly diagnosed cancers and 10% of cancer mortality(1). Environmental risk factors for GC include smoking, and infectious agents (1), including the bacterium Helicobacter pylori (2), and Epstein Barr Virus (3). Consistent with its complicated etiology (e.g., diet) and anatomical environment, GC is clinically and pathologically highly heterogeneous (4), with a large variation in 5-y survival rates in different countries, and even different cities in the same country (5, 6). This clinical heterogeneity is mirrored by concomitant heterogeneous molecular signatures in GC mRNA, protein, and miRNA expression profiles (7, 8). Standard treatment strategies have largely ignored the heterogeneity and individuality of different subtypes of GC. The current approach entails surgical removal of the tumor followed by adjuvant fluoropyrimidine, taxane, and platinum-based chemotherapy doublets or triplets, especially for advanced GC, and this is exacerbated by the lack of reliable markers to predict response. Recently, the US Food and Drug Administration and the European Medicines Agency have approved Trastuzumab for patients with HER2-overexpressing metastatic GC, which represent less than 15% of the disease population. The high incidence of GC in Asian countries and its increasing incidence in Western countries point to a clear need for developing more effective therapies for GC as well as the discovery of markers that predict their therapeutic response.

Genome sequencing has emerged as a powerful tool to identify potential driving oncogenic targets for therapeutic intervention. Wang et al. sequenced 22 samples from Hong Kong GC patients and identified mutations in genes involved in chromatin modification
(e.g., ARID1A) (9). Two recent and more comprehensive genomic characterizations of Russian and Hong Kong GC by the The Cancer Genome Atlas (TCGA) and Wang et al. (10, 11), respectively, reported anatomical and histological subtype-specific genetic and epigenetic alterations. Both studies highlighted RHOA (ras homolog family member A) mutation as a novel oncogenic driver for the diffuse subtype of GCs, which was functionally validated by Wang et al. (11) and Kakieuchi et al. (12). Hotspot mutations in RHOA were shown to result in defective RHOA signaling and promote escape from anoikis in organoid cultures. These studies revealed new insights into the genetic and epigenetic alterations in GC, but their lack of sufficient clinical follow-up data limited their potential translation into clinical utility.

Our international consortium sought to determine whether novel or clinically actionable events for Chinese GC could be identified by focusing on its heterogeneity (anatomical location, known histologic subtypes, sex) and analyzing samples with comprehensive clinical follow-up. This strategy allows us to report herein several clinically important findings. First, we identified a new lethal subtype of GC that is characterized by high levels of clonal heterogeneity. Second, through whole-genome sequencing (WGS) of multiple samples from two patients, we found that clonal divergence occurs intratumorally and mutations accumulate during clonal metastasis. Third, we uncovered key novel mutations in the targetable neuregulin-1 (NRG1) and ERBB4 ligand-receptor pair in 10% of GC cases. Finally, we show that BRCA2 mutations serve as new genetic markers to predict longer survival for GC.

Results

Frozen tissue samples of primary gastric adenocarcinoma from 294 northern Chinese patients not treated with chemotheraphy or radiotherapy prior were derived from surgical resection specimens and obtained from the Tianjin Medical University Cancer Institute and Hospital-National Foundation for Cancer Research Joint Tissue Banking Facility (Dataset S1). Germ-line DNA was available from matching blood samples and used as a sequence reference to detect somatic alterations. All patients signed informed consent, and the study protocol was approved by the Institutional Review Board from Tianjin Medical University Cancer Institute and Hospital. We performed whole-exome sequencing (WES) on 78 samples (discovery cohort) and WGS on 110 additional cases with discrepant diagnosis among pathologists (five cases, dependent review by at least two senior pathologists, and in this study cases with a median follow-up interval (CI), 1.62–3.16; P = 0.0043), indicating that the clonal subtype is an independent prognostic factor for GC. The distinct clinical behavior of the HiC and LoC subtypes suggest they may have fundamental differences in tumorigenesis, as was recently proposed to be manifested by various nucleotide level mutational signatures (17, 18). Detailed mutational signature analysis of these two subtypes showed a significant difference in the percentages of C-to-T transition and C-to-T transversion mutations. HiC GCs had a significantly smaller fraction of C-to-T transversion mutation (45% in HiC vs. 52% in LoC) but greater fraction of C-to-G transition mutation (25% in HiC vs. 9% in LoC, P = 0.002). Specifically, the HiC subtype had a significantly higher fraction of C > G mutation in almost all contexts, except for the G in the 5′-adjacent to C (Dataset S4 A–C).

