Cancer Research

Distinct Subtypes of Gastric Cancer Defined by Molecular Characterization Include Novel Mutational Signatures with Prognostic Capability

Xiangchun Li[1,2], William K.K. Wu[1,3], Rui Xing[4], Sunny H. Wong[1], Yuexin Liu[5], Xiaodong Fang[2], Yanlin Zhang[6], Mengyao Wang[2], Jiaqian Wang[2], Lin Li[2], Yong Zhou[2], Senwei Tang[1], Shaoliang Peng[7], Kunlong Qiu[2], Longyun Chen[2], Kexin Chen[5], Huanming Yang[2], Wei Zhang[8], Matthew T.V. Chan[3], Youyong Lu[4], Joseph J.Y. Sung[1], and Jun Yu[1]

Abstract

Gastric cancer is not a single disease, and its subtype classification is still evolving. Next-generation sequencing studies have identified novel genetic drivers of gastric cancer, but their use as molecular classifiers or prognostic markers of disease outcome has yet to be established. In this study, we integrated somatic mutational profiles and clinicopathologic information from 544 gastric cancer patients from previous genomic studies to identify significantly mutated genes (SMG) with prognostic relevance. Gastric cancer patients were classified into regular (86.8%) and hypermutated (13.2%) subtypes based on mutation burden. Notably, TpCpW mutations occurred significantly more frequently in regular, but not hypermutated, gastric cancers, where they were associated with APOBEC expression. In the former group, six previously unreported (*XIRP2*, *NBEA*, *COL14A1*, *CNBD1*, *ITGAV*, and *AKAP6*) and 12 recurrent mutated genes exhibited high mutation prevalence ($\geq$3.0%) and an unexpectedly higher incidence of nonsynonymous mutations. We also identified two molecular subtypes of regular-mutated gastric cancer that were associated with distinct prognostic outcomes, independently of disease staging, as confirmed in a distinct patient cohort by targeted capture sequencing. Finally, in diffuse-type gastric cancer, *CDH1* mutation was found to be associated with shortened patient survival, independently of disease staging. Overall, our work identified previously unreported SMGs and a mutation signature predictive of patient survival in newly classified subtypes of gastric cancer, offering opportunities to stratify patients into optimal treatment plans based on molecular subtyping. *Cancer Res; 76(7); 1724–32. ©2016 AACR.*

Introduction

Despite a decrease in both incidence and mortality owing to progresses in *Helicobacter pylori* eradication and cancer screening, gastric cancer remains the seventh most common cancer and the third leading cause of cancer-related death worldwide with a 5-year survival rate of 29.6% (1). According to Lauren classification, gastric cancer can be divided into two distinct subtypes (i.e., intestinal and diffuse type) with substantial differences in histology and pathogenesis (2). The intestinal type is featured by well-differentiated, glandular neoplastic cells that are structurally analogous to intestinal cells and more likely to occur in an area with a high risk of gastric cancer (2), whereas the diffuse type is characterized by poorly differentiated neoplastic cells that morphologically resemble signet ring cells and exhibit deeper invasion and infiltration of the whole stomach wall (3–4).

Cancer is a genetic disease arising from changes in DNA sequences (5–6). Many endogenous and exogenous factors can cause somatic mutations, such as defective DNA repair, infidelity in DNA replication, and mutagenic exposures. The landscape of somatic mutations bears the signatures of mutagenic factors (i.e., mutational processes) that have acted upon the genome (7). For example, a vast majority of somatic mutations in skin and lung cancers are associated with ultraviolet and smoking exposure, respectively. Until recently, with the benefit of vast amount of genome data, researchers began to elucidate mutational signatures in human cancers. Alexandrov and colleagues identified 21 mutational processes of 30 different types of primary cancers, some of which were linked to known factors, including DNA mismatch repair deficiency and APOBEC

[1]Institute of Digestive Disease and Department of Medicine & Therapeutics, State Key Laboratory of Digestive Disease, Li Ka Shing Institute of Health Sciences, CUHK Shenzhen Research Institute, The Chinese University of Hong Kong, Hong Kong, Hong Kong. [2]Beijing Genomics Institute, Shenzhen, Guangdong, P.R. China. [3]Department of Anaesthesia and Intensive Care, The Chinese University of Hong Kong, Hong Kong, Hong Kong. [4]Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing, China. [5]Department of Epidemiology and Biostatistics, Tianjin Medical University Cancer Institute and Hospital, Tianjin, P.R. China. [6]Department of Computer Science, City University of Hong Kong, Hong Kong, Hong Kong. [7]School of Computer Science, National University of Defense Technology, Changsha, Hunan, P.R. China. [8]Department of Pathology, University of Texas MD Anderson Cancer Center Informatics Center, Houston, Texas.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Authors: Jun Yu, The Chinese University of Hong Kong, Rm707A, Li Ka Shing Medical Sciences Building, Prince of Wales Hosptial, Shatin, N.T., Hong Kong, Hong Kong. Phone: 852-3763-6099; Fax: 852-2144-5330; E-mail: junyu@cuhk.edu.hk; and William K.K. Wu, E-mail: wukakei@cuhk.edu.hk

doi: 10.1158/0008-5472.CAN-15-2443

overactivity (8). Several independent studies also reported that APOBEC-mediated mutagenesis is widespread across human cancers, and APOBEC3B is responsible for clustered mutation in cancer genomes, as well as *PIK3CA* hotspot mutations (9–11).

Next-generation sequencing has become a powerful tool to elucidate potential driver genetic aberrations underlying cancer development. In gastric cancer, Wang and colleagues and Zang and colleagues reported frequent somatic mutations in chromatin remodeling and cell adhesion genes (e.g., *ARID1A*, *MLL*, *MLL3*, and *FAT4*), respectively (12, 13). Comprehensive genomic characterizations conducted by The Cancer Genome Atlas (TCGA; ref. 14) and Wang and colleagues (15) further revealed genetic alterations underlying gastric carcinogenesis with multi-omics profiling. Together with the study conducted by Kakiuchi and colleagues (16), these three studies highlighted a novel driver gene *RHOA* in diffuse-type gastric cancer with varying mutational prevalence. Through whole-exome deep sequencing, we reported that high clonality is correlated with poorer prognosis in gastric cancer in the Chinese population (17). These studies also implicated that gastric cancer with mutator phenotype has a higher mutational burden and is more likely to exhibit microsatellite instability. Nevertheless, our knowledge of cancer driver genes, especially those mutated at low or intermediate frequencies, in gastric cancer is far from complete because of limited sample size in these studies (18). The mutational processes underlying gastric cancers with or without mutator phenotype are still obscure and have not been investigated. Moreover, gastric cancer is genetically and clinically heterogeneous. Thus, the lack of systematic clinical correlation analyses impedes translation of these findings into clinical benefits. Therefore, deeper insights into gastric carcinogenesis, particularly related to clinical properties, are needed.

