Clonal architectures predict clinical outcome in gastric adenocarcinoma based genomic variation and tumor evolution and heterogeneity in gastric adenocarcinoma

Chenxia Ren

Changzhi Medical College

Cuiling Wu

Changzhi Medical College

Niuniu Wang

Changzhi Medical College Affiliated Heping Hospital

Changhong Lian

Changzhi Medical College Affiliated Heping Hospital

Changqing Yang (✉ Young@czmc.edu.com)

Changzhi Medical College Affiliated Heping Hospital

Research article

Keywords: Tumor immunity, tumor evolution, tumor heterogeneity, gastric cancer

Posted Date: February 26th, 2020

DOI: <https://doi.org/10.21203/rs.2.24649/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Gastric cancer is a highly heterogeneous disease. Due to the lack of effective molecular markers and personalized treatment, the prognosis of patients with gastric cancer is still rather poor. This paper attempts to explore the genomic instability and intratumoral heterogeneity of gastric cancer through bioinformatics analysis.

Methods: According to the RNA-Seq data, copy number variation data and clinical follow-up information data of TCGA, brunet algorithm in NMF was used to identify SNV signature. ABSOLUTE algorithm was used to screen and identify CNV signature. The relationship between clonal state and survival of genes was analyzed with Kaplan-Meier method. The Spearman method was used to evaluate the correlation between clonal/sub-clonal events and TMB and neoantigen.

Results: Mutation analysis of SNV mutation and CNA data in the TCGA gastric cancer dataset divided mutation of TCGA gastric cancer samples into three signatures with significant differences. Furthermore, the clonal/sub-clonal events of SCNA and SNV were analyzed to identify the clonal/sub-clonal status of mutations in STAD samples. It was further found that the number of mutations in TP53, TTN and MUC16 genes was the highest (> 20%) among the samples. The number of CNV (Gain) in MIEN1, GRB7 and PNMT genes was the highest (> 10%) among the samples. A series of early gastric cancer genes, such as TP53, USH2A and GLI3 as well as advanced genes such as CTNNB1, LRP1B and ERBB4 were further identified. Ultimately, the clonal/sub-clonal status of 5 early genes, 12 intermediate genes and 8 advanced genes was significantly correlated with the overall survival rate of patients. In addition, there was a significant correlation between clonal/sub-clonal events and TMB/Neoantigens.

Conclusion: In this paper, the relationship between intratumoral heterogeneity and genomic instability was evaluated based on developmental data of gastric cancer cloning system. A series of molecular features determined by screening can be used as a marker of potential gastric cancer heterogeneity, which is of great significance to the personalized treatment of gastric cancer.

Background

Gastric cancer (GC) is one of the four most common cancers in the world and the second leading cause of cancer-related deaths in the world [1]. The incidence of gastric cancer lacks typical clinical features, which poses great challenges to the early diagnosis of gastric cancer [2]. At the same time, the recurrence rate and metastasis rate of gastric cancer patients are higher, which is an important reason for the short survival time and poor prognosis of gastric cancer patients [3]. Currently, the treatment mode of gastric cancer is combination of surgery, chemotherapy and targeted treatment [4]. Despite some progress in the treatment of gastric cancer in recent years, the prognosis of patients is still poor [5]. Therefore, the research on the etiology and pathogenesis of gastric cancer is an important means to prevent and treat gastric cancer.

The occurrence and development of gastric cancer is the result of a combination of environmental factors and genetic factors [6]. Epidemiological data have confirmed that: *Helicobacter pylori* (HP), high-salt diet, inadequate dietary fiber intake, smoking and alcohol consumption are common pathogenic factors of gastric cancer [7]. Extensive studies have shown that cancer is a genetic disease caused by mutations in cell genome. Cancer cells are produced by the gradual accumulation of genomic mutation, including somatic cell mutations and copy number variations [8, 9]. SNV variation is caused by the replacement of one base in the DNA sequence by another base [10]. Copy number variation refers to the variation of large genomic regions in cancer [11]. SNV mutation and copy number variation play an important role in the occurrence and development of tumors. For example, the SNV variation of single base G/T substitution in HRAS gene can cause changes in the corresponding protein products. The imbalance of the protein's involvement in the process can cause cancer [10]. The copy number variation in the region of tumor suppressor genes, such as BRCA1, CDKN2A and PTEN, can lead to low expression of these genes, which brings selective growth advantage for cancer cells [12, 13]. In gastric cancer, the amplification of copy number of cell growth regulatory gene HER2 can significantly promote the proliferation of gastric cancer cells [14]. Clinical studies have also shown that Trastuzumab combined with chemotherapy can significantly improve the survival rate of gastric cancer patients with HER2 gene amplification [15]. This not only indicates the importance of genomic variation in the occurrence and development of gastric cancer, but also suggests the potential value of targeted genomic mutation in the treatment of gastric cancer.