Identification of High Clonality and Low Clonality Subtypes in GC by Clonality Analysis. The 78 GC samples in discovery panel and their matching blood cells were subjected to WES. A mean coverage depth of 167X was achieved for GC tumor DNA, and 170X for matched normal blood cells. Eighty-five percent of exons were covered at 20X or greater depth for both tumor and normal samples (Fig. S1 and Dataset S2). We applied MuTect (13) and VarScan2.2.5 (14) to whole-exome data to detect somatic point and indel mutations. A total of 13,866 mutations were detected in this cohort, including 3,421 synonymous, 8,558 missense, 576 nonsense, 241 splice site, 967 frameshift, and 103 nonframeshift indels. The overall mutation frequency was 112.5 mutations (range of 1–1,486 mutations) with a median of 82.5 for nonsilent mutations per tumor (Dataset S3).

To gain insight into GC heterogeneity, we performed SNV clonality analysis using a recently reported algorithm SciClone (15), which estimates the clusters of clones in each sample. This analysis identified cluster numbers ranging from 1 to 8 in our cohort with a bimodal distribution dipping at a cluster number of 4. We thus defined those GC cases with cluster numbers larger than 4 as high clonal or HiC (nine cases) (Fig. L4 and Dataset S4) and cases with cluster numbers equal to or fewer than 4 as low clonal or LoC (68 cases) (Fig. 1B and Dataset S4). Association analyses with clinical parameters revealed that the HiC subtype was associated with older disease onset age (Fig. 1C) (Pearson correlation coefficient = 0.31, P = 0.02). There were a total of 11,768 protein-coding subclonal SNVs (92.3% of all detected mutations, with a median of 106 mutations per sample) (Dataset S3). Gene-level mutation analysis revealed that the HiC subtype had a significantly higher frequency of TP53 mutations (Fig. 1D), whereas the LoC subtype was enriched for ARID1A mutations (Fig. 1E), both thought to be the driving events in GC tumorigenesis. Previous studies of leukemia (16) have reported that the higher clonal complexity was associated with poor survival, likely because of chemoresistance. Among our cohort of 78 GC patients, 32 (41.03%) died from GC with a median follow-up of 25.1 months (range from 0.20 to 78.95 months) (Dataset S1). Survival analysis showed that the HiC subtype was significantly associated with shorter survival compared with the LoC subtype (P = 0.02, log-rank test) (Fig. 1F). There were no significant differences in the treatment regimens between the HiC and LoC subtypes. This association remained significant after adjusting for age, sex, stage, Lauren subtype, TP53 mutation, and ARID1A mutation [adjusted hazard ratio (HR), 4.69; 95% confidence interval (CI), 1.62–13.6; P = 0.0043], indicating that the clonal subtype is an independent prognostic factor for GC. The fact that a subset of the HiC and LoC subtypes suggest they may have fundamental differences in tumorigenesis, as was recently proposed to be manifested by various nucleotide level mutational signatures (17, 18). Detailed mutational signature analysis of these two subtypes showed a significant difference in the percentages of C-to-G transition and C-to-T transversion mutations. HiC GCs had a significantly smaller fraction of C-to-T transversion mutation (45% in HiC vs. 52% in LoC) but greater fraction of C-to-G transition mutation (25% in HiC vs. 9% in LoC, P = 0.002). Specifically, the HiC subtype had a significantly higher fraction of C > G mutation in almost all contexts, except for the G in the 5′-adjacent to C (Dataset S4 A–C).