The purposes of this study are to characterize mutational processes operative in gastric cancer and to identify previously unreported significantly mutated genes (SMG) and prognosticators for patients with gastric cancer.

Materials and Methods

Genome data

All somatic mutations, including single-nucleotide substitution and short insertion/deletion, were collected from recent publications (Supplementary Table S1), representing five geographically different studied cohorts and annotated by ANNOVAR (19). All somatic mutations were examined in a panel of 442 sequenced normal samples. Mutations present in this panel were removed. Clinical data were also acquired from these publications (Supplementary Dataset S1). However, only TCGA and Tianjin cohorts have the follow-up and vital status of patients.

Clustering based on mutation loads

We applied the optimal k-means clustering via dynamic programming to the number of somatic mutations identified in each case (20). It can guarantee that the within-cluster distance for each cluster is always minimal. Given that the exact cluster number is unknown, we ran this algorithm with different cluster numbers ranging from 1 to 9. The optimal cluster number was selected on the basis of Bayesian information criterion. Finally, an exact cluster number of 6 was selected. Through visual inspection, we divided 544 cases into regular-mutated and hypermutated groups for follow-up analyses. The optimal k-means clustering was performed with *R* package Ckmeans.1d.dp (20).

Identifying genes overrepresented in the hypermutated gastric cancer

To identify genes overrepresented in the hypermutated group, we need to take into account different background mutation rates between the hypermutated and regular-mutated groups, as well as different background mutation rates among mutational categories. Herein, we used goodness-of-fit test with Poisson distribution to address this issue. The background mutation rate for the ith category in the regular- and hypermutated groups are α_i and β_i, respectively. For a given gene, the number of nucleotides in the ith category is n_i. The expected number of mutations given α_i, β_i, and n_i across all categories ϕ can be estimated by Poisson distribution in regular- and hypermutated gastric cancers, i.e., $\lambda_\alpha = \sum_\phi \alpha_i * n_i$ and $\lambda_\beta = \sum_\phi \beta_i * n_i$, respectively (21). The α_i and β_i can be obtained after running MutSigCV, with π_α and π_β as the numbers of observed mutations in the regular- and hypermutated gastric cancers, respectively. To test whether a gene is significantly overrepresented in the hypermutated group, we examined the difference between π_β/π_α and $\lambda_\beta/\lambda_\alpha$. If π_β/π_α is significantly greater than $\lambda_\beta/\lambda_\alpha$, this gene was considered to be significantly overrepresented in the hypermutated group.

Identification of significantly mutated genes

We identified significantly mutated genes with three algorithms using MutSigCV, MutSigCL, and MutSigFN. MutSigCV (21) quantifies significance of nonsilent mutations in a gene with background mutation rate estimated by silent mutations, with other confounding covariates taken into account. MutSigCL and MutSigFN measure the significance of clustered mutations and the functional impact of mutation, respectively (18). In MutSigFN analysis, we separately used CADD and Polyphen2 scores available from dbNSFP database (22) to measure the functional impact of somatic mutations. For efficient computation, we carried out a 2-step permutation. We performed 999 times in the first step to define candidate SMGs (i.e., $P < 0.05$), followed by extensive permutation with 1,000,000 times. We then combined *P* values obtained from the first and second steps. *P* values were then FDR-corrected (*q* values) using the method of Benjamini and Hochberg. For the final analysis of SMGs, we applied additional filtering steps to eliminate possible false positives that may result from the batch effect via combining somatic mutations from different studies. In the regular-mutated group, a gene was considered to be a SMG if it satisfied these conditions: (i) statistically significant (q value < 0.1) by at least one of MutSig algorithms; (ii) expressed in the TCGA pan-cancer dataset (23), human cancer cell lines (24), and/or reported in previous studies (18,23,25,26); (iii) mutated in at least 3 of 5 cohorts; and (iv) mutational prevalence comparable among different cohorts (if mutated). In the hypermutated group, we employed similar but more stringent criteria to select SMGs, including statistically significant by at least two MutSig algorithms and mutated in at least 2 of the 3 available cohorts (i.e., Hong Kong, TCGA, and Tianjin China). This produced a final list of 66 SMGs in the hypermutated group.

Molecular typing

We applied nonnegative matrix factorization (NMF; refs. 27–29) to perform molecular subtyping. A binary matrix *A* describing mutations of SMGs (rows) across cancer samples (columns) was constructed. Specifically, $a(i, j) = 1$ if gene *i* was mutated in sample *j*, otherwise $a(i, j) = 0$; then, *A* was factorized into two nonnegative matrices *W* and *H* (i.e., $A \approx WH$). Matrix *H* was used to group samples into clusters. Optimal number of

clusters was selected on the basis of cophenetic coefficient and dispersion value (29–30).