Genomic mutations are different in different cancer cell samples, which is known as tumor heterogeneity [16]. In the process of tumor occurrence and development, the cloning types of tumor cells will continue to change due to genomic variation and other factors. These different tumor cell clones show significant differences in growth rate, invasive ability and immune features. They constitute the heterogeneity between cells in the tumor. These highly heterogeneous tumor cells lead to differential responses to various treatments [17, 18]. Therefore, an in-depth understanding of the heterogeneity of tumors help reveal the mechanism of tumor formation and evolution, which provides clues and basis for the development of accurate diagnosis, prevention and treatment of tumors. Due to the high degree of spatio-temporal heterogeneity [19], there are still few targeted drugs for gastric cancer. HER2 remains the only therapeutic target with a clear benefit group in the field of gastric cancer. Even if initial anti-HER 2 therapy was effective, drug resistance mostly occurred within one year, preventing sustained benefit from patients [20]. Therefore, in-depth exploration of the heterogeneity and evolutionary mechanism of gastric cancer will help to deepen the understanding of gastric cancer, in order to provide more treatment methods.

In this study, the clone (subclone) composition of each tumor sample was inferred from the group library of somatic mutation (SNV) and copy number variations (SCNV) from the genome data of gastric adenocarcinoma (STAD) from TCGA. Using the tumor clonal phylogenetic data, the relationship between intratumoral heterogeneity and genomic instability was further evaluated. These features will deepen our understanding of gastric cancer and serve as a potential marker of tumor complexity.

Methods

Data download and preprocessing

TCGA GDC API was used to download the latest clinical follow-up information on March 14, 2019 (S1_Table.txt). The following steps of preprocessing were conducted on the clinical sample data of TCGA: 1) samples with no clinical information or OS < 30 days were removed; 2) normal tissue sample data were removed. The mutated data were preprocessed in the following steps: 1) silent and intron mutated sites are removed; 2) hyper-mutated samples were removed. Hyper-mutated samples were defined as samples with more than 11.4 mutations per Mb [21]. The following steps were performed on the CNV data: 1) data with the interval > 500 kb were removed;

2) gencode.v 22 of the GRh38 version was applied to map the CNV interval to the corresponding genes. The statistics information of the preprocessed STAD data set is shown in Table 1.

Table 1
Clinical information of the
preprocessed STAD data set

Event	
Alive	207
Dead	142
Stage_T	
T1	16
T2	74
T3	161
T4	94
Un	4
Stage_N	
N0	103
N1	93
N2	72
N3	71
Un	10
Stage_M	
M0	311
M1	23
Un	15
Stage	
I	46
II	109
III	145
IV	35
Un	14
Grade	

Event	
G1	9
G2	124
G3	207
Un	9
Age	
0 ~ 50	29
50 ~ 60	83
60 ~ 70	103
70 ~ 80	113
80 ~ 100	21
Gender	
Female	124
Male	225
median clonal	48
median subclonal	6

Clone state classification of SNV

First, the ABSOLUTE algorithm was used to calculate the purity, ploidy and absolute DNA copy number of each sample. For SCNV and SSNV, they were classified as clonal events if $p(\text{CCF} > 0.85) > 0.5$. Otherwise, they were classified as sub-clonal events. CCF referred to cancer cell fraction. Specifically, for each mutated site (including CNV and SNV), the numbers of mutated reads, unmutated reads, tumor purity and local CNV could be used to evaluate the probability distribution density of CCF. In the first step, the tumor DNA proportion was calculated first, and then the allele fraction (AF) probability was calculated according to the binomial probability density distribution. In this step, the influence of normal cell components was cleared and the $p(\text{AF})$ was obtained. The second step was to integrate all possible mutation multiplicities (m : 1 to local absolute copy number) and to evaluate CCF's probability by $p(\text{AF})$ [22].

Analysis of mutation characteristics

Non-negative matrix factorization (NMF) is an unsupervised clustering method, which is widely used in the discovery of tumor molecular subtypes based on genomics [23–24]. The mutation characteristics in CRC were further observed, and the NMF method was used to cluster the samples based on SNV data to identify the mutation characteristics of the samples, in which the NMF method selected the standard

"brunet" for 50 iterations. The clustering number k is set to 2 to 10, and the average contour width of the common member matrix is calculated using the R software package NMF [25]. and the minimum member of each subclass is set to 10. The cophentic, dispersion and rss indexes of $K = 2-10$ are evaluated respectively. According to these three indicators, the optimal number of clusters is selected. In order to evaluate the heterogeneity of mutant signature, we calculated the contribution of each mutant signature to each sample, using the 30 known mutant signatures, provided by COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures_v2). We calculated the similarity between mutant signatures and COSMIC mutational signatures [26].

