

# Clinical Significance of Four Molecular Subtypes of Gastric Cancer Identified by The Cancer Genome Atlas Project



Bo Hwa Sohn<sup>1</sup>, Jun-Eul Hwang<sup>2</sup>, Hee-Jin Jang<sup>1,3</sup>, Hyun-Sung Lee<sup>3</sup>, Sang Cheul Oh<sup>4</sup>, Jae-Jun Shim<sup>5</sup>, Keun-Wook Lee<sup>6</sup>, Eui Hyun Kim<sup>1</sup>, Sun Young Yim<sup>1</sup>, Sang Ho Lee<sup>7</sup>, Jae-Ho Cheong<sup>8</sup>, Woojin Jeong<sup>9</sup>, Jae Yong Cho<sup>10</sup>, Joohee Kim<sup>1</sup>, Jungsoo Chae<sup>11</sup>, Jeeyun Lee<sup>12</sup>, Won Ki Kang<sup>12</sup>, Sung Kim<sup>13</sup>, Sung Hoon Noh<sup>8</sup>, Jaffer A. Ajani<sup>14</sup>, and Ju-Seog Lee<sup>1</sup>

## Abstract

**Purpose:** The Cancer Genome Atlas (TCGA) project recently uncovered four molecular subtypes of gastric cancer: Epstein–Barr virus (EBV), microsatellite instability (MSI), genomically stable (GS), and chromosomal instability (CIN). However, their clinical significances are currently unknown. We aimed to investigate the relationship between subtypes and prognosis of patients with gastric cancer.

**Experimental Design:** Gene expression data from a TCGA cohort ( $n = 262$ ) were used to develop a subtype prediction model, and the association of each subtype with survival and benefit from adjuvant chemotherapy was tested in 2 other cohorts ( $n = 267$  and 432). An integrated risk assessment model (TCGA risk score) was also developed.

**Results:** EBV subtype was associated with the best prognosis, and GS subtype was associated with the worst prognosis.

Patients with MSI and CIN subtypes had poorer overall survival than those with EBV subtype but better overall survival than those with GS subtype ( $P = 0.004$  and  $0.03$  in two cohorts, respectively). In multivariate Cox regression analyses, TCGA risk score was an independent prognostic factor [HR, 1.5; 95% confidence interval (CI), 1.2–1.9;  $P = 0.001$ ]. Patients with the CIN subtype experienced the greatest benefit from adjuvant chemotherapy (HR, 0.39; 95% CI, 0.16–0.94;  $P = 0.03$ ) and those with the GS subtype had the least benefit from adjuvant chemotherapy (HR, 0.83; 95% CI, 0.36–1.89;  $P = 0.65$ ).

**Conclusions:** Our prediction model successfully stratified patients by survival and adjuvant chemotherapy outcomes. Further development of the prediction model is warranted. *Clin Cancer Res*; 23(15); 4441–9. ©2017 AACR.

<sup>1</sup>Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas. <sup>2</sup>Department of Hematology-Oncology, Chonnam National University Medical School, Gwangju, Korea. <sup>3</sup>Division of Thoracic Surgery, Michael E. DeBakey Department of Surgery, Baylor College of Medicine, Houston, Texas. <sup>4</sup>Division of Hemato-Oncology, Department of Internal Medicine, Korea University Medical Center, Korea University College of Medicine, Seoul, Korea. <sup>5</sup>Department of Internal Medicine, Kyung Hee University School of Medicine, Seoul, Korea. <sup>6</sup>Department of Internal Medicine, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, Korea. <sup>7</sup>Department of Surgery, Kosin University College of Medicine, Busan, Korea. <sup>8</sup>Department of Surgery, Yonsei University College of Medicine, Seoul, Korea. <sup>9</sup>Department of Life Science, Research Center for Cellular Homeostasis, Ewha Womans University, Seoul, Korea. <sup>10</sup>Department of Medical Oncology, Yonsei University College of Medicine, Seoul, Korea. <sup>11</sup>Department Obstetrics and Gynecology, Kyung Hee University School of Medicine, Kyung Hee University Hospital at Gangdong, Seoul, Korea. <sup>12</sup>Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Gangnam-Gu, Seoul, Korea. <sup>13</sup>Department of surgery, Samsung Medical Center, Gangnam-Gu, Seoul, Korea. <sup>14</sup>Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas.

**Note:** Supplementary data for this article are available at Clinical Cancer Research Online (<http://clincancerres.aacrjournals.org/>).

B.H. Sohn, J.-E. Hwang, H.-J. Jang, and H.-S. Lee contributed equally to this article.

**Corresponding Author:** Ju-Seog Lee, MD Anderson Cancer Center, 6565 MD Anderson Blvd, Houston TX 77054. Phone: 713-834-6154; Fax: 713-563-4235; E-mail: jlee@mdanderson.org

doi: 10.1158/1078-0432.CCR-16-2211

©2017 American Association for Cancer Research.

## Introduction

Gastric cancer is the fourth most common cancer and third leading cause of cancer-related death worldwide, accounting for an estimated annual 723,100 deaths (1–3). Surgical resection with subsequent adjuvant chemotherapy has been established as an effective treatment for patients with early-stage gastric cancer (4–7). However, recurrence occurs in up to 30% to 40% of patients within 5 years (4, 8–10), suggesting that gastric cancer is a clinically heterogeneous disease. The inherent clinical heterogeneity is most likely due to differences in molecular characteristics of cancer cells. In past years, molecular profiling studies of gastric cancers at a whole-genome level have revealed several genetic and epigenetic changes underlying gastric carcinogenesis (11–13). However, these findings have not been translated to clinical practice, largely owing to the lack of a strong genetic-clinical association.