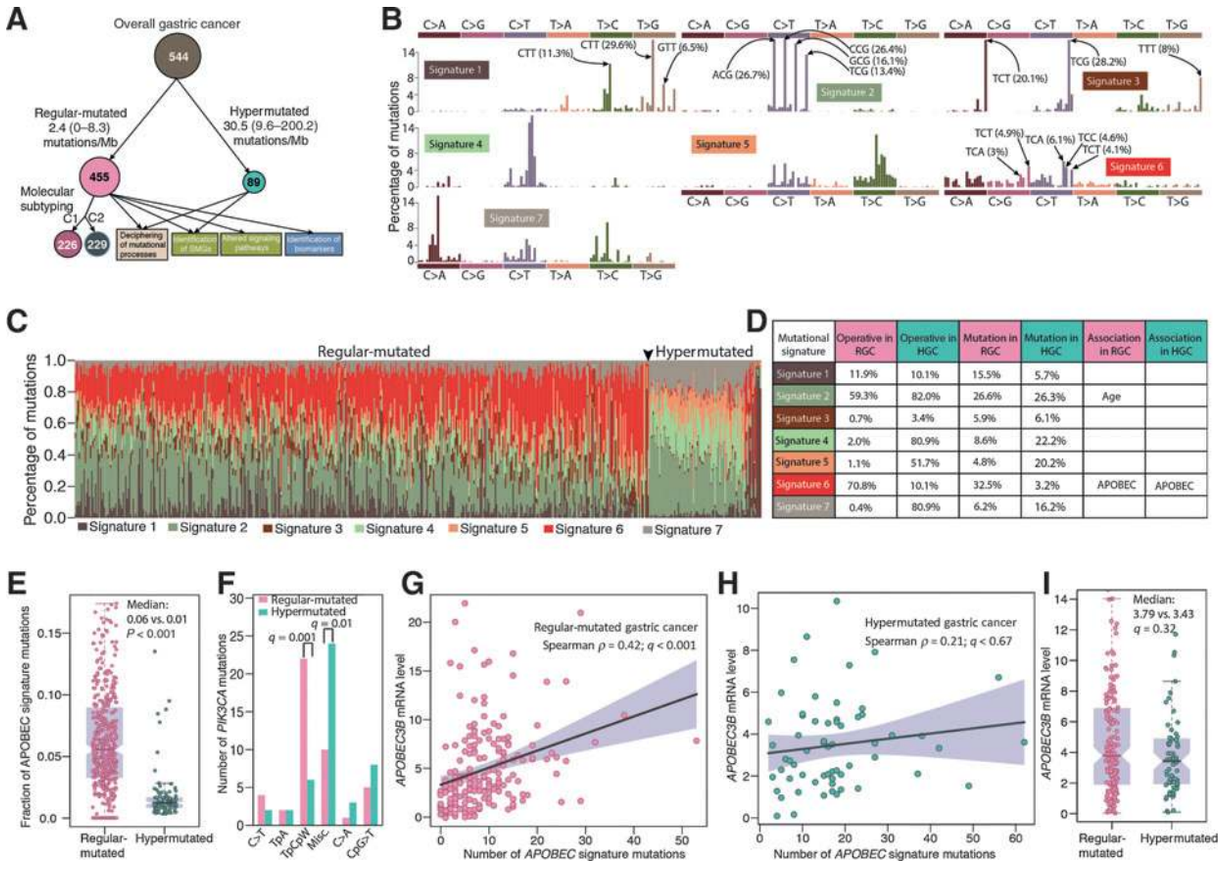
Targeted capture sequencing

Genomic DNA from gastric cancers and lymphocytes was fragmented and hybridized to commercially available capture arrays for enrichment. All samples were collected from patients diagnosed with primary gastric cancer with long-term survival data (15–60 months). Data were analyzed using a bioinformatic pipeline as previously described (31). All patients had given written informed consent, and the study protocol was approved by the clinical ethics committee of the Peking University Cancer Hospital & Institute (Beijing, China).

Results

Classification of gastric cancer based on mutation loads

Somatic mutational profiles of 544 gastric cancer patients from previous genomic studies were aggregated. These samples exhibited large variation in mutation density, ranging from 0 to 200.2 mutations per megabase pair (Mb). To avoid undue effects of samples with high-burden mutation on subsequent analyses (32–33), unsupervised clustering of these 544 gastric cancer samples was performed according to their number of somatic mutations. Two distinct clusters with varying mutation burdens were identified (Fig. 1A and Supplementary Fig. S1A), which were thereafter referred to as regular- (n = 455; 2.4 mutations/Mb; range, 0–8.3) and hyper-mutated (n = 89; 20.5 mutations/Mb; range, 9.6–200.2) gastric cancer, the latter of which showed marked overrepresentation of samples with microsatellite instability (Fisher exact test; OR = 1,012.4; P < 0.001). Although several genes (*BRCA2*, *FANCM*, *PRKDC*, *MSH3*, etc.) that are involved in maintaining genomic integrity were frequently mutated in the hypermutated group, these genes were not significantly enriched when the heterogeneity of background mutation rates was considered (Supplementary Dataset S2).



Mutational signature	Operative in RGC	Operative in HGC	Mutation in RGC	Mutation in HGC	Association in RGC	Association in HGC
Signature 1	11.9%	10.1%	15.5%	5.7%		
Signature 2	59.3%	82.0%	26.6%	26.3%	Age	
Signature 3	0.7%	3.4%	5.9%	6.1%		
Signature 4	2.0%	80.9%	8.6%	22.2%		
Signature 5	1.1%	51.7%	4.8%	20.2%		
Signature 6	70.8%	10.1%	32.5%	3.2%	APOBEC	APOBEC
Signature 7	0.4%	80.9%	6.2%	16.2%		

Figure 1.
Mutational signatures of human gastric cancer. A, mutation burden stratified gastric cancers into the regular- and hypermutated types. Regular-mutated could be further classified into two subgroups (i.e., C1 and C2) based on mutation patterns. B, seven mutational signatures (i.e., signatures 1–7), indicative of distinct underlying mutational processes, were derived from 544 gastric cancer genomes. C, mutational exposures (number of mutations) were attributed to each mutation signature. There were significantly more mutations attributable to signature 6 (i.e., APOBEC signature) in regular-mutated as compared with hypermutated gastric cancer. D, prevalence and proportion of mutations associated with each mutational signature are also shown. A signature is considered to be operative in a tumor if it contributed to more than 100 SNVs or more than 25% of all SNVs in that sample. RGC, regular-mutated gastric cancer; HGC, hypermutated gastric cancer. E, proportion of APOBEC signature mutations. F, mutational patterns of *PIK3CA* in regular- and hypermutated gastric cancer. G and H, relationships between *APOBEC3B* mRNA levels and number of APOBEC signature mutations in regular- and hypermutated gastric cancer. I, mRNA expression level of *APOBEC3B* in regular- and hypermutated gastric cancer.