Although TP53 mutations were commonly found in the HiC group, they were present only in the minor clones (mutation frequency < 15%) in all six positive HiC cases (Dataset S4 D and E). Conceivably, any of the minor clones can become a major clone after a single agent therapy, which has been illustrated in a study in leukemia (16). Thus, in principle, all of the subclones in a tumor should be eliminated for curative therapy. We used a drug–gene interaction database, DGIdb (19), to search for druggable targets for the clones present in the HiC tumors (Dataset S4F). As expected, many gene mutations and the clones harboring them do not have corresponding drugs available. However, for tumors like PGM71 and PGM32, targetable mutations were identified for almost all subclones [e.g., ALK, ABL2, SMAD4 (SMAD family member 4), FANCG, NRG1, KRAS (Kirsten rat sarcoma viral oncogene homolog), and others]
(Dataset S4G). In these cases, combination chemo- and targeted therapies might be much more effective than those using single agents. For other HiC cases the targetable mutations were observed in only two or three clones, underscoring the need for drug development strategies that test multiple combinations (Dataset S4H).

Clonal Evolution of GC During Metastasis. To define the type and extent of intratumoral heterogeneity as GC progresses to metastasis, we performed WGS of two cases (Pt1 and Pt2) where, for each case, we obtained the information by dissection and anatomical mapping of three different areas of primary tumor and two lymph node metastases. Pt1 had a total of 4,082 nonsilent mutations, including 2,809 missense, 143 nonsense, 132 splice site, 958 frameshift, and 40 in-frame indels. Pt2 had a lower mutation rate with 287 nonsilent mutations (253 missense, 14 nonsense, 10 splice site, 3 frameshift, and 7 in-frame). Consistent with the higher mutation rate in Pt1, mutations were found in several DNA-mismatch repair and chromatin remodeling genes, including **MSH6, TGFBR2, KDM5A, and MLL4**. In Pt2, the **SETBP1** gene (p.S944N) was mutated, as it is in myeloid malignancies (20), together with amplification of **AKAP9 and CDK6** (cyclin-dependent kinase 6). Intriguingly, mutations of the gene encoding the voltage-dependent calcium channel protein, **CACNA1D** p.V529G, was found in both Pt1 and Pt2. Copy number alteration analysis showed that a large proportion of the genome was altered in both Pt1 and Pt2, accounting for 516.6 Mb and 706.8 Mb, respectively. Pt2 exhibited consistent genomic amplification of 7q21.2–7q21.3, and Pt1 had amplifications across chr7 and chr8. Pt1 and Pt2 had a comparable number of structural variations with 3,812 and 3,390 breakpoints, respectively.

Because clonal evolution is closely linked to cancer metastasis, we applied phylogenetic tree analysis ([SI Methods](#)) to the five samples for each GC case and observed considerable divergence among the three primary tumor samples in each (Fig. 2 and Fig. S2). Interestingly, only one part of the primary tumor was closely related to the two highly similar lymph-node metastases, suggesting they were derived from a common clonal ancestor. The multiple clones uncovered within primary tumor and lymph-node metastases suggests the necessity of combined therapy that targets different genetic changes found within the patient’s combined tumor burden. We therefore used DGIdb (19) to search for potentially druggable targets in different parts of the same tumor and their resultant metastases. Results of the drug–sample interaction analysis suggest that Pt1 might benefit from combined targeting of PI3K, CDK7, and Notch (as well as immunotherapy) (Fig. 2), whereas combined AURKC and CDK6 inhibitors might benefit Pt2 (Fig. S2).

**Validated NRGs-ERBBs Mutations in GC.** Among the total of 13,866 mutations, we also sought to identify clinically actionable mutations in GC. We analyzed the WES data of the 78 GCs with statistical approaches including MutSigCV ([Methods](#)) and identified 16 significantly (q value < 0.2) mutated genes. Of

---

**Fig. 1.** Identification of HiC and LoC subtypes by clonality analysis and their clinical outcomes. (A and B) SciClone analysis with cluster number 4 as cut-off identified 9 HiC and 68 LoC cases. Each peak infers one subclone by SciClone algorithm. (C) Association between clonality (y axis) and age at diagnosis (x axis). (D) Fractions of TP53 mutation carriers (y axis) in different clonality GC (x axis). (E) Fraction of ARID1A mutation carriers (y axis) in different clonality GC (x axis). (F) Kaplan–Meier survival curves for HiC and LoC cases.