Recently, as part of The Cancer Genome Atlas (TCGA) project, the genome and proteome of gastric cancer have been extensively characterized to uncover molecular subtypes and identify dysregulated pathways and potential therapeutic targets (14). Integrative analysis of multiple genomic and proteomic data from gastric cancer tissues revealed four molecular subtypes: (i) Epstein–Barr virus (EBV) subtype with extreme DNA hypermethylation, (ii) microsatellite instability (MSI) subtype with elevated mutation rates and hypermethylation, (iii) genomically stable (GS) subtype with less distinctive

### Translational Relevance

Molecular classification of cancers has significantly improved patient care by the development of treatments tailored to the genetic or epigenetic abnormalities specific to molecular subtypes as evidenced by improved clinical care in breast cancer after the discovery of ER-positive and *HER2*-positive subtypes and subsequent development of tailored treatment targeting ER or *HER2*. For gastric cancer, such important subtypes with characteristic molecular features are likely to exist. The Cancer Genome Atlas project recently uncovered four molecular subtypes of gastric cancer. However, their clinical significances are currently unknown. Thus, we analyzed multiple data sets and uncovered clinical association of each molecular subtype. This may help clinicians stratify patients according to molecular characteristics and develop tailored treatments for each subtype in future.

genomic alterations, and (iv) chromosomal instability (CIN) subtype with marked aneuploidy and frequent focal amplification of receptor tyrosine kinases. Although distinct gastric cancer subtypes with striking genomic features would provide a roadmap for the development of personalized treatment strategies, the clinical relevance of these subtypes, such as their implications for prognosis or response to standard treatment, is currently unknown owing to a lack of sufficient follow-up data from patients in the TCGA cohort. Furthermore, because molecular stratification of gastric cancer in the TCGA study was based on highly complicated integrative analysis of multiple genomic and proteomic data sets, it is very unlikely that these findings will be easily translated into clinical practice.

In the current study, we analyzed gene expression data from TCGA project, uncovered gene expression signatures specific to each of the four molecular subtypes, developed prediction models for stratification of patients with gastric cancer by subtype using these signatures, and tested our model in two large independent cohorts. We also found that the subtypes were predictors of survival outcomes and response to standard adjuvant chemotherapy.

### Materials and Methods

#### TCGA cohort data and the 4 subtypes of gastric cancer

Genomic data from the TCGA gastric cancer cohort were downloaded from the TCGA data portal site (<http://cancergenome.nih.gov/>) and processed as described in previous studies (14–22). By applying integrative analysis of multiple genomic and proteomic data from gastric cancer tissues, including somatic mutations, mRNA expression, miRNA expression, promoter methylation, somatic copy-number alteration, and protein expression data from reverse phase protein arrays, the TCGA classification scheme employed a decision tree whereby gastric tumors were divided into four subtypes (14). Briefly, tumors were first categorized by the presence of EBV features (EBV subtype), then by the presence of high MSI (MSI subtype). The remaining tumors were further grouped by the number of somatic copy-number alterations: genomically stable (GS subtype) or chromosomal instability (CIN subtype). In the TCGA

cohort, mRNA expression data were generated by RNA sequencing, and this information was available for 262 tumor tissue samples. Because most tissues in the TCGA cohort were recently collected, follow-up time for patients in the TCGA cohort was very short and incomplete. Thus, TCGA cohort data from 262 patients were used to generate our model (see below) but were not used for our survival analyses.

#### Patients and tissue samples for validation cohorts

Demographic information and clinical data, as well as tissue samples, were obtained from 267 patients with gastric cancer who had undergone gastrectomy as primary treatment between 1999 and 2006 at Korea University Guro Hospital, Kosin University College of Medicine, or Yonsei University Severance Hospital, South Korea. All patients underwent a D2 gastrectomy, and all tissues were snap-frozen and stored in  $-80^{\circ}\text{C}$  freezer. Informed consent for sample collection had been obtained from all patients. Generation and analysis of genomic data were carried out at The University of Texas MD Anderson Cancer Center (MDACC, Houston, TX), and these data were designated as the MDACC cohort. Our study was approved by the Institutional Review Boards of MD Anderson Cancer Center and each institute that provided tissues. Of the 267 patients in MDACC cohort, 155 had received standard adjuvant chemotherapy [either single-agent 5-fluorouracil (5-FU) or a combination of 5-FU and cisplatin/oxaliplatin, doxorubicin, or paclitaxel]. Patients in stage I and stage IV with metastasis were not included in subset analysis for assessing benefit of adjuvant chemotherapy.

For an external validation cohort, we used tumor specimen data collected from patients with gastric cancer at the Samsung Medical Center (SMC;  $n = 432$ ), as described in a previous study (23).

#### Generation of gene expression data from human gastric cancer tissues

All experiments and analyses were done in the Department of Systems Biology at the MDACC. Gene expression data from the 267 patients in MDACC cohort were generated by hybridizing labeled RNAs to HumanHT-12 v3.0 Expression BeadChips (Illumina). Briefly, total RNA was extracted from the fresh-frozen tissues using a mirVana RNA isolation labeling kit (Ambion). We used 500 ng of total RNA for labeling and hybridization according to the manufacturer's protocols. The microarray data were normalized using the quantile normalization method in the Linear Models for Microarray Data package in the R language environment (24). The expression level of each gene was transformed into a  $\log_2$  base before further analysis. Primary microarray data from the MDACC cohort are available in the Gene Expression Omnibus (GEO) database of the National Center for Biotechnology Information (NCBI; accession numbers GSE13861 and GSE26942).

Gene expression data from the SMC cohort were generated using HumanRef-8 WG-DASL v3.0 (Illumina) that contained a subset (24,526 gene features) of probes in Human HT-12. Primary microarray data are available in the GEO database of the NCBI (accession number GSE26253).

#### Generation of gene expression data, subtype prediction, and TCGA risk score for recurrence

Gene expression data were generated and analyzed as explained in previous studies (25–29). BRB-ArrayTools

(National Institutes of Health) were used for all statistical analyses of gene expression data (30). We first generated a subtype prediction model using data from the TCGA cohort. For selection of subtype-specific gene sets, multiple two-class *t* tests were performed for all possible combinations of the four subtypes. Gene expression differences were considered statistically significant if the *P* value was less than 0.001. Only genes with significant differences in expression in all 3 possible comparisons were considered subtype-specific genes, yielding 349 significant genes for the EBV subtype, 455 for the MSI subtype, 1,513 for the GS subtype, and 143 for the CIN subtype. The top 200 significant genes in each subtype and all 143 genes for the CIN subtype were further selected for development of the prediction model.