Analysis of genomic variation in cloning and subcloning

The copy number variation of STAD was evaluated by the ABSOLUTE algorithm. Firstly, the CNV obtained by ABSOLUTE was screened, and the CNV interval meeting the following conditions was retained: 1) modal $CN < 2$ (Loss) or modal $CN > 2$ (Gain); 2) CNV interval < 0.5 Mb. Then the SCNA was corresponding to the gene by using the coordinate position of gencode V22, and the relationship between the cloned and subcloned gene and CNV was analyzed.

Analysis of temporal sequence relationship between mutation and tumor evolution

With a sample's clonal events and sub-clonal events, the possible timing of mutations in tumor evolution was constructed. When clonal events and sub-clonal events appear in the same sample, an edge was established between the two to analyze all the samples in the same way, and finally a gene network with a specific direction was obtained. The node of the network represented the gene, and the edge indicated that there was a clonal and sub-clonal relationship between the two genes. Enrichment analysis was performed according to the number of in-edges and out-edges of each node (gene), fisher exact test was used for significance test, and the BH method was used to calculate FDR. For SSNV and SCNV, the nodes (genes) with $FDR < 0.05$, 0.2 and out-edges $>$ in-edges were defined as early genes; similarly, the nodes with $FDR < 0.05$, 0.2 and in-edges $>$ out-edges were defined as late genes; and the genes of other cases were defined as intermediate genes. Because we calculated the genes that produced SCNA based on the CNV interval and gff interval of the chip data, some genes of SCNA might present "false positive", affecting the results of SSNV, we inferred the temporal order of SCNA and SSNV respectively. In order to facilitate the display, we removed some conflicting edges and finally got SCNV pairs and SNV pairs.

Statistic Analysis

The Kaplan-Meier method was used to perform survival curves for the subgroups in each data set, and logrank test was used to determine the statistical significance of the differences, significantly defined as $P < 0.05$, Chisq test was used to test the significance of sample overlap between clinical features and clonal or subclonal events. The Benjamini-Hochberg method converts the P value to FDR. All of the above

analyses were completed under R 3.5.1. If there are no special instructions, * means $p < 1e-5$ in this article, $p < 0.01$, and * means $p < 0.05$.

Results:

Evaluation of samples' purity and ploidy

The purity and ploidy of the STAD samples calculated by the above method is shown in S2_Table. The CCF of SSNV is demonstrated in S3 _ Table, and the CCF of SCNV is shown in S4 _ Table. In the end, 24359/10480 clonal/sub-clonal events (69.9%) were obtained for SNV, and 129048/5191 clonal/sub-clonal events (96.1%) for SCNV.

Genomic mutant signature analysis

Mutated signatures can reflect the potential effects of previous exposure to different carcinogens, as well as some characteristic changes associated with DNA damages and repair in STAD tumors. Here the brunet algorithm in NMF was used to identify the SNV signature. In order to ensure that the optimal number of SNV signatures could be identified, we evaluated cophenetic and rss when $k = 2-10$ (that is, there were 2–10 SNV signatures), respectively. According to these two indexes, $k = 3$ (that is, three SNV signatures) was chosen as the optimal quantity (S1_Fig). According to the trinucleotide mutation pattern, three SNV signatures were obtained, which were defined as Signature A-C. According to the base mutation pattern, signature A was mainly composed of "C > T", while signature B was mainly composed of "C > A", "C > G" and "C > T", and the mutation pattern of "C > G" only appeared on signature B. Signature C mainly consisted of "T > G" (Fig. 1A). SNV was divided into clonal events and sub-clonal events based on CCF. No significant difference was observed in the contribution of the two types of SSNV to the three mutated signatures (S2 Fig A), showing that clonal events and sub clonal events were similar in mutation patterns.