Mutation signatures of gastric cancer in relation to clinicopathologic features and APOBEC3B expression

To gain further insights into the mutational processes operative in regular- and hyper-mutated gastric cancer, we delineated their mutation signatures using the computational framework proposed by Alexandrov and colleagues (34). Seven mutation signatures were extracted from all 544 gastric cancer samples, namely signatures 1 to 7, each of which contributed to different proportion of mutations in the regular- and hypermutated groups (Supplementary Fig. S1B and S1C). In kernel principal component analysis, we found that mutation signatures corresponding to different studies were intermingled, whereas the regular- and hypermutated groups were clearly distinguishable (Supplementary Fig. S1C), suggesting a minimal impact of batch effect on mutation signatures. General linear regression was performed to analyze the relationship between mutation signatures and clinicopathologic features. We observed that age at diagnosis was associated with signature 2 (Fig. 1C), which is concordant with a previous report (8). Of interest, signature 6, which was dominated by mutations at TpCpW DNA motif (where W=A or T; mutated nucleotide underlined), accounted for 32.5% mutations in regular- but only 3.2% in hyper-mutated gastric cancer. In this regard, regular-mutated gastric cancer harbored 6 times more mutations at TpCpW DNA motif than the hypermutated group (Wilcoxon test; median, 0.06 vs. 0.01; $P < 0.001$; Fig. 1D). This mutation pattern is also known as APOBEC signature, which is widespread across multiple human cancer types (10). The dominance of TpCpW mutations in regular-mutated gastric cancer could be exemplified by the mutation pattern of *PIK3CA* (Fig. 1E), which is a driver gene in gastric cancer. The RNA-editing enzyme APOBEC3B has been reported to contribute to mutations at TpCpW motif in cancer genome (10). To interrogate the contribution of APOBEC3B to mutations in gastric cancer, we analyzed the relationship between its expression levels and number of TpCpW mutations. We observed that APOBEC3B mRNA levels were positively correlated with the number of TpCpW in regular- (Spearman $\rho = 0.42$; $P < 0.001$; Fig. 1F) but not hypermutated gastric cancer (Spearman $\rho = 0.21$; $q = 0.67$; Fig. 1F) despite similar expression levels in both groups ($q = 0.32$; Fig. 1F), signifying their difference in mutagenesis. As regular- and hypermutated gastric cancers were characterized by distinct molecular features, these two subgroups were analyzed separately in subsequent analyses.

Significantly mutated genes in gastric cancer

A total number of 39,891 and 101,189 nonsilent somatic mutations, including missense, nonsense, splice-site, and frameshift mutations, were detected in 455 and 89 cases of regular- and hypermutated gastric cancer, respectively. To identify SMGs that are causally linked to tumorigenesis, we used three algorithms, namely MutSigCV (21), MutSigCL, and MutSigFN, to identify genes whose mutations were positively selected, clustered in hotspots or of functional consequences. In regular-mutated gastric cancer, 31 SMGs were identified (Fig. 2A; Supplementary Dataset S2), among which 12 reported [*TP53* (48.4%), *ARID1A* (13.8%), *CDH1* (11.6%), *PIK3CA* (8.4%), *APC* (6.8%), *RHOA*

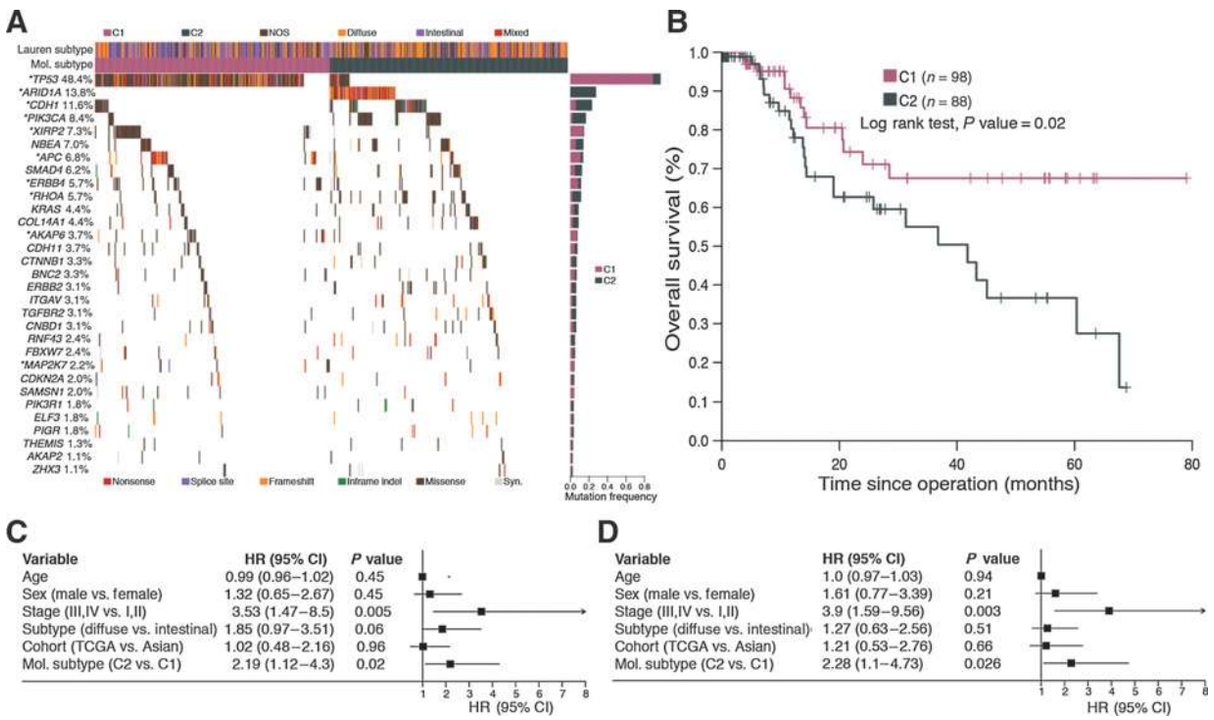


Figure 2.
Mutational landscape and prognostic significance of molecular subtyping in regular-mutated gastric cancer. A, mutational landscape of SMGs ordered by overall mutation frequencies in regular-mutated gastric cancer samples. Molecular classification into C1 and C2 based on mutation status of 31 SMGs was performed using NMF. Asterisks indicate SMGs with preferential mutations in either molecular subtype. B, Kaplan–Meier survival curves displaying survival outcomes of C1- and C2-type regular-mutated gastric cancer. Univariate (C) and multivariate (D) Cox regression analyses for age, sex, TNM staging, Lauren classification, and molecular subtypes. HR, 95% CI, and *P* values are displayed.