To develop a subtype prediction model, we adopted a previously developed model using Bayesian compound covariate predictor algorithms (25–29). Briefly, gene expression data for each subtype gene signature (i.e., the 200 significant genes for each subtype, as described above) were used to generate the Bayesian probability of each tissue sample belonging to a particular subtype. We applied 0.4 as the cutoff of Bayesian probability for each predictor. With this cutoff, the sensitivity and specificity of each predictor ranged from 0.8 to 1 in the training set (the TCGA cohort). Receiver operating characteristic (ROC) analysis of this training set indicated the following order of strength for each predictor: EBV > MSI > GS > CIN (Supplementary Fig. S1); therefore, we adopted a TCGA classification scheme employing a decision tree whereby tumors are grouped into the four subtypes. Briefly, new samples in the test cohorts (i.e., the MDACC and SMC cohorts) were assigned to one of the four subtypes according to Bayesian probability scores. When new samples had more than 2 probability scores above the cutoff value, samples were assigned according to the predetermined strength of the predictors. Samples lacking probability scores above the cutoff value were not assigned

to any subtype. Same prediction algorithm was applied to gene expression data from gastric cancer cell lines.

#### Development of the TCGA risk score for recurrence

We developed an integrated risk assessment model by pooling the probabilities of the four predictors (subtypes). Because EBV and MSI were associated with good prognosis, we used the inverse of the probability for these subtypes to determine risk of recurrence. GS probability was weighted by a factor of 2 to reflect its strong association with poor prognosis. CIN probability was not modified because it was only moderately associated with poor prognosis.

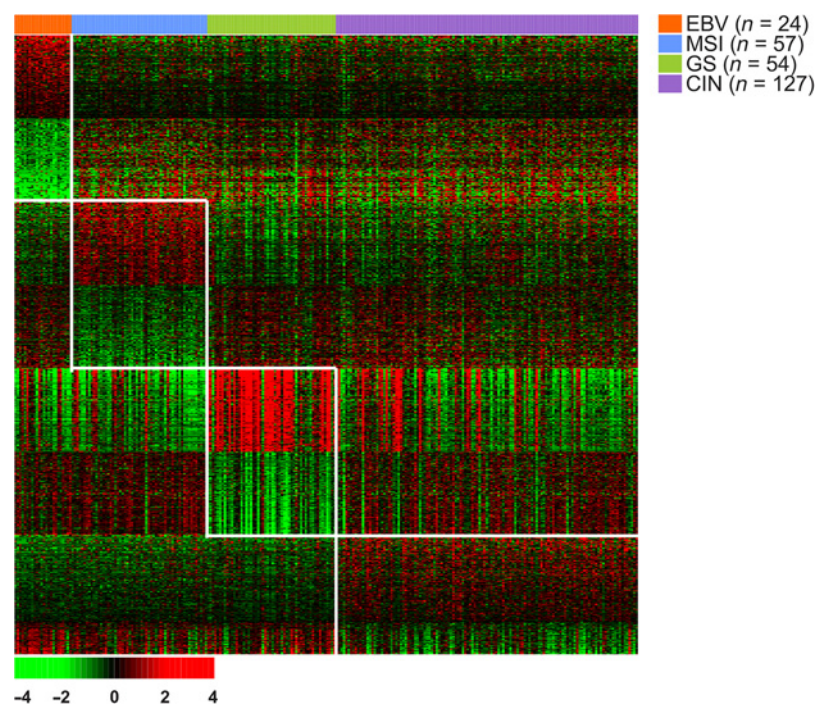
TCGA Risk Score raw (TRSraw) = (1 – EBV probability) + (1 – MSI probability) + (GS probability × 2) + CIN probability. To generate a dynamic range of scores from 0 to 100, we reformulated TRSraw: TRS =  $e^{\text{TRSraw}}$ . This generated TRS values ranging from 3.2 to 85.27. Cutoff points were specified to reflect prognostic differences: low risk (<20), intermediate risk (20–30), and high risk of recurrence (>30).

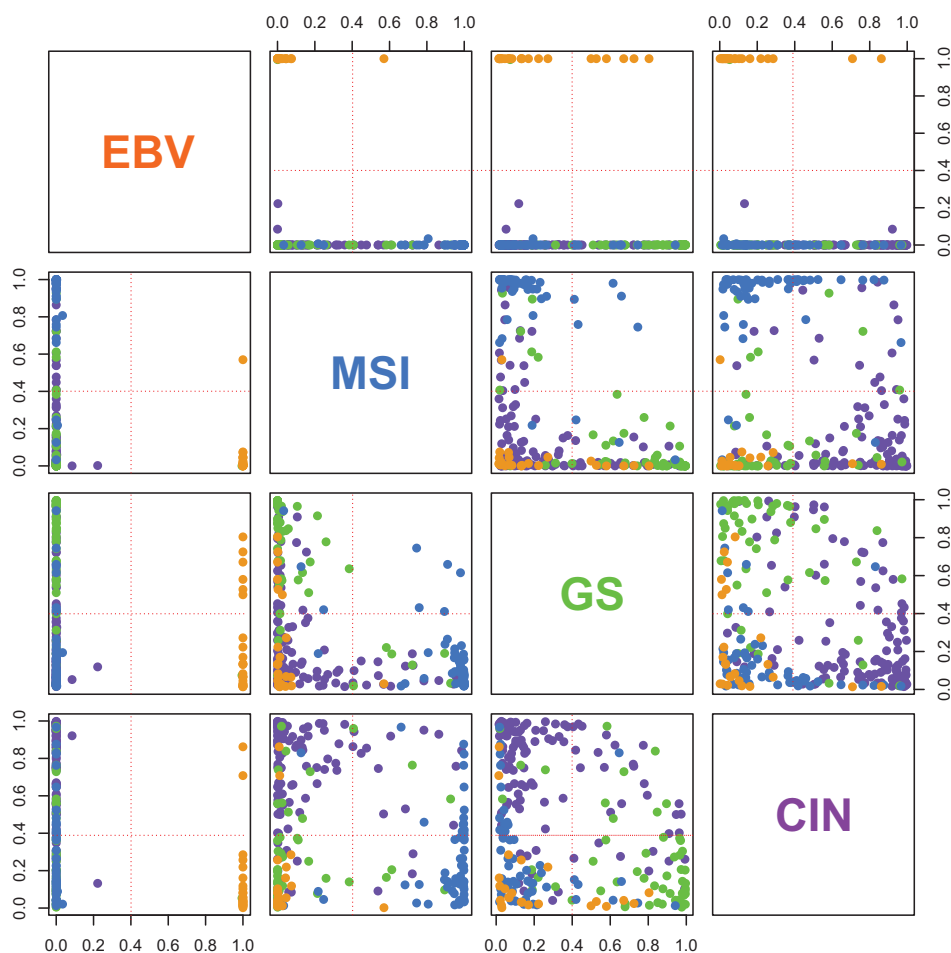
#### Statistical analysis

The association of each subtype with overall survival and recurrence-free survival (RFS) in the MDACC cohort was estimated using Kaplan–Meier plots and log-rank tests. Overall survival was defined as the time from surgery to death, and RFS was defined as the time from surgery to the first confirmed recurrence. Data were censored when a patient was alive without recurrence at last follow-up. Multivariate Cox proportional hazards regression analysis was used to evaluate independent prognostic factors associated with RFS and overall survival, including TRS, tumor stage, and pathologic characteristics as covariates. A *P* value of less than 0.05 was considered statistically significant. To assess the association of each molecular subtype with benefit from adjuvant chemotherapy, we fitted a Cox proportional hazards model to