Identification and distribution of mutant signatures

In order to evaluate the heterogeneity of mutant signatures, contributions of signature A-C were calculated in each sample (the larger the contribution, the higher the proportion of the signature in the sample). It was found that signature A accounted for a large proportion in most samples, while signature B and C accounted for a high proportion only in specific samples (Fig. 1B). By using the known 30 mutant signatures provided by COSMIC, we calculated the cosine similarity between three signatures and COSMIC mutational signatures (expressed by correlation coefficient), finding that signature B had high similarity with signature 3, signature13, signature C and signature 17(Fig. 1C). The similarity between signature A and signature 1 was the strongest (S2 Fig B). Signature 3 was associated with the repair failure of double strand breaks of homologous recombined DNA, which was mainly found in breast, ovarian, and pancreatic cancers. Apart from gastric cancer, Signature 17 was also found in esophageal cancer, breast cancer, liver cancer, lung adenocarcinoma, B-cell lymphoma and melanoma, but its relationship with tumor was still unknown. Signature 1 had been found in all cancer types and most

cancer samples. It was generated by the endogenous mutation caused by spontaneous deamination of 5-methylcytosine. Its number of mutations was related to the age when cancer was diagnosed.

(https://cancer.sanger.ac.uk/cosmic/signatures_v2)

Variation analysis of cloning and subcloning genomes

The clonal/sub-clonal events data of SCN and SSN were integrated, and the clonal and sub-clonal structures of STAD samples were analyzed. The SCN and SSN genes with more than 5% occurrence times in all samples were selected, and 46 SCN genes with the highest occurrence frequency and 101 SSN genes (Fig. 2, S5_Table) were obtained, respectively. The results showed that the number of mutations of TP53, TTN and MUC16 genes in the samples was the highest (> 20%), and the major mutation was clonal events (S3_Fig, enrichment $p < 0.05$, S6_Table), indicating that these genes were more likely to occur as early mutation events. The number of clonal and subclonal mutations in common proto-oncogenes such as PIK3CA and CDH1 was relatively small (< 10%). MIEN1, GRB7 and PNMT genes had the largest number of CNV (Gain) appearing in the samples (> 10%), while for ERBB2, MYC, KRAS and HRAS, the number of CNV was small (6%-10%, S4_Fig).

Analysis of temporal sequence relationship between mutation and tumor evolution

In order to analyze the mutations involved in the occurrence and development of STAD, 46 SCN and 101 SSN with the highest mutation frequency were sorted according to CCF value (Fig. 3A, S5_FigA-B). On the whole, the CCF of SCN was significantly higher than that of SNV (rank test $p < 1e-5$, mean ccf: 0.9287/0.9003). Gain was the major result for SCN, and Loss accounted for a very small proportion. (Gain/Loss:1214/3).

In order to facilitate the demonstration, only the gene pairs with intergenic edges ≥ 2 were retained, and finally 369 SCN pairs (S7_Table) and 119 SSN pairs (S8_Table, S10_Table) were obtained. Five early SCN genes and eight early SSN genes (S9_Table, S10_Table) were obtained by edges enrichment analysis. In the temporal order results of SSN, it was found that TP53, USH2A and GLI3 appeared the earliest in STAD, which could be the driver events of STAD. On the other hand, CTNNB1, LRP1B and ERBB4 appeared the latest in STAD, which might be related to the progress of STAD (Fig. 3B). The results of all SSN's edges are shown in S5_FigC (edges were not filtered). In the temporal order results of SCN, MYC, KRT14 and KRT16 were defined as early genes and metaphase genes, and KRAS, ERBB2 and CCNE1 were late genes (Fig. 3C). For the results of all SCN edges, see S5_FigD (edges were not filtered).

Relationship between cloning or sub-cloning events and prognosis

In order to study the effect of clonal or sub-clonal events on the survival of patients, the kaplan-meier method was adopted to analyze the prognosis relationship between the clonal status and overall survival

of 46 high frequency SCNA genes and 101 high frequency SSNV genes (the number of mutations > 5%). When log rank test $p < 0.1$, five early genes (Fig. 4A-B, S6_Fig) which were obviously related to overall survival, 12 metaphase genes (Fig. 4C-D, S7_Fig) and eight late genes (Fig. 4E-F, S8_Fig) were obtained. From the KM curve of the overall survival rate, it could be found that the clonal events of early genes of MYC and DNAH9 had a more significant effect on the prognosis than the sub-clonal events did. On the other hand, in the intermediate and late genes, clonal events and sub-clonal events showed significant effects on the prognosis of OS. Interestingly, the clonal events of the intermediate gene OBSCN corresponded to better prognosis, while the sub-clonal events corresponded to poor prognosis, which was different from other genes.