**Fig. 2.** Clonal evolution in GC metastasis revealed by phylogenetic tree. WGS data from five samplings (three primary and two lymph node metastases) of one GC case were used in the similarity analysis of somatic subatracutions. Branch and trunk lengths are proportional to the number of non synonymous mutations acquired on the corresponding branch or trunk. Driver mutations and copy number events were acquired by the indicated genes in the branches the arrows indicate. The potential drugs which target each tumor portion are shown as nodes with arrows pointing to the samples.
these, 13 were mutated in 5% or more of the tumors. The analysis identified previously reported gene mutations in GC: TP53, ARID1A, CDH1 (cadherin 1, type 1, E-cadherin (epithelial)), APC (adenomatous polyposis coli), RHOA, PIK3CA (phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha), SMAD4, MYC (v-myc avian myelocytomatosis viral oncogene homolog), and KRA. Further pathway analysis of recurrent mutated genes revealed that mutated genes in our cohort belonged to the TP53 pathway, Wnt pathway, ERBB pathway, and HR pathway, and so forth (Dataset S5). Based on mutation frequency and novelty (Dataset S3), we selected 103 recurrently mutated genes in Wnt (Fig. S3), ERBB (Fig. S4), and HR (Fig. S5) pathways for targeted sequencing verification in an additional 216 GC samples (Dataset S1) using the Ion-Torrent platform (Methods). Among the 103 genes, 86 (83%) were found have at least one mutation in the validation cohort (Dataset S3 and S6–S10).

The extensive follow-up data associated with our GC cohort allowed us to evaluate the clinical significance of many of the mutations identified in this and previous studies. We first evaluated if there are some genes specifically associated with different anatomical locations in GC. Using 78 WES samples, we identified 10, 3, and 4 genes, which showed moderate enrichment in antrum [i.e., CSMD3 (30.7%), TSHZ3 (15.4%), PCDHAI1 (19.2%), ARHGAP28 (15.4%), DST (23.1%), SORCS1 (19.2%), MUC17 (15.4%), PCDH20 (19.2%), USP9X (19.2%), CACNA2D1 (15.4%), body [i.e., FLG2 (15.4%), GRM1 (15.4%), SYNE2 (15.4%),] and cardia [i.e., CDH1 (23.1%), GFRAL (15.4%), SBNO1 (15.4%), STAB2 (15.4%)], respectively (P < 0.05). Seven (the underlined ones) of those genes remained significant when we included the extra 216 targeted sequencing cases. Intriguingly, none of these gene mutations were observed to significantly associate with same anatomical location in TCGA data, suggesting different genetic background (Russian vs. Chinese) or even diets might contribute to the mutation distribution. In addition to anatomical location, we also interrogated the association between gene mutation and Lauren subtypes of GC. We found that CDH1 somatic mutations (21 of 294 cases, 7.1%) were strongly enriched in the diffuse GC (Fig. S6A) and 8 of these 21 somatic CDH1 mutations were loss-of-function types (e.g., nonsense and frameshift mutations). RHOA mutations were also enriched in diffuse GC, consistent with recent reports (11, 12). Interestingly, Wnt pathway mutations showed a moderate enrichment among intestinal GC. Survival analysis showed that the CDH1-mutated GCs had shorter survival, whereas Wnt pathway mutated GCs had longer survival (log-rank P = 0.04) (Fig. S6B). This difference was not a result of their association with the two Lauren classifications because we did not observe significantly different survival of these latter two groups in our cohort. Interestingly, RHOA-mutated cases were not associated with shorter survivals similar to CDH1-mutated GCs, suggesting RHOA driven and CDH1 driven diffuse GCs are different clinically.