**Figure 1.**

Prediction signatures for four molecular subtypes of gastric cancer in TCGA project cohort. Subtype-specific gene expression signatures were identified by applying multiple *t* tests ( $P < 0.001$ ). Among the significant genes in each subtype, the top 200 genes were selected for development of prediction models (all 143 significant genes were used for the CIN subtype). Data are presented in matrix format in which each row represents an individual gene and each column represents a tissue sample. Each cell in the matrix represents the expression level of a gene feature in an individual tissue sample. The coloring in the cells reflects relatively high (red) and low (green) expression levels, as indicated in the scale bar ( $\log_2$  transformed scale).





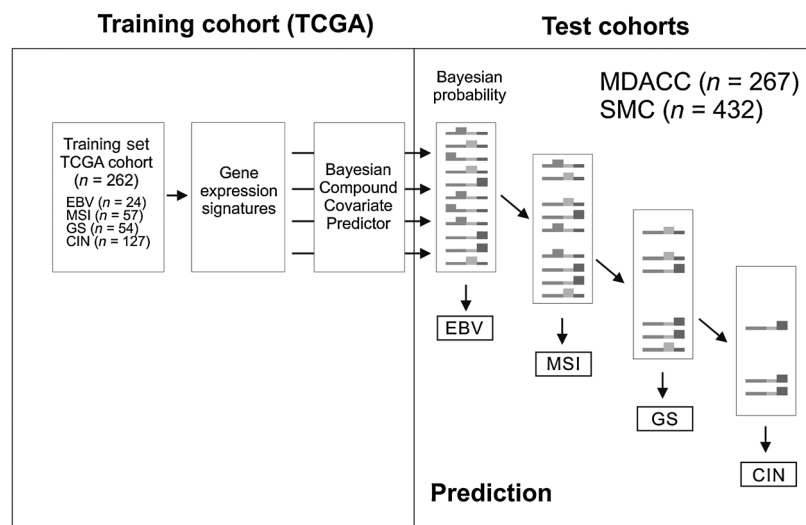
**Figure 2.** Scatter plot matrix of Bayesian probability for each gastric cancer subtype predictor in the TCGA cohort.

data from patients in MDACC cohort. All statistical analyses were conducted in the R language environment (<http://www.r-project.org>). Ingenuity pathway analysis (Ingenuity) was used for gene set enrichment analysis and gene network analysis to identify enriched gene sets and upstream regulators in each subtype.

## Results

### Subtype-specific predictors

The subtype-specific gene signatures for each of the four subtypes are shown in Fig. 1 and Supplementary Table S1. We next constructed a prediction model using a Bayesian compound



**Figure 3.** Schematic diagram of TCGA project prediction model. A decision tree approach was employed for categorizing patients in test cohorts into the four subtypes of gastric cancer according to the Bayesian probability of each predictor.

covariate predictor algorithm and tested the strength of each signature (29). When ROC analysis was carried out to the Bayesian probability of each sample, the EBV signature showed highest sensitivity and specificity (100% AUC), and the CIN signature showed the lowest sensitivity and specificity (89.6% AUC). The MSI and GS signatures showed sensitivity and specificity that were lower than EBV but higher than CIN (MSI: 97.7% AUC; GS: 95.4% AUC; Supplementary Fig. S1).

Using the scatter plot matrix approach (Fig. 2), we determined the cutoff of Bayesian probability for each predictor to have reasonably high sensitivity and specificity. Under a cutoff of 0.4, the sensitivity and specificity of each predictor ranged from 0.8 to 1 in the training set (Supplementary Table S2). For construction of the prediction model for the four subtypes, we adopted a TCGA classification scheme that employs a decision tree whereby tumors are grouped into the four subtypes according to Bayesian probability scores: first EBV, then MSI, then GS, and lastly CIN (Fig. 3).

### Prognostic significance of the TCGA gastric cancer subtypes

After establishing a robust, highly sensitive and specific prediction model from the TCGA cohort data to categorize patients with gastric cancer by subtype, we examined the association of each subtype with prognosis using gene expression data from the MDACC cohort ( $n = 267$ ; Supplementary Table S3). When

patients in the MDACC cohort were classified by subtype using our prediction model, the EBV subtype was associated with the best prognosis, for both RFS ( $P = 0.006$  by the log-rank test) and overall survival ( $P = 0.004$ ; Fig. 4A). The GS subtype was associated with the worst prognosis. Patients with the MSI and CIN subtypes had a moderate prognosis that was worse than that of patients with the EBV subtype but better than that of patients with the GS subtype (Fig. 4A).