Relationship between cloning or sub-cloning events and clinical characteristics

Based on the previous method, clonal events of SCNA and SSNV were obtained. The relationship between clonal events and sub-clonal events and clinical characteristics was analyzed with the clinical information provided by TCGA. The differences of clonal/sub-clonal events in TNM, stage, age, gender and organization types were analyzed. The results showed that there were significant differences in the number of sub-clonal events in T stage (Fig. 5A). N stage, gender and tissue types have significant differences in clonal events (Fig. 5B-D). The risk of gastric adenocarcinoma in males was higher than that in females, which was consistent with our observation that clonal events in males was significantly higher than that in females. Papillary and tubular mutation in tissue types was significantly higher than that in other types. There was no significant difference observed in clonal/sub-clonal events among M, stage, age and grade (S9_Fig).

Relationship between cloning or sub-cloning events and TMB/Neoantigens

Tumor mutation burden (TMB) and neoantigen are important biomarkers in immunity checkpoint therapy, and the appearance of clonal/sub-clonal events also has an important effect on the occurrence and progression of tumor. Therefore, the relationship between clonal/sub-clonal events and TMB and neoantigen was analyzed. Because the distributions of TMB, neoantigen and clonal/sub-clonal events did not conform to normal distribution (shapiro test $p < 1e-5$), the spearman method was used to evaluate the correlation among them. According to the significance test, there is a highly significant relationship among clonal events and TMB and neoantigen (Fig. 6A-C), but the correlation among sub-clonal events and TMB and neoantigen was weak (Fig. 6D), which seemed to indicate that the emergence of clonal events contributed greatly to tumor mutation burden and new antigen production. The mutation of mismatch repair genes (MMR) has an important effect on the mutation burden of genome. The clonal/sub-clonal difference between MMR's mutated samples (Mut) and non-mutated samples (WT) was further evaluated. It was found that the clonal/sub-clonal events in the Mut group was higher than that in the WT group (Fig. 6E), but there was no significance observed, which might be related to the

smaller sample size of the Mut group. TMB and neoantigens in the Mut group were significantly higher than those in the WT group (Fig. 6F), but there was no significant difference between them in OS (Fig. 6G), indicating that although the abnormality of the mismatch repair system had an important effect on genomic stability, its relationship with prognosis was complicated.

Discussion

Gastric cancer is one of the most common cancers in the world and the second leading cause of cancer death [27]. The occurrence of gastric cancer is affected by both environmental and genetic factors. Environmental risk factors for gastric cancer include a high-salt diet, smoking and infection, *Helicobacter pylori*, etc. Because of its complex etiology and anatomical structure, GC is highly heterogeneous in clinic and pathology [19], and its 5-year survival rate varies greatly in different regions [28, 29]. The standard treatment regimen supplemented by operation combined with chemotherapy largely ignores the heterogeneity and features of gastric cancer. Therefore, there is an urgent need to explore biomarkers on the basis of heterogeneity of gastric cancer to develop more effective treatment methods.

Existing studies have shown that cancer is mainly caused by genomic mutations. The identification of mutations from the cancer genome is helpful to the development of cancer pathogenesis and clinical treatment. In recent years, researchers have been studying the molecular mechanism of gastric cancer. In particular, second-generation sequencing, as a high-throughput method, can systematically identify genomic variations in the cancer genome [30]. A number of studies have been carried out on the genomic variation of gastric cancer. For example, studies have reported genomic mutations in gastric cancer such as TP53, PIK3CA, CTNNB1, and CDH1 [31]. In addition, studies have shown that 59% of gastric cancer has mutations in chromatin remodeling genes such as ARID1A, PBRM1 and SETD2 [32]. These studies suggest that the complexity of genomic variation in gastric cancer needs to be fully discussed. As two important sources of genomic variation, SNV variation and copy number variation have played an important role in gastric cancer. This study attempted to analyze the mutation of SNV variation and CNV data in the TCGA gastric cancer dataset, and divided the mutations in TCGA gastric cancer samples into three signatures with significant differences and combined with the COSMIC database, which indicated that they had good similarity with the database. These results deepen our understanding of gastric cancer mutations.

Tumorigenesis is a dynamic evolution process. Tumor heterogeneity comes from the evolution of different subclones in tumor evolution [17]. In the process of tumor occurrence and development, the clone types of tumor cells continue to change due to the accumulation of mutations [33]. According to cancer cell fraction (CCF), the tumor cell population can be divided into clones and subclones. Studies on the clonal evolution model of tumor cells and the influencing factors of cloning help increase the understanding of tumor heterogeneity, which is helpful to clinical individualized therapy [34]. This study attempts to explore the evolutionary models of cloning and subcloning in gastric cancer. Firstly, the clonal/sub-clonal state of STAD mutation was identified by clonal/sub-clonal events analysis of SCNA and SSNV. In addition, there was no significant difference in the status of clonal/sub-clonal among the

three mutant signatures identified in this paper. Furthermore, by selecting the genes with mutations in more than 5% of the samples, it was found that the number of mutations was the highest in TP53, TTN and MUC16 genes (> 20%), and the main mutation was clonal events. Ultimately, a series of early stage genes such as TP53, USH2A, and GLI3 and advanced stage genes such as CTNNB1, LRP1B and ERBB4 were identified by analyzing the relationship between cloning state and tumor evolution. Based on the above results, this paper revealed the cloning status of gastric cancer and identified a series of genes related to the evolution of gastric cancer.