(5.7%), *SMAD4* (6.2%), *ERBB4* (5.7%), *KRAS* (4.4%), *ERBB2* (3.2%), and *CTNNB1* (3.1%)] and 6 previously unreported [*XIPR2* (7.3%), *NBEA* (7.0%) *COL14A1* (4.4%), *AKAP6* (3.7%), *CNBD1* (3.1%), and *ITGAV* (3.1%)] SMGs exhibited moderate to high mutation prevalence (≥3.0%). In particular, four genes, namely *TP53*, *CDH1*, *SMAD4*, and *CTNNB1*, were ranked as significantly mutated by all 3 algorithms. The 6 previously unreported SMGs were frequently mutated across multiple human cancer types (Supplementary Dataset S3; refs. 35, 36). *ITGAV*, upstream regulator of PI3K signaling pathways and located in 2q32.1, was significantly deleted in Hong Kong and TCGA cohorts (14, 15). *NBEA* (located in 13q12.11) and *CNBD1* (located in 8q22) were focally deleted in TCGA cohort, with the latter significantly mutated in 1% of previous pan-cancer dataset, but not in individual cancer types (14, 18). *AKAP6*, protein kinase anchoring factor, was reported to be significantly mutated in esophageal adenocarcinoma (37). *COL14A1* was reported to be frequently mutated and downregulated in other cancer types (38, 39). In the hypermutated group, we first used 3 MutSig algorithms to define SMGs by evaluating only single-nucleotide substitutions, giving rise to 4 SMGs, including *ARID1A* and *TP53* (Supplementary Dataset S2). By including short insertions/deletions, the final list of SMGs was expanded to 66, including *MLL2*, *RNF43*, *B2M*, *ACVR2A*, and *RNF43* (Supplementary Dataset S2).

SMG mutation patterns predictive of survival in regular-mutated gastric cancer

To determine if SMG mutation status could be used for molecular typing, NMF-based unsupervised clustering on binary mutation matrix of SMGs was performed, which yielded two subgroups (hereafter referred to as C1 and C2; Supplementary Fig. S1D). The first subgroup C1 was enriched with mutations in *TP53* (89.9%; $q < 0.001$), *XIRP2* (13.7%; $q < 0.001$), *APC* (11.1%; $q = 0.01$), *ERBB4* (8.4%; $q = 0.047$), and *AKAP6* (6.6%; $q = 0.004$), whereas the second subgroup C2 was overrepresented by mutations in *ARID1A* (27.5%; $q < 0.001$), *CDH1* (17.5%; $q < 0.001$), *PIK3CA* (14.4%; $q < 0.001$), and *RHOA* (9.2%; $q = 0.007$). C2 was also featured with more gastric body–located [OR = 1.71; 95% confidence interval (CI), 1.0–2.93; $P = 0.05$] and diffuse-type (OR = 2.45; 95% CI, 1.49–4.06; $P < 0.001$) gastric cancer. In comparison with TCGA molecular subtypes, we found that C1 was more enriched with chromosome instability (CIN) subtype (Fisher exact test, $P < 0.001$), whereas C2 had even distribution of CIN and genome stable (GS) subtypes (45 CINs vs. 42 GSs). Importantly, C1 was associated with a significantly better prognostic outcome ($P = 0.02$; Fig. 2B). Univariate analysis indicated that advanced tumor–node–metastasis (TNM) staging (stage III/IV) and C2 subtype but not other factors, including tumor location (antrum, body, or cardia), were significantly associated with poorer survival outcome (Fig. 2C). Multivariate analysis revealed that the prognostic significance of C1/2 was independent of age, sex, Lauren classification, TNM staging, and studied cohorts (Fig. 2D). Nevertheless, subgroup analysis stratified by Lauren classification indicated that the prognostic significance of C1/2 was more apparent in diffuse-type than intestinal-type gastric cancer (Supplementary Fig. S2A).

C1/2 signature validation in an independent cohort

We observed that 8 (*TP53*, *ARID1A*, *CDH1*, *PIK3CA*, *XIRP2*, *APC*, *ERBB2*, and *RHOA*) of the abovementioned 31 SMGs exhibited differential mutation frequencies with mutational prevalence >5%. Importantly, we found that these 8 SMGs could achieve comparable prediction accuracy in discriminating C1 from C2 as by the 31 SMGs (Supplementary Fig. S2B). By performing targeted capture sequencing in an independent cohort of gastric cancer patients and classifying these patients into C1 and C2 subgroups based on these 8 SMGs, we found that C1/2 remained a significant independent prognostic marker in gastric cancer (Supplementary Fig. S2C).

Mutation distribution in cancer signaling pathways

By mapping SMGs and other well-known genes to cell signaling pathways, we observed that several pathways were frequently altered in regular-mutated gastric cancer, including genotoxic/oncogenic stress response (57.6%), histone modification/chromatin remodeling (26.6%), and growth factor receptor signaling (22.4%) and Wnt signaling (24.4%; Fig. 4).

Recurrent point mutations in gastric cancer

Recurrent point mutations could help to identify cancer driver genes and druggable targets (40). Herein, we depicted the panorama of recurrent point mutations in gastric cancer (Supplementary Dataset S2). Of the 31 SMGs in regular-mutated gastric cancer, twenty of them were significantly enriched for recurrent point mutations, including *TP53*, *PIK3CA*, *CDH1*, *KRAS*, *RHOA*, *ERBB2*, and *ERBB4*. Nevertheless, only three SMGs, namely *TP53*, *PIK3CA*, and *KRAS*, were identified in hypermutated gastric cancer. Notably, two recurrent mutations (p.E542K and p.E545K) in *PIK3CA* are significantly overrepresented in the regular-mutated gastric cancer (OR = 4.26; Fisher exact test, $P = 0.004$), which were exclusively located at TpCpW motif (Fig. 3A). Conversely, another *PIK3CA* hotspot mutation (p.H1047R) was significantly enriched in the hypermutated group (OR = 4.46; Fisher exact test, $P = 0.02$), which occurred in ApTpG rather than TpCpW motif (Fig. 3B). For other SMGs, recurrent point mutations of *CDH1* and *RHOA* were only observed in regular-mutated but not hypermutated gastric cancers.