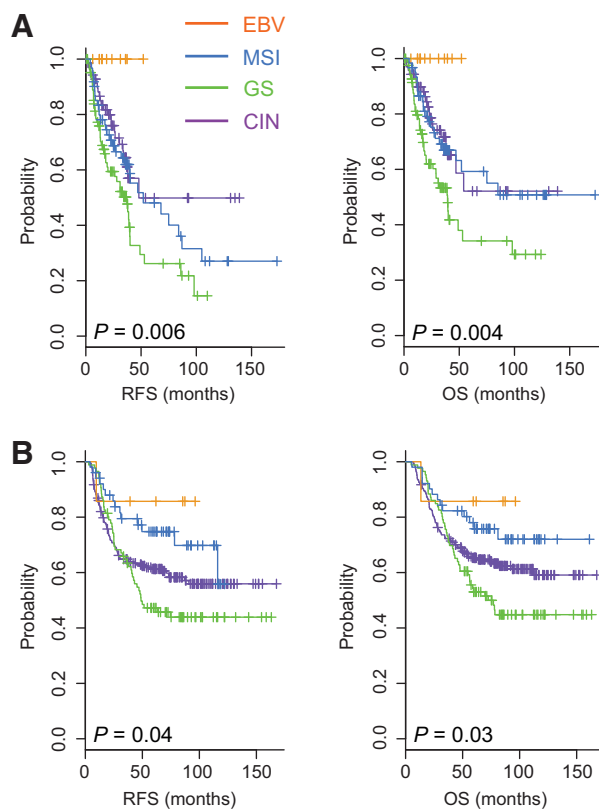
When the prediction model was applied to the SMC cohort ( $n = 432$ ; Supplementary Table S3), the GS subtype was associated with the worst prognosis and the EBV subtype was associated with the best prognosis (Fig. 4B). Consistent with the MD Anderson cohort, the MSI subtype was associated with moderate prognosis. However, the prognosis of patients with the CIN subtype was poorer in the SMC cohort than in the MD Anderson cohort, suggesting that the CIN subtype might represent a less homogeneous subgroup.

Because favorable prognosis of patients in EBV subtype is most intriguing, we next carried out gene set enrichment analysis of EBV-specific gene expression signature using ingenuity pathway analysis. It revealed that genes involved in energy production and metabolism are most significantly altered in EBV subtype (Supplementary Table S4). More interestingly, vast majority of metabolic genes are downregulated in EBV subtype (Supplementary Fig. S2).

Patients with the MSI subtype were diagnosed at older ages (median age 60) relative to patients with the other subtypes, whereas those with the GS subtype were diagnosed at relatively younger ages (median age, 52;  $P = 9.4 \times 10^{-7}$  by the Student *t* test; Supplementary Fig. S3). In addition, most patients with the EBV subtype were male (79%) and diagnosed at younger ages (median age, 53;  $P = 0.01$  by Student *t* test; Supplementary Fig. S3). This observation is highly consistent with reports from TCGA and previous studies (14, 31, 32), suggesting that our prediction model grouped patients with similar genetic and clinical characteristics to those observed in each subtype in the TCGA data. Taken together, these findings support the robustness of our prediction model and indicate an association between the molecular subtypes and clinical outcomes, suggesting that molecular characteristics of gastric cancer reflected in genomic and proteomic patterns may dictate clinical outcomes.

### Biological characteristics of subtypes

We next carried out gene network analysis to uncover potential upstream regulators of subtype-specific genes which may contribute to clinical and biological characteristics of each subtype. Not surprisingly, vast majority of predicted upstream regulators of EBV subtype genes are cytokines such as IL1B, IL2, IL3, IL21, IL27, and INFG (Supplementary Table S5), suggesting that infiltrated immune cells in EBV tumors are highly active due to viral infection. Interestingly, in contrast to EBV subtype, many of these cytokines are suppressed in CIN tumors, suggesting that high copy-number alteration may play roles in cancer immunity by suppressing activation of immune cells. This observation is in good agreement with recent study showing strong correlation of tumor aneuploidy with immune evasion across 12 different cancer types (33). MSI subtype is characterized by activation of EZH2, subunit of histone methyltransferase Polycomb repressive complex 2 (PRC2) that functions in the regulation of gene silencing (34), suggesting that further silencing of gene expression by EZH2 may play roles in progression of MSI subtype.



**Figure 4.** Prognosis associated with each of the four subtypes of gastric cancer in two independent patient cohorts. Patients in the MDACC cohort (A) and SMC cohort (B) were stratified by subtype. RFS and overall survival (OS) were plotted for each subtype.

Interestingly, GS subtype is characterized by activation of mir-21, a best known oncomir gene (35), as evidenced by downregulation of its many targets genes in GS subtype (Supplementary Fig. S4).

#### TCGA subtypes and adjuvant chemotherapy

Because adjuvant chemotherapy is the standard treatment for gastric cancer and more than half of patients in the MD Anderson cohort had received adjuvant chemotherapy (5–7), we next sought to determine whether specific subtypes were associated with an increased clinical benefit from adjuvant chemotherapy. We carried out a subset analysis of patients in the MD Anderson cohort with American Joint Commission on Cancer (AJCC) stage II, III, or IV disease without distant metastasis ( $n = 157$ ); patients with advanced gastric cancer have been shown to benefit most from adjuvant chemotherapy (7). Of 157 patients, 116 received adjuvant chemotherapy. Patients with the CIN subtype exhibited the greatest benefit from adjuvant chemotherapy, as evidenced by significantly increased RFS rates ( $P = 0.03$ ; Fig. 5A). The 3-year RFS rate was 58.7% for those who received chemotherapy, compared with 33.5% for those who did not. The HR for recurrence among those who received adjuvant chemotherapy was 0.39 [95% confidence interval (CI), 0.16–0.94,  $P = 0.03$ ; Fig. 5B]. However, no benefit from adjuvant chemotherapy was observed among patients with the GS subtype ( $P = 0.66$ ; Fig. 5A). The HR for recurrence among those who received adjuvant chemotherapy was 0.83 (95% CI, 0.36–1.89,  $P = 0.65$ ; Fig. 5B). Patients with the MSI subtype showed only moderate benefit from adjuvant chemotherapy ( $P = 0.18$  by log-rank test; HR, 0.55; 95% CI, 0.22–1.3,  $P = 0.18$ ; Fig. 5). The benefit of adjuvant chemotherapy could not be assessed for patients with the EBV subtype because all patients received chemotherapy.