Another important part of this article is to explore the relationship between prognosis and clinical features of patients with gastric cancer. Cloning and subclonal status play an important role in tumor progression [35], but the relationship between cloning and gastric cancer patients is not clear. In this study, 5 early stage genes, 12 metaphase genes and 8 advanced stage genes were identified. The clonal/sub-clonal state of these genes was more significantly related to the overall survival rate of patients. Furthermore, the relationship between clonal/sub-clonal events and clinical pathology parameters is more complex. T stage had significant difference in the number of sub-clonal events, while N stage, gender and tissue type had significant difference in the number of clonal events. No significant differences were observed in the M, stage, age, and grade factors. Finally, this paper focuses on two important biomarkers in the immune checkpoints therapy of tumor mutation load (TMB) and neoantigen [36]. The results showed a significant correlation between clonal/sub-clonal events and TMB/Neoantigens, indicating that the emergence of clonal/sub-clonal events had important contributions to TMB and neoantigens. The above results have shown the relationship between clonal status and clinical characteristics coupled with prognosis of patients with gastric cancer. In addition, it also suggests its great potential in immunotherapy.

Conclusion

In summary, this study has collected data from the (STAD) genome of gastric adenocarcinoma from TCGA to infer the clone (subclone) composition of each tumor sample. The relationship between intratumoral heterogeneity and genomic instability was evaluated using the tumor clonal phylogenetic data. A series of molecular features were screened and identified as markers of potential heterogeneity of gastric cancer, which is of great significance for personalized treatment of gastric cancer.

Abbreviations

TCGA
The Cancer Genome Atlas
NMF
Non-negative matrix factorization
SNV
somatic mutation variation
TMB
Tumor mutation burden

GC
gastric cancer
HP
helicobacter pylori
CNV
copy number variation
CCF
cancer cell fraction
AF
allele fraction

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

All authors declare no conflict of interest.

Funding

This study was supported by the Fund for Shanxi “1331 Project” Key Innovative Research Team and College Students’ Innovation Project (No. D2018385).

Authors’ contributions

CXR, CLW, NNW, CHL, and CQY contributed to the conception and design of the study. CXR collected and analysed the data. DY, CLW and NNW wrote the manuscript. All authors have read and approved the manuscript.

Acknowledgements

Not Applicable

References

1. E. Van Cutsem, X. Sagaert, B. Topal, K. Haustermans, H. Prenen. Gastric cancer. *Lancet*.2016; 388(10060): 2654-2664.
2. H. Yoon, N. Kim. Diagnosis and management of high risk group for gastric cancer. *Gut & Liver*. 2015; 9(1): 5.
3. A. Thiel, A. Ristimäki. Gastric cancer: basic aspects. *Helicobacter*. 2012; 17(s1): 26-29.
4. D.H. Ilson. Advances in the treatment of gastric cancer. *Curr Opin Gastroenterol*. 2017; 33(6): 473-476.
5. U. Harjes. Gastric cancer: Risk analysis. *Nat Rev Cancer*.2018; 18(2): 66.
6. P. Karimi, F. Islami, S. Anandasabapathy, N.D. Freedman, F. Kamangar. Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention, *Cancer epidemiology, biomarkers & prevention* : a publication of the American Association for Cancer Research. American Society of Preventive Oncology.2014; 23(5): 700-13.
7. T.P. Slavin, J.N. Weitzel, S.L. Neuhausen, K.A. Schrader, C. Oliveira, R. Karam. Genetics of gastric cancer: what do we know about the genetic risks? *Translational gastroenterology and hepatology*.2019; 4: 55.
8. T. Shimizu, H. Marusawa, Y. Endo, T. Chiba. Inflammation-mediated genomic instability: roles of activation-induced cytidine deaminase in carcinogenesis. *Cancer Sci*. 2012; 103(7): 1201-6.
9. B. Vogelstein, K.W. Kinzler. Cancer genes and the pathways they control. *Nature medicine*.2004; 10(8) :789-99.
10. M.R. Stratton, P.J. Campbell, P.A. Futreal. The cancer genome. *Nature*.2009; 458(7239): 719-24.
11. R. Beroukhim, C.H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*.2010; 463(7283): 899-905.
12. N.L. Solimini, Q. Xu, C.H. Mermel, A.C. Liang, M.R. Schlabach, J. Luo, et al. Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science*.2012; 337(6090):104-9.
13. W. Xue, T. Kitzing, S. Roessler, J. Zuber, A. Krasnitz, N. Schultz, et al. A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proc. Natl. Acad. Sci. U.S.A*. 2012; 109(21): 8212-7.
14. L. Liang, J.Y. Fang, J. Xu. Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. *Oncogene*.2016; 35(12): 1475-82.
15. Y.J. Bang, E. Van Cutsem, A. Feyereislova, H.C. Chung, L. Shen, A. Sawaki, et al. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet*.2010; 376(9742): 687-97.
16. Comprehensive molecular portraits of human breast tumours. *Nature*.2012; 490(7418):61-70.