CDH1 mutation as a prognostic factor in diffuse-type gastric cancer

To discover single-gene prognosticators, we analyzed the association between nine SMGs (mutated at >5%) and survival data. We found that *CDH1* and *SMAD4* mutations were significantly associated with shortened survival in patients with gastric cancer as revealed by Kaplan–Meier analysis. In subgroup analysis stratified by Lauren subtype, *CDH1* mutation(s) was a significant prognostic factor in diffuse-type but not intestinal-type gastric cancer independent of TNM staging (Supplementary Fig. S3A). In contrast, *SMAD4* mutation was associated with shortened survival only in the intestinal-type gastric cancer. We then verified the prognostic significance of *CDH1* and *SMAD4* in an independent cohort (17), in which these two genes were sequenced on an orthogonal-sequencing platform. We observed that *CDH1* mutations remained an independent factor for poor survival in the diffuse-type gastric cancer (Supplementary Fig. S3B). However, the significance of *SMAD4* mutations could not be verified. The prognostic significance of *CDH1* in the diffuse-type gastric cancer by combined analysis of two cohorts was shown in Fig. 5.

Discussion

In this study, we performed systematic analyses on 544 gastric cancers and correlated genetic events with clinicopathologic

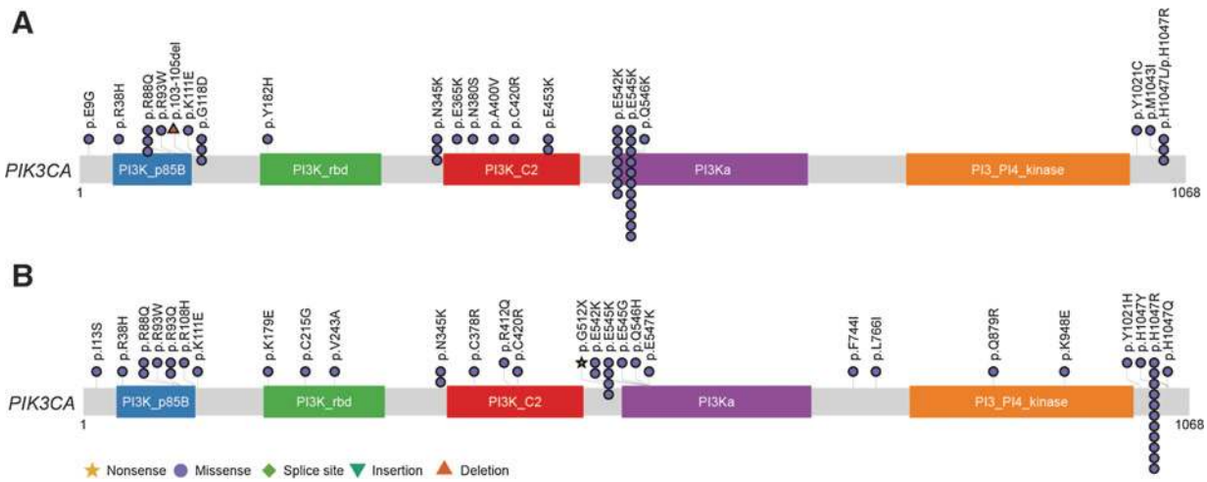


Figure 3.
Mutation plots of *PIK3CA* in the regular- (A) and hypermutated (B) gastric cancers. The distribution of different classes of mutations (different shapes) and functional domains of *PIK3CA* is shown.

features. Major findings derived from our study include (i) there are ubiquitous and specific mutational processes underlying the pathogenesis of different subtypes of gastric cancer with varying mutation burdens; (ii) several previously unreported SMGs that are mutated at intermediate or low prevalence were identified; (iii) regular-mutated gastric cancer can be further stratified into

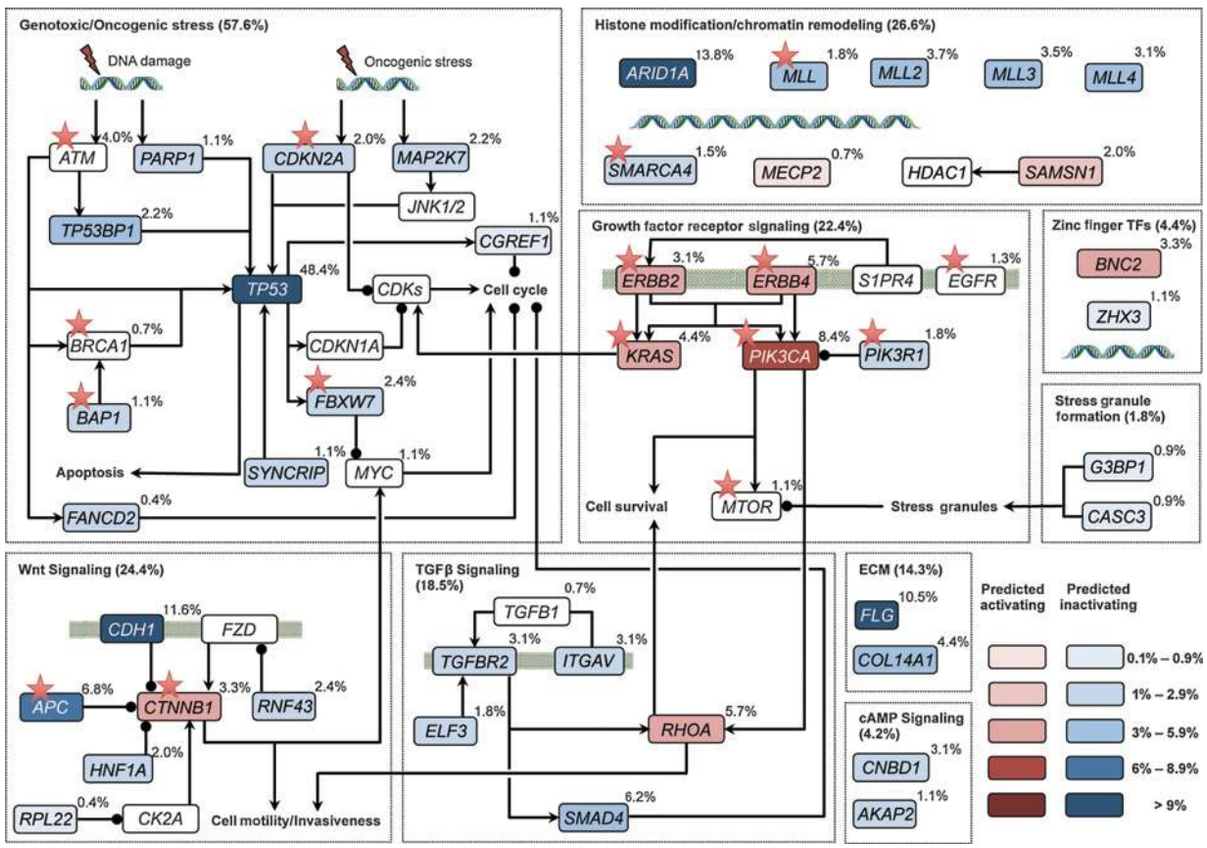


Figure 4.
Altered signaling pathways in regular-mutated gastric cancer. Key pathways and inferred functions are summarized. Red and blue colors denote SMGs with activating and inactivating mutations, respectively, whereas genes in white are not identified as SMGs but known to have tumorigenic roles. Potential druggable targets are marked with stars. TF, transcription factor.