To further demonstrate association of subtypes with chemoresistance/chemosensitivity, we used gene expression data and  $IC_{50}$  values of 5-FU from 24 gastric cancer cell lines that are

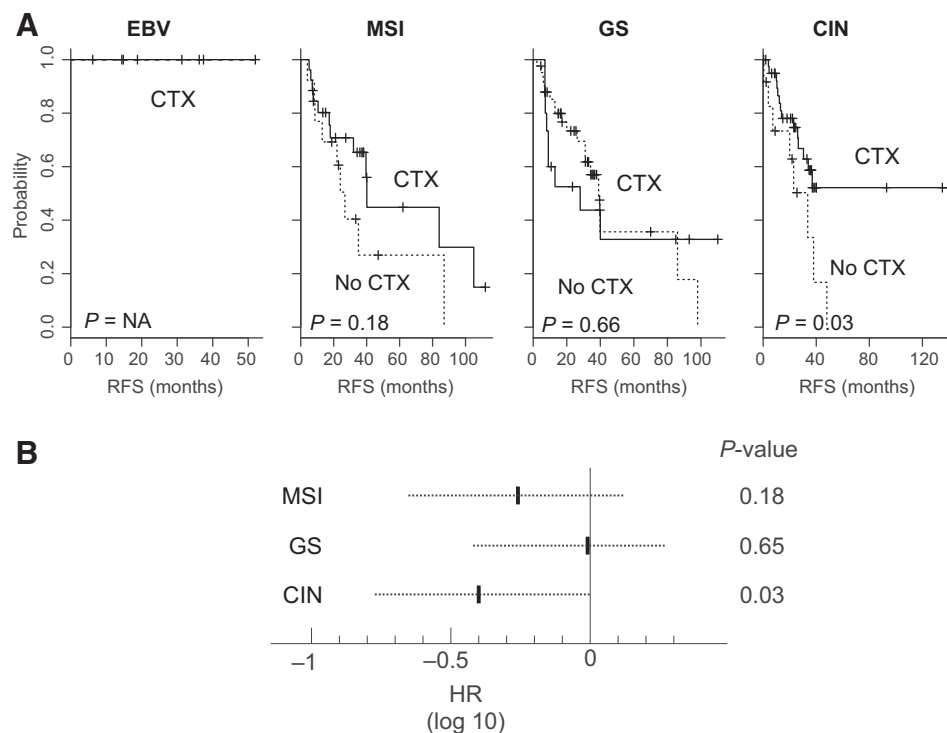
available from Genomics of Drug Sensitivity in Cancer project (36). As seen in gene expression data from tumor tissues, similar fraction of gastric cell lines were grouped into four subtypes according to subtype-specific gene expression signatures (Supplementary Fig. S5). In good agreement with clinical observation, GS cell lines have highest  $IC_{50}$ , whereas CIN cell lines have very low  $IC_{50}$ , suggesting that GS cells are indeed resistant to 5-FU.

#### TCGA risk score

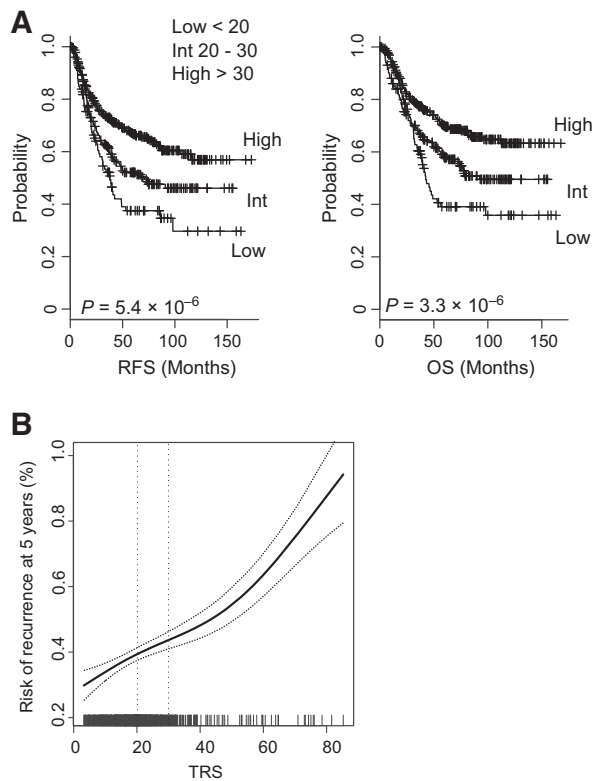
We next developed an integrated risk assessment model (TRS) by pooling the probabilities of the four predictors (subtypes). When patients in the MDACC and SMC cohorts were pooled into a single cohort ( $n = 699$ ) and stratified according to TRS, the 5-year RFS rate was 66.7% for low-risk patients, 52.1% for intermediate-risk patients, and 37.5% for high-risk patients ( $P = 5.4 \times 10^{-6}$  by log-rank test; Fig. 6A).

To evaluate the prognostic value of the TRS in combination with other clinical variables, we next carried out univariate and multivariate Cox proportional hazards regression analyses with combined clinicopathologic variables in the pooled cohort (MD Anderson + SMC). In addition to T stage, N stage, distant metastasis, and AJCC stage, which are well-known prognostic factors, TRS was a significant predictor of RFS in the univariate analysis (Supplementary Table S6). When we included all relevant clinical variables in a multivariate Cox regression analysis, TRS remained a significant prognostic factor (HR, 1.5; 95% CI, 1.2–1.9,  $P = 0.001$ ).

Because stage II gastric cancer is considered to be heterogeneous (37), we carried out a subset analysis in patients with stage II disease. In both univariate and multivariate analyses, TRS was a significant predictor of RFS (Supplementary Table S7), providing independent predictive value beyond tumor location, T stage, number of nodes examined, and histologic findings.



**Figure 5.** Benefit of chemotherapy among patients with each subtype of gastric cancer. **A**, Kaplan-Meier plots of RFS among patients who received adjuvant chemotherapy (CTX) and those who did not (no CTX) for each subtype.  $P$  values were obtained using the log-rank test. **B**, Cox proportional hazards regression analysis estimating the benefit of adjuvant chemotherapy for patients with each subtype. The dotted line represents the 95% CIs of the HRs.



**Figure 6.**

TCGA project risk score (TRS). **A**, Kaplan-Meier plots of RFS and overall survival (OS) among patients stratified by TRS. *P* values were obtained using the log-rank test. **B**, Relationship between continuous TRS values and 5-year recurrence risk estimated by a Cox proportional hazards model. The dotted lines represent the 95% CIs of the HRs.

TRS showed linear correlation with probability of recurrence at 5 years (Fig. 6B). The distribution of TRS values across the four subtypes reflected the prognostic nature of each subtype; patients with the GS subtype had the highest TRS values and those with the EBV subtype had the lowest TRS values (Supplementary Fig. S6). Taken together, our data strongly suggest that TRS revealed underlying biology not captured by the traditional clinical and pathologic features.