17. I. Dagogo-Jack, A.T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*.2018; 15(2): 81-94.
18. A. Sharma, R. DasGupta. Tracking tumor evolution one-cell-at-a-time. *Mol Cell Oncol*.2019; 6(3): 1590089.
19. M.A. Shah, J.A. Ajani. Gastric cancer—an enigmatic and heterogeneous disease. *Jama*.2010; 303(17): 1753-4.
20. M. Apicella, S. Corso, S. Giordano. Targeted therapies for gastric cancer: failures and hopes from clinical trials. *Oncotarget*.2017; 8(34): 57654-57669.
21. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*. 2014 Sep 11;513(7517).
22. Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature*. 2015; Oct 22;526:(7574).
23. Mirzal, A. Nonparametric Tikhonov Regularized NMF and Its Application in Cancer Clustering. *IEEE/ACM Trans Comput Biol Bioinform*.2014; 11:1208-1217.
24. Yu N, Gao YL, Liu JX, Shang J, Zhu R, Dai LY.Co-differential Gene Selection and Clustering Based on Graph Regularized Multi-View NMF in Cancer Genomic Data. *Genes (Basel)*.2018; 9.
25. Ye C, Toyoda K, Ohtsuki T. Blind Source Separation on Non-contact Heartbeat Detection by Non-negative Matrix Factorization Algorithms. *IEEE Trans Biomed Eng*. 2020; Feb:67(2).
26. Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, Kwiatkowski DJ, et al.Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet*. 2016; 06:48(6).
27. J. Ferlay, H.R. Shin, F. Bray, D. Forman, C. Mathers, D.M. Parkin. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer*.2010; 127(12):2893-917.
28. W. Chen, R. Zheng, S. Zhang, P. Zhao, G. Li, L. Wu, J. He. The incidences and mortalities of major cancers in China, 2009. *Chin J Cancer*.2013; 32(3): 106-12.
29. V. Catalano, R. Labianca, G.D. Beretta, G. Gatta, F. de Braud, E. Van Cutsem. Gastric cancer. *Critical reviews in oncology/hematology*.2009; 71(2):127-64.
30. I.P. Ribeiro, J.B. Melo, I.M. Carreira. Cytogenetics and Cytogenomics Evaluation in Cancer. *Int J Mol Sci*.2019; 20(19).
31. H.H. Li-Chang, K. Kasaian, Y. Ng, A. Lum, E. Kong, H. Lim, S.J. Jones, D.G. Huntsman, D.F. Schaeffer, S. Yip. Retrospective review using targeted deep sequencing reveals mutational differences between gastroesophageal junction and gastric carcinomas. *BMC Cancer*. 2015; 15: 32.
32. K. Wang, J. Kan, S.T. Yuen, S.T. Shi, K.M. Chu, S. Law, et al. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nature genetics*.2011; 43(12): 1219-23.
33. S. Turajlic, A. Sottoriva, T. Graham, C. Swanton. Resolving genetic heterogeneity in cancer, *Nature reviews. Genetics*.2019; 20(7):404-416.

34. M.W. Fittall, P. Van Loo. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome medicine*.2019; 11(1) (2019) 20.
35. Y. Assenov, D. Brocks, C. Gerhäuser. Intratumor heterogeneity in epigenetic patterns. *Semin. Cancer Biol.* 2018; 51:12-21.
36. M.J. Duffy, J. Crown. Biomarkers for Predicting Response to Immunotherapy with Immune Checkpoint Inhibitors in Cancer Patients. *Clinical chemistry*.2019; 65(10):1228-1238.

Supplementary Figure Captions

S1 Fig. NMF rank survey. A-C: The distribution of cophenetic, rss and silhouette when rank=2-10 D-F: Consensus matrix when ran=2-4.