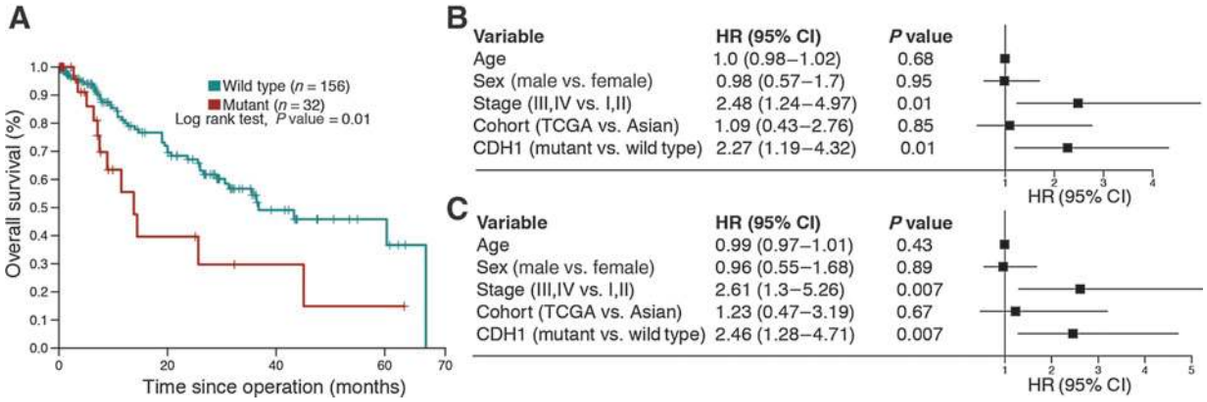


Figure 5.
Prognostic significance of *CDH1* mutation(s) in a combined cohort of diffuse-type gastric cancer. A, Kaplan–Meier survival curves showing survival outcomes for two subgroups stratified by *CDH1* mutation status. B and C, univariate (B) and multivariate (C) Cox regression analyses for *CDH1* mutation in relation to age, sex, and TNM staging.

two subtypes (i.e., C1 and C2) with distinct clinical outcomes; and (iv) *CDH1* mutation is an independent prognostic factor for poorer survival in patients with diffuse-type gastric cancer.

In line with the previous observation (8,10), we revealed that APOBEC-mediated mutation pattern is ubiquitous in gastric cancer. Although hypermutated gastric cancer that showed defects in DNA damage repair is more likely to generate single-strand DNA breakage, which is an ideal substrate for APOBEC family of cytidine deaminases (41); this mutational process contributed to the majority of somatic mutations in regular-mutated but not hypermutated gastric cancer. Besides, we observed that *PIK3CA* mutations (p.E542K and p.E545K) of regular-mutated gastric cancer are more likely to occur in TpCpW motif in comparison with the hypermutated group. These discrepancies have not been reported in gastric cancer. Concordant with our findings, a recent study on esophageal squamous cell carcinoma reported that these two *PIK3CA* hotspot mutations were overrepresented in APOBEC signature tumors (42). In addition, a supervised random forest clustering (43, 44), with mutational exposures as input, ranked signature 6 (APOBEC signature) as the most important predictive feature in distinguishing the regular-mutated from the hypermutated gastric cancer with high prediction accuracy (Supplementary Fig. S1C and S1E). Taken together, these findings suggest that APOBEC-mediated mutagenic activity is operative to greater extent in regular-mutated gastric cancer, and alternative underlying mutagenic factors, presumably defective DNA proofreading and repair, have substantially greater impact on the hypermutated gastric cancer.

According to a previous study, nearly 600 samples per cancer type are required to achieve a complete catalog of cancer genes (18). Our study encompassed the largest number of gastric cancer samples available ($n = 544$) by combining data from previous whole-genome and whole-exome sequencing studies on gastric cancer (14–17, 45). Thus, we are able to identify SMGs that are mutated less frequently with a higher statistical power (18). For instance, we found that *FBXW7*, widely reported as cancer driver in multiple human malignancies but not gastric cancer (23), was significantly mutated at 2.4% in our gastric cancer study; besides, there are other previously unreported SMGs (e.g., *XIRP2*, *NBEA*, *COL14A1*, *CNBD1*, *AKAP6*, and *ITGAV*) that are reported for the first time in gastric cancer, underscoring the importance of increasing sample size to identify potential cancer genes. However, the statistic power to identify cancer genes mutated at 2% of samples is merely 12% in regular-mutated gastric cancer (18); thus, more rare SMGs are awaited to be discovered.

Batch effect is a common phenomenon in meta-analysis. To exclude artifacts in our SMG list, we employed stringent criteria to filter out possible false positives, including genes showing bias towards specific cohorts. However, cancer development is closely related to environmental exposures, which can provide unique selective pressure over specific cancer genes. As different cohorts may have their unique exposures, SMGs displaying cohort-specific features could be discarded in this study, leading to underestimation of the number of SMG. Another limitation of our study is the lack of functional validation of the identified SMGs. Until such data are available, whether these SMGs are indeed drivers in gastric tumorigenesis remains uncertain.

A major clinically relevant finding of this study is the use of 8 SMGs to classify regular-mutated gastric cancer into two prognostically distinct subgroups, namely C1 and C2 (Fig. 2 and Supplementary Fig. S1C). The prognostic significance of C1/2 was validated in an additional cohort and independent of age, sex, Lauren classification, TNM staging, and studied cohorts. Together with other genome-based molecular subtyping studies (46–48), this finding highlighted the importance of molecular subtyping based on genome data in clinical utility. However, the lack of clinical and surgical data (residual tumor, number of removed lymph nodes, and more refined clinical tumor staging) may be a potential limitation of this study.