Although ERBB2 has been recently recognized as a therapeutic target for GC because of its gene amplification and mutation (confirmed in our study), our analysis revealed that another member of the ERBB pathway, ERBB4, was significantly mutated in both our discovery and validation cohorts. Mutation of ERBB4 or HER4 has recently been shown to be an oncogenic driver in melanoma (21), although this has not been previously described for GC. Mutated ERBB4 activates both ERBB4 and PI3K-AKT signal transduction and can be inhibited by the dual-tyrosine kinase inhibitor Lapatinib, for which efficacy is presently being evaluated in phase III clinical trials in ERBB4-mutated melanoma (clinicaltrials.gov/ct2/show/NCT01264081). In our discovery cohort, ERBB4 was mutated in seven samples. In the validation cohort, another 13 samples had ERBB4 mutations. Among these mutations, ERBB4 p.R50C, was also observed in melanoma (Fig. 3A). Unlike ERBB2, which has no known ligand, ERBB4 binds to NRGs, particularly NRG1, forming homo- and heterodimers, and leading to the subsequent activation of kinase activity. Previous studies have shown that overexpression of NRG1 leads to the activation of ERBBs and consequent Lapatinib-sensitivity (22). However, no somatic mutation of NRG1 has been reported previously. In our discovery cohort, NRG1 was mutated in five samples. In the validation cohort, NRG1 mutations were observed in another 16 cases. Taking these data together, we find that NRG1 and ERBB4 were mutated in 34 GC samples (11.6%), with 3 samples having mutations in both genes and 31 samples having mutations in either NRG1 or ERBB4 (Fig. 3). Detailed analysis showed that 7 of 21 (33%) ERBB4 mutations occurred in the kinase domain and 4 (20%) in the receptor domains, suggesting these mutations might impact receptor ligand interaction and kinase activity. Interrogation of the 289 GC cases in the TCGA database also revealed ERBB4 mutations present in both the receptor domain (i.e., p.R106H/C) and the kinase domain (i.e., p.V744L and p.774N/G). We also found NRG1 mutations in the TCGA cohort, including the mutations...
(i.e., p.A221T and p.A225P) in the EGF-like domain (predicted as the functional domain to bind ERBB4), as well as in regions adjacent to the EGF-like domain (i.e., p.E223G, p.R224Q, and p.S226P). These findings suggest that the EGF-like domain is a hotspot mutation for NRG1. Mutations were also present in other NRGs and ERBB family proteins (Dataset S11). Mutual exclusivity analysis was conducted for ERBB and NRG gene members in the combined Tianjin and TCGA cohorts. In 1,000-times permutation, the ERBBs and NRGs’ mutations were mutually exclusive of each other (P = 0.02).

GCs with BRCA2 Mutations Show a Trend for Longer Survival. Adjunctive chemotherapy as the standard of care for GC has limited success and there is no reliable marker to predict response. Among our 78 WES cohort of GC cases, 5 had BRCA2 mutations (interestingly, there were no BRCA1 mutations). Correspondingly, in the 216 targeted sequencing cohort cases, another 12 had BRCA2 mutations (10 somatic and 2 germ line) (Dataset S3 and S5). A higher frequency (28 of 289 cases) of BRCA2 mutations was also found in the TCGA Russian GC cohort (Fig. 4A). This finding is consistent with epidemiological studies showing that GC is the third major cancer type, joining breast and ovarian cancer, with BRCA2 mutations (23). We hypothesized that the BRCA2 mutation might be clinically important for GC because of accumulating evidence demonstrates that such mutations are associated with improved response to platinum-based chemotherapy and longer survival in ovarian cancer (24). BRCA2 mutant carrier patients are marginally correlated with longer survival in ovarian cancer (24). This finding is of potential significance because inhibitors for ERBB signaling are available and in clinical trials. Our study also revealed some mutation events that were not reported previously.

The most important finding to emerge from this study is the description of intratumoral heterogeneity, which was revealed by clonality analysis. The observation that mutational signatures of two subtypes of GC are distinct from each other indicates there might be fundamental difference in etiology (such as mutagens) of two subtypes, which cause different mutational patterns. The GCs with the highest heterogeneity, which we termed the HiC subtype, have the poorest survival. This result could be because of the escape of specific subclones from the therapeutic effect of single agents. Gene-level analysis further identified TP53 and ARID1A mutation as markers for HiC and LoC, respectively, which may have translational significance in clinic, although this will need to be validated in the future.