## Discussion

The comprehensive integrative analysis of the genome and proteome of gastric cancer tissues from TCGA uncovered four molecularly distinct subtypes (14). Because the clinical value of subtype classification has been demonstrated in other cancers (38, 39), we developed a genomic prediction model for the subtype classification of gastric cancer using a statistically defined set of multiple genes and gene expression data, and we tested our model in two large independent cohorts (total of 699 patients). This allowed us to properly demonstrate the robustness of the model, as well as the prognostic value of the subtypes in relation to conventional clinical variables.

Our analysis showed that all of the subtypes were present in both test cohorts, clearly demonstrating the reproducibility of tumor classification developed by TCGA and the robustness of the

prediction models we developed in the current study. Moreover, clinical characteristics of patients in predicted subtypes are in good agreement with previous observations, further supporting the robustness of our prediction model. More importantly, the four subtypes showed different clinical courses in terms of overall survival and RFS. We found that patients with the EBV subtype had a better prognosis than patients with other subtypes in both cohorts, which is also consistent with a previous report showing that positive immunostaining for EBV was associated with a good prognosis (32, 40, 41). Previous study suggested increased immune response due to viral infection might prevent outgrowth of cancer cells (32). In addition to increased immune response, our data suggested that reduced metabolic activity leading to slow growth of cancer cells might contribute to better prognosis.

Previous TCGA data showed that mutation rates are highest in the MSI subtype and genomic copy-number alterations are highest in the CIN subtype (Supplementary Fig. S7). The EBV subtype is characterized by very high promoter methylation and frequent mutations in *PIK3CA*, and the GS subtype is characterized by low mutation rates, low copy-number alterations, and frequent mutations in *CDH1* and *RHOA*. Taken together with these TCGA data, our findings suggest that the molecular characteristics in each subtype may dictate clinical outcomes of patients with gastric cancer.

The subtype-specific genetic signatures used in our model predicted not only survival outcomes, but also the relative benefit of adjuvant chemotherapy. Our findings suggested that patients with the CIN subtype most benefitted from adjuvant chemotherapy. Adjuvant chemotherapy was associated with improved outcomes in patients with the CIN subtype but not in patients with the GS subtype. These findings suggest that gastric cancer cells of the GS subtype are resistant to chemotherapy, and frequently altered genes in the GS subtype might account for chemoresistance. For example, *NUPR1* is an activated transcription regulator in GS subtype (Supplementary Table S5), and recent studies demonstrated that it enhances chemoresistance in multiple cancers (42–44). However, although our data revealed an association between the CIN subtype and benefit from adjuvant chemotherapy, the retrospective design of our study limited the predictive nature of this association. Therefore, additional studies are warranted to fully assess the ability of this genomic prediction model to identify patients most likely to benefit from adjuvant chemotherapy and those who may need additional treatments.

Using our model, we also developed TRS, an intuitive scoring system ranging from 0 to 100 that can predict risk of recurrence after treatment. Our findings strongly suggest that TRS provides an independent value beyond conventional clinical variables. First, TRS was an independent predictor of recurrence in multivariable analyses, both in the pooled cohort and in the subset of patients with stage II disease. Second, TRS had a strong linear relationship with probability of recurrence at 5 years. Third, the TRS values for each of the subtypes were highly correlated with survival outcomes for the subtypes. Our study has some limitation such as retrospective nature in clinical data and sample collection and lack of diverse ethnic groups in our cohort. Because all patients in current study were from Korea, the clinical association of subtypes needs to be further investigated in larger prospective cohort.

In summary, by analyzing genomic data from TCGA and new data generated in the current study, we demonstrated the clinical significance of the four subtypes of gastric cancer and developed prediction models that can reliably stratify patients with gastric

cancer into these four subtypes. Our prediction model could be used to identify not only patients with a poor prognosis (GS subtype), but also those who would most benefit from adjuvant chemotherapy (CIN subtype). Further development of the prediction model will be necessary before it can be implemented into routine clinical practice. Simpler, more robust standardized tools such as qRT-PCR for measuring expression of these genes will be required for clinical use. Increasing the applicability of the prediction model in the clinic will also require selecting fewer genes that well represent each subtype. Nevertheless, the validation of our prediction model in two independent patient cohorts and the fact that the model reflects the biological characteristics associated with each subtype indicate that this prediction model could be used to develop rational therapy recommendations. If confirmed in prospective studies, the association between subtype and adjuvant chemotherapy outcomes might improve patient selection for treatment.

### Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

### Authors' Contributions

Conception and design: H.-S. Lee, W. Jeong, J.-S. Lee

Development of methodology: J.-E. Hwang, H.-S. Lee, J.-S. Lee

### References

- Jung KW, Won YJ, Kong HJ, Oh CM, Lee DH, Lee JS. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2011. *Cancer Res Treat* 2014;46:109–23.
- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin* 2016;66:7–30.
- Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;65:87–108.
- Bang YJ, Kim YW, Yang HK, Chung HC, Park YK, Lee KH, et al. Adjuvant capecitabine and oxaliplatin for gastric cancer after D2 gastrectomy (CLASSIC): A phase 3 open-label, randomised controlled trial. *Lancet* 2012;379:315–21.
- Cunningham D, Allum WH, Stenning SP, Thompson JN, van de Velde CJ, Nicolson M, et al. Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *N Engl J Med* 2006;355:11–20.
- Macdonald JS, Smalley SR, Benedetti J, Hundahl SA, Estes NC, Stemmermann GN, et al. Chemoradiotherapy after surgery compared with surgery alone for adenocarcinoma of the stomach or gastroesophageal junction. *N Engl J Med* 2001;345:725–30.
- Sasako M, Sakuramoto S, Katai H, Kinoshita T, Furukawa H, Yamaguchi T, et al. Five-year outcomes of a randomized phase III trial comparing adjuvant chemotherapy with S-1 versus surgery alone in stage II or III gastric cancer. *J Clin Oncol* 2011;29:4387–93.
- Aoyama T, Yoshikawa T, Watanabe T, Hayashi T, Ogata T, Cho H, et al. Survival and prognosticators of gastric cancer that recurs after adjuvant chemotherapy with S-1. *Gastric Cancer* 2011;14:150–4.
- Lee J, Lim dH, Kim S, Park SH, Park JO, Park YS, et al. Phase III trial comparing capecitabine plus cisplatin versus capecitabine plus cisplatin with concurrent capecitabine radiotherapy in completely resected gastric cancer with D2 lymph node dissection: The ARTIST trial. *J Clin Oncol* 2012;30:268–73.
- Sakuramoto S, Sasako M, Yamaguchi T, Kinoshita T, Fujii M, Nashimoto A, et al. Adjuvant chemotherapy for gastric cancer with S-1, an oral fluoropyrimidine. *N Engl J Med* 2007;357:1810–20.
- Wang K, Kan J, Yuen ST, Shi ST, Chu KM, Law S, et al. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet* 2011;43:1219–23.
- Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic muta-