S2 Fig. A: Signatures comparison between clonal and sub clonal mutations. Fisher was used to test the preference distribution of signatures on clonal and sub clonal. B: The similarity between Signature A-C and cosmic signature. The relative contribution value was used to calculate the correlation, and the correlation coefficient represents similarity. Figure B only shows cosmic signatures with significant correlation ($P < 0.001$), and the circle size represents the correlation test-log10 (P).

S3 Fig. The clonal and sub-clonal significant enrichment results of genes with mutation rates $> 5\%$. * indicates significant enrichment (enrichment $p < 0.05$).

S4 Fig. The clonal and sub-clonal events frequency distribution of genes with mutation rates $> 5\%$. Un suggests that it cannot be identified as a clonal mutation or sub-clonal mutation.

S5 Fig. A: The CCF distribution of top 5% SSNV mutation genes; B: The CCF distribution of top5% SCNV genes; C: The temporal maps of SSNV acquisitions in STAD. D: The temporal maps of SCNV acquisitions in STAD. The C and D maps contain all the genes with mutation rates higher than 5%, and genes without edges were not deleted.

S6 Fig. The KM curve of early genes' clonal/sub-clonal events and OS

S7 Fig. The KM curve of intermediate genes' clonal/sub-clonal events and OS

S8 Fig. The KM curve of late genes' clonal/sub-clonal events and OS

S9 Fig. The KM curve of early genes' clonal/sub-clonal events and OS

Figures

Figure 1

Distribution of mutant signatures of the TCGA STAD samples. A. the four mutant signatures identified by NMF; B. the composition of the STAD samples' four signatures; C. the composition of COSMIC'S 30 signatures of STAD samples.

Figure 2

Frequency distribution of clonal and sub-clonal events in STAD samples (showing the first 30 genes). Top panel: Quantitative statistics of clonal and sub-clonal events. Bottom panel: the clonal and sub-clonal status of the STAD samples. The genes tagged with red stars represent that they have significant difference in clonal/sub-clonal events. Un indicates that the identification of clonal or sub-clonal events cannot be made.

Figure 3

The temporal order of mutation acquisitions during STAD evolution. A: cancer cell fraction (CCF) distribution of top 5% mutant genes. Genes with only clonal mutations are removed. B: The temporal maps of SSNV acquisitions in STAD. C: The temporal maps of SCNVS acquisitions in STAD. B and C diagrams delete some genes without edges. The arrow indicates that the two genes appeared in the same sample, and the arrow width indicates the number of times the event occurred. The genes labeled red correspond to the early genes, the gray correspond to the intermediate genes, and the blue correspond to the late genes. FDR < 0.05 was selected for temporal order significance test.

Figure 4

Relationship between clonal/sub-clonal status and overall survival rate of SCNV and SSVN (OS). A-B: the KM curve of clonal/sub-clonal status and OS of early genes; C-D: the KM curve of clonal/sub-clonal status and OS of intermediate genes; E-F: the KM curve of clonal/sub-clonal status and OS of late genes. P values are calculated by log rank sum test.

Figure 5

A: The quantitative distribution of clonal events and sub-clonal events on T stage; B: the quantitative distribution of clonal events and sub-clonal events on N stage; C: the quantitative distribution of clonal events and sub-clonal events on gender. D: The quantitative distribution of clonal events and sub-clonal events on different tissue types.

Figure 6

The correlation between clonal/sub-clonal events and TMB and neoantigens. A: The scatter diagram of clonal events and TMB; B: The scatter diagram of sub-clonal events and TMB; C: The scatter diagram of clonal events and neoantigens; D: The scatter diagram of sub-clonal events and neoantigens; E: Clonal/sub-clonal events comparison between the MMR Mut group and the WT group; F: Comparison of

TMB and neoantigens between the MMR Mut group and the WT group; G: The OS KM curves of the MMR Mut group and the WT group. The correlation test used the spearman method (Rho value).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [S1Table.txt](#)
- [S1Fig.tif](#)
- [S2Table.txt](#)
- [S2Fig.tif](#)
- [S5Table.txt](#)
- [S4Table.txt](#)
- [S8Table.txt](#)
- [S7Table.txt](#)
- [S6Table.txt](#)
- [S10Table.txt](#)
- [S4Fig.tif](#)
- [S6Fig.tif](#)
- [S9Table.txt](#)
- [S3Fig.tif](#)
- [S7Fig.tif](#)
- [S9Fig.tif](#)
- [S5Fig.tif](#)
- [S8Fig.tif](#)
- [S3Table.txt](#)