Conclusion and Discussion
The present study of Chinese GC identified a number of commonly mutated genes that have been reported in other populations (Hong Kong and Russian) including TP53, ARID1A, CDH1, APC, RHOA, PIK3CA, SMAD4, MYC, and KRAS. Our study also revealed some mutation events that were not reported previously. Notably, the NGR1 and ERBB4 genes were commonly mutated in our cohort. Whether this is more specific to the Chinese population compared with others remains to be replicated. However, this finding is of potential significance because inhibitors for ERBB signaling are available and in clinical trials. Our study also revealed BRCA2 mutation as a predictor of longer survival, similar to that

**Fig. 4.** BRCA2 mutation carriers have longer survival. (A) Distribution of coding mutations in BRCA2 gene. Red: mutations in Tianjin cohort; green: mutations in TCGA cohort. The number of total mutations is more than the number of patients is because some patient has multiple BRCA2 mutations. (B) Kaplan–Meier survival curves for BRCA2 wild-type and BRCA2 mutated GCs. (C) Forest plot of multivariable Cox-regression model including clinical and BRCA2 mutations.

**Table 1.** Distribution of Mutations in the EGF-like Domain of ERBBs and NRGs. Each domain contains three subdomains: EGF-like, ESD-like, and ETV-like.
reported in ovarian cancer (24). It has been proposed that the reason BRCA2 mutations are associated with improved survival is that BRCA2 mutated tumor cells have a diminished capacity to repair damaged DNA caused by chemotherapy. Interestingly, platinum-based chemotherapy is frequently a component of frontline therapy for GC, similar to ovarian cancer. One notable distinction of BRCA2 mutation profiles in GC is that most are nonsense or frameshift mutations (Fig. 4A), whereas mutations in triple-negative breast cancer and ovarian cancer are nonsense or frameshift. This finding suggests that most of the BRCA2 mutations found in GC are deleterious mutations, although this has to be functionally validated in the future. Nevertheless, our study identified BRCA2 mutations as a prognostic marker that could be clinically useful for GC management and clinical trial design.

Methods

**Illumina-Based Whole-Exome Sequencing.** The whole-exome capture procedure was performed with Agilent’s SureSelect Human All Exon Kit protocol. Resulting DNA libraries with an insert size of 200 bp on average were sequenced using the 90-bp paired-end technology on Illumina HiSeq. 2000. Real-time image analysis and base calling were performed by HiSeq Control Software v1.1.37 and Real Time Analysis v1.7.45 using standard parameters, respectively.

**PGM Sequencing.** Next-generation sequencing was performed using the Personal Genome Machine (PGM, LT).

**Statistical Analysis.** Student’s t test, analysis of variance, χ², Wilcoxon rank-sum test, Fisher’s exact test, Kaplan–Meier estimate, and Mantel–Cox survival analyses were performed using R 2.10.0. Significance was defined as P < 0.05. Benjamini–Hochberg multiple testing correction was used to estimate the false-discovery rate when multiple testing correction was applied.

**ACKNOWLEDGMENTS.** This research was funded by the grants from The National Foundation for Cancer Research and the Asian Fund for Cancer Research.

**Supporting Information**

**Chen et al. 10.1073/pnas.1422640112**

**SI Methods**

**Identifying Significantly Mutated Pathways.** We obtained mutation rates for each category (i.e., C = *CpG→T, *Cp(A/C/T)→T, A→G, transver, and null+indel) from MutSigCV when identifying significantly mutated genes to estimate the expected number of somatic mutations for every gene in a given pathway, and aggregate them to obtain somatic mutation counts in this pathway under the null hypothesis. Specifically, the mutation rates for these five categories are denoted as \( r = \{ r_{*CpG→T}, r_{*Cp(A/C/T)→T}, r_{A→G}, r_{\text{transver}}, r_{\text{null+indel}} \} \), and a pathway \( P \) with \( k \) genes \( P = \{ g_1, g_2, \ldots, g_k \} \), and the number of bases for each category in gene denoted as \( n_g = \{ n_{g*CpG→T}, n_{g*Cp(A/C/T)→T}, n_{gA→G}, n_{g\text{transver}}, n_{g\text{null+indel}} \} \); whereas \( r \) and \( n_g \), in this case the number of expected somatic mutations (\( \lambda_P \)) in pathway \( P \), can be computed as:

\[
\lambda_P = \sum \sum n_g * r.
\]  