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J.-E. Hwang, H.-J. Jang, H.-S. Lee, S.C. Oh, J.-J. Shim, S.H. Lee, J.-H. Cheong, W.K. Kang, S. Kim, S.H. Noh, J.-S. Lee  
 Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J.-E. Hwang, H.-J. Jang, H.-S. Lee, E.H. Kim, J.-H. Cheong, J. Kim, J. Chae, J.A. Ajani, J.-S. Lee  
 Writing, review, and/or revision of the manuscript: J.-E. Hwang, H.-J. Jang, H.-S. Lee, J.-J. Shim, K.-W. Lee, J.-H. Cheong, J.A. Ajani, J.-S. Lee  
 Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): E.H. Kim, J.-H. Cheong, J.Y. Cho, S. Kim, J.-S. Lee  
 Study supervision: H.-S. Lee, W.K. Kang, J.-S. Lee

### Grant Support

This work was supported by grants from the NIH (CA127672, CA129906, CA138671, and CA150229); 2016 Institutional Research Grant (IRG) and 2016 Sister Institute Network Fund (SINF) grant from The University of Texas MD Anderson Cancer Center; Korea National Research Foundation (2014M3C1A3051981) funded by the Ministry of Science, ICT & Future Planning; Scientific Research Center Program (2014R1A2A2A01003983); and Research Initiative grant from Korean Research Institute of Bioscience and Biotechnology (KRIBB).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received September 2, 2016; revised September 28, 2016; accepted April 3, 2017; published OnlineFirst July 26, 2017.

- tions in cell adhesion and chromatin remodeling genes. *Nat Genet* 2012;44:570–4.
- Zouridis H, Deng N, Ivanova T, Zhu Y, Wong B, Huang D, et al. Methylation subtypes and large-scale epigenetic alterations in gastric cancer. *Sci Transl Med* 2012;4:156ra40.
- The Cancer Genome Atlas. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 2014;513:202–9.
- The Cancer Genome Atlas. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;455:1061–8.
- The Cancer Genome Atlas. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;474:609–15.
- The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61–70.
- The Cancer Genome Atlas. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;487:330–7.
- The Cancer Genome Atlas. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;489:519–25.
- The Cancer Genome Atlas. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 2013;368:2059–74.
- The Cancer Genome Atlas. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013;499:43–9.
- The Cancer Genome Atlas. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 2014;507:315–22.
- Lee J, Sohn I, Do IG, Kim KM, Park SH, Park JO, et al. Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. *PLoS ONE* 2014;9:e90133.
- Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;19:185–93.
- Lee JS, Chu IS, Heo J, Calvisi DF, Sun Z, Roskams T, et al. Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology* 2004;40:667–76.
- Lee JS, Chu IS, Mikaelyan A, Calvisi DF, Heo J, Reddy JK, et al. Application of comparative functional genomics to identify best-fit mouse models to study human cancer. *Nat Genet* 2004;36:1306–11.
- Lee JS, Grisham JW, Thorgerirsson SS. Comparative functional genomics for identifying models of human cancer. *Carcinogenesis* 2005;26:1013–20.



28. Oh SC, Park YY, Park ES, Lim JY, Kim SM, Kim SB, et al. Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. *Gut* 2012;61:1291–8.
29. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A* 2003; 100:9991–6.
30. Simon R, Lam A, Li M-C, Ngan M, Meneses S, Zhao Y. Analysis of gene expression data using BRB-Array Tools. *Cancer Informatics* 2007;3:11–7.
31. Fukayama M, Hino R, Uozaki H. Epstein-Barr virus and gastric carcinoma: Virus-host interactions leading to carcinoma. *Cancer Sci* 2008; 99:1726–33.
32. van BJ, zur HA, Klein KE, van de Velde CJ, Middeldorp JM, van den Brule AJ, et al. EBV-positive gastric adenocarcinomas: A distinct clinicopathologic entity with a low frequency of lymph node involvement. *J Clin Oncol* 2004;22:664–70.
33. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* 2017;355.
34. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. *Nature* 2011;469:343–9.
35. Medina PP, Nolde M, Slack FJ. OncomiR addiction in an in vivo model of microRNA-21-induced pre-B-cell lymphoma. *Nature* 2010;467:86–90.
36. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2013;41:D955–61.
37. Ajani JA, Bentrem DJ, Besh S, D'Amico TA, Das P, Denlinger C, et al. Gastric cancer, version 2.2013: featured updates to the NCCN Guidelines. *J Natl Compr Canc Netw* 2013;11:531–46.
38. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817–26.
39. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* 2011;378: 1812–23.
40. Camargo MC, Kim WH, Chiaravalli AM, Kim KM, Corvalan AH, Matsuo K, et al. Improved survival of gastric cancer with tumour Epstein-Barr virus positivity: an international pooled analysis. *Gut* 2014;63:236–43.
41. Wu MS, Shun CT, Wu CC, Hsu TY, Lin MT, Chang MC, et al. Epstein-Barr virus-associated gastric carcinomas: relation to *H. pylori* infection and genetic alterations. *Gastroenterology* 2000;118:1031–8.
42. Chowdhury UR, Samant RS, Fodstad O, Shevde LA. Emerging role of nuclear protein 1 (NUPR1) in cancer biology. *Cancer Metastasis Rev* 2009;28:225–32.
43. Neira JL, Bintz J, Arruebo M, Rizzuti B, Bonacci T, Vega S, et al. Identification of a drug targeting an intrinsically disordered protein involved in pancreatic adenocarcinoma. *Sci Rep* 2017;7:39732.
44. Vincent AJ, Ren S, Harris LG, Devine DJ, Samant RS, Fodstad O, et al. Cytoplasmic translocation of p21 mediates NUPR1-induced chemoresistance: NUPR1 and p21 in chemoresistance. *FEBS Lett* 2012;586: 3429–34.