Another key finding that emerged from our study is the identification of *CDH1* mutation(s) as an independent prognostic marker for poor prognosis in diffuse-type gastric cancer. Somatic mutations in *CDH1* are frequently reported for sporadic gastric cancer with predilection towards the diffuse type (15, 16), which is associated with dismal survival (3). Interestingly, we revealed that *CDH1* mutation(s) allows further stratification of diffuse-type gastric cancer into distinct subgroups with significantly different survival outcomes. In summary, our findings may be leveraged to speed up interpretation of cancer genome data that may foreshadow clinical outcomes and thus potentially help to guide gastric cancer intervention.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: X. Li, W.K.K. Wu, L. Li, H. Yang, W. Zhang, Y. Lu, J.J.Y. Sung, J. Yu

Development of methodology: X. Li

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): X. Li

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): X. Li, W.K.K. Wu, S.H. Wong, Y. Liu, X. Fang, Y. Zhang, M. Wang, L. Li, Y. Zhou, S. Tang, S. Peng, L. Chen, W. Zhang

Writing, review, and/or revision of the manuscript: X. Li, W.K.K. Wu, S.H. Wong, Y. Liu, W. Zhang, M.T.V. Chan, J. Yu

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): X. Li, R. Xing, M. Wang, K. Qiu, K. Chen, Y. Lu, J.J.Y. Sung

Study supervision: W.K.K. Wu, W. Zhang, J. Yu

Grant Support

This work was supported by research funds from RGC GRF Hong Kong (766613), Shenzhen Virtual University Park Support Scheme to CUHK Shenzhen Research Institute.

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 14, 2015; revised October 23, 2015; accepted November 10, 2015; published OnlineFirst February 8, 2016.

References

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. GLOBOCAN 2012: estimated cancer incidence, mortality and prevalence worldwide in 2012, Section of cancer information. IARC, Lyon, France. v1.0
2. Lauren P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. Acta Pathol Microbiol Scand 1965; 64:31–49.
3. Yamashita K, Sakuramoto S, Katada N, Futawatari N, Moriya H, Hirai K, et al. Diffuse type advanced gastric cancer showing dismal prognosis is characterized by deeper invasion and emerging peritoneal cancer cell: The latest comparative study to intestinal advanced gastric cancer. Hepatogastroenterology 2009;56:276–81.
4. Marrelli D, Roviello F, De Manzoni G, Morgagni P, Di Leo A, Saragoni L, et al. Different patterns of recurrence in gastric cancer depending on Lauren's histological type: longitudinal study. World J Surg 2002;26: 1160–5.
5. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719–24.
6. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.
7. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet 2014;15: 585–98.
8. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013; 500:415–21.
9. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet 2013;45:977–83.
10. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet 2013;45:970–6.
11. Henderson S, Chakravarthy A, Su X, Boshoff C, Fenton TR. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. Cell Rep 2014;7: 1833–41.
12. 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.
13. Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet 2012; 44:570–4.
14. Bass AJ, Thorsson V, Shmulevich I, Reynolds SM, Miller M, Bernald B, et al. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202–9.
15. Wang K, Yuen ST, Xu J, Lee SP, Yan HH, Shi ST, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat Genet 2014;46:573–82.
16. Kakiuchi M, Nishizawa T, Ueda H, Gotoh K, Tanaka A, Hayashi A, et al. Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. Nat Genet 2014;46:583–7.
17. Chen K, Yang D, Li X, Sun B, Song F, Cao W, et al. Mutational landscape of gastric adenocarcinoma in Chinese : implications for prognosis and therapy. Proc Natl Acad Sci 2015;6:1–6.
18. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495–501.
19. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:1–7.
20. Wang H, Song M. Ckmeans.1d.dp: Optimal *k*-means clustering in one dimension by dynamic programming. R J 2011;3:29–33.
21. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.
22. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 2013;34:2393–402.
23. Kandoth C, Mclellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013; 502:333–9.
24. Klijn C, Durinck S, Stawiski EW, Havery PM, Jiang Z, Liu H, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol 2015;33:306–12.
25. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep 2013;3:2650.
26. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4: 177–83.
27. Brunet J, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 2004;101:4164–9.
28. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics 2005;21: 3970–5.
29. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 2007;23:1495–502.
30. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010;11:367.
31. Yu J, Wu WKK, Li X, He J, Li XX, Ng SS, et al. Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. Gut 2015;64:636–45.
32. Cancer T, Atlas G. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.
33. Bass AJ, Laird PW, Shmulevich I, Thorsson V, Schultz N, Sheth M, et al. Comprehensive Molecular Characterization of Gastric Adenocarcinoma. Nature 2014;513:202–9.
34. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013;3:246–59.

35. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401–4.
36. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:pl1.
37. Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat Genet 2013;45: 478–86.
38. Stirzaker C, Zotenko E, Song JZ, Qu W, Nair SS, Locke WJ, et al. Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. Nat Commun 2015;6: 1–11.
39. Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc Natl Acad Sci USA 2013;110:21083–8.
40. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
41. Smith HC, Bennett RP, Kizilyer A, McDougall WM, Prohaska KM. Functions and regulation of the APOBEC family of proteins. Semin Cell Dev Biol 2012;23:258–68.
42. Zhang L, Zhou Y, Cheng C, Cui H, Cheng L, Kong P, et al. Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. Am J Hum Genet 2015;96:597–611.
43. Breiman L. Random Forests. Mach Lean 2001;45:5–32.
44. Liaw A, Wiener M. Classification and regression by random forest. R News 2002;2:18–22.
45. Wong SS, Kim KM, Ting JC, Yu K, Fu J, Liu S, et al. Genomic landscape and genetic heterogeneity in gastric adenocarcinoma revealed by whole-genome sequencing. Nat Commun 2014;5:5477.
46. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 2014;158:1–16.
47. Gross AM, Orosco RK, Shen JP, Egloff AM, Carter H, Choueiri M, et al. Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss. Nat Genet 2014;46:939–43.
48. Akbani R, Akdemir KC, Aksoy BA, Albert M, Ally A, Amin SB, et al. Genomic classification of cutaneous melanoma. Cell 2015;161:1681–96.