The \( P \) value of observing \( X_P \geq K \) somatic mutations given the expected mutation count \( \lambda_P \) is:

\[
p(X_P \geq K | P) = \frac{\lambda_P^K e^{-\lambda_P}}{K!}
\]  

The Benjamini–Hochberg method of false-discovery rate control was used to convert \( P \) values into \( q \) values.

**Phylogenetic Tree Reconstruction for a Patient with Multiple Tumors.** Phylogenetic tree analysis can be used to trace evolution relationship among a patient with multiple tumors. Here we used two strategies to reconstruct phylogenetic trees for two GC patients each with three primary and two lymph node metastatic tumors.

For the first method, we aggregated all of somatic SNVs in each patient to construct a mutation space (\( M \)) (i.e., the union of all of somatic SNVs identified in five tumors). We then determined mutation status for each somatic SNV in \( M \) for each tumor, 1 for mutation-positive, and 0 mutation-negative. In this case, we calculated the pairwise distances between two tumors in each patient, and hierarchical clustering algorithm was used to construct a dendrogram to infer evolution history.

In the second method, we derived a mutation sequence space for each tumor in \( M \), which covers all of somatic SNV sites, alternative nucleotide for mutation-positive and corresponding reference nucleotide for mutation-negative. Specifically, supposed in \( M \) we have 10 somatic SNVs, reference nucleotide corresponding to these 10 sites consists of DNA mutation sequence, for example, ACGTACGAAC.

Reference sequence: ACGTACGAAC:

- Tumor 1: ATGAACGAAC (C→T and T→A in the second and fourth positions);
- Tumor 2: ACCTACTAAC (G→C and G→T in the third and seventh positions);
- Tumor 3: AAGTTCGAAC (C→A and A→T in the second and fifth positions);
- Tumor 4: ACGTAGGTAC (C→G and A→T in the sixth and eighth positions);
- Tumor 5: ACTTACTAAC (G→T and G→T in the third and seventh positions).

After replacing reference nucleotides with corresponding alternative nucleotides, we obtained the mutation sequences for each tumor and applied standard phylogenetic tree analyses with functions provided in Matlab Bioinformatics Toolbox.

When comparing structure of phylogenetic trees from these two methods, we found that they share similar structure patterns, indicating the phylogenetic trees reconstructed for these two GC cases is quite robust.
**Fig. S1.** Quality control of deep seq data.

**Fig. S2.** Clonal evolution in GC metastasis revealed by phylogenetic tree. WGS data from five samplings (three primary and two lymph node metastases) of one GC case were used in the similarity analysis of somatic substitutions. Branch and trunk lengths are proportional to the number of nonsynonymous mutations acquired on the corresponding branch or trunk. Driver mutations and copy number events were acquired by the indicated genes in the branches the arrows indicate. The potential drugs which target each tumor portion are shown as nodes with arrows pointing to the samples.
Fig. S3. Distribution of coding mutations in Wnt pathway. The heatmap shows the matrix of synonymous and nonsynonymous Wnt pathway mutations in 78 WES and 216 targeted sequencing tumors. Each column denotes an individual tumor, and each row represents a gene. Mutations are colored by the types of mutations. The left column shows the number of mutations in each gene.

Fig. S4. Distribution of coding mutations in PIK3CA-ERBB pathway. The heatmap shows the matrix of synonymous and nonsynonymous PIK3CA-ERBB pathway mutations in 78 WES and 216 targeted sequencing tumors. Each column denotes an individual tumor, and each row represents a gene. Mutations are colored by the types of mutations. The left column shows the number of mutations in each gene.
Fig. S5. Distribution of coding mutations in DNA repair pathway. The heatmap shows the matrix of synonymous and nonsynonymous homologous-combination DNA repair pathway mutations in 78 WES and 216 targeted sequencing tumors. Each column denotes an individual tumor, and each row represents a gene. Mutations are colored by the types of mutations. The left column shows the number of mutations in each gene.

Fig. S6. Molecular diffused and intestinal gastric subtype defined by CDH1/RHOA and Wnt pathway mutation. (A) Fractions of CDH1/RHOA mutation carriers (x axis) in different GC subtype (y axis). (B) Kaplan–Meier survival curves for CDH1 and Wnt mutation carriers. (C) Fractions of Wnt pathway mutation carriers (x axis) in different GC subtype (y axis).

Dataset S1. Clinical information of GC sample cohort for WES and validation

Dataset S2. Coverage and mutation frequencies for individual samples generated by WES

Dataset S3. All somatic mutation detected by WES and validation cohort
Dataset S4. Clonality plot for each sample

(A) Mutation signature plots of nine HiC samples. (Left) Base substitutions from WES are divided into six categories to represent the six possible base changes color-coded in the pie chart. (Right) Base substitutions are further divided by the 16 possible flanking nucleotides surrounding the mutated base as listed in the trinucleotide context labels. (B) Mutation signature plots 68 LoC samples. G > G transitions are denoted by asterisks. The difference between HiC and LoC in A was significant, \( P < 0.01 \), \( \chi^2 \) test. (C) The mean frequency and SD for 16 possible C > G transition with flanking nucleotides. The mutation frequency is calculated separately for each case and Wilcoxon-rank sum test were applied to calculate the P values, \(* P < 0.05\), \(** P < 0.01\), \(*** P < 0.001\). All error bars, SD. (D and E) Clonality plot for HiC (D) and LoC (E) subtype GC case. (Upper) Schematic depiction of all 78 tumor samples consisting of multiple clones in tumor, each peak indicates a clone. (Lower) Schematic depiction of mutated allele frequency and coverage in each tumor sample. The cells of the secondary clone contain the majority of mutations present in the founder clone but have acquired a distinct set of new mutations not shared with the founder. (F) An ideogram of a normal karyotype is shown in the outer ring. Each ring inside represents mutated allele frequency and their genomic coordinates for one HiC case. On each ring, the size and color of dots indicate the frequencies of mutated alleles. Red color denotes most-abundant mutated alleles and yellow the least-abundant mutated alleles. The known drugable genes (in blue) and drugs (in black) are labeled at the corresponding gene locations outside ideogram. (G) Subclone and drugable genes for potentially treatable three representative HiC cases. (H) Subclone and potential drugable genes for currently untreatable three representative HiC cases.

Dataset S5. Pathway analysis result in WES data

Dataset S6. Mutation correlated with different anatomical locations in GC

Dataset S7. Mutation rates and transition-transversion ratios in regions of the genome as detected by WES

Dataset S8. Mutation rates of C > G only and P value in HiC and LoC samples

Dataset S9. Statistically significant protein-coding mutations in GC generated by WES

Dataset S10. Germ-line mutation detected by Ion Torrent in validation cohort

Dataset S11. Distribution of protein-coding mutations in significantly mutated genes
Supplementary Dataset 1. Clonality plot for each sample
Supplementary Dataset 1. Clonality plot for each sample
Supplementary Dataset 1 (Continued). Clonality plot for each sample.
Supplementary Dataset 1 (Continued). Clonality plot for each sample
Supplementary Dataset 1 (Continued). Clonality plot for each sample
Supplementary Dataset 1 (Continued). Clonality plot for each sample
Supplementary Dataset 2. Distribution of protein coding mutations in significantly mutated genes
Supplementary Dataset 2 (continued). Distribution of protein coding mutations in significantly mutated genes
Supplementary Dataset 2 (continued). Distribution of protein coding mutations in signifcantly mutated genes