Gastric Cancer
https://doi.org/10.1007/s10120-025-01585-y

ORIGINAL ARTICLE

Multi-omics sequencing of gastroesophageal junction adenocarcinoma reveals prognosis-relevant key factors and a novel immunogenomic classification

Zhao Ma[1,6] · Mengting Li[2,4,5] · Fuqiang Li[2,4,9] · Kui Wu[2,4,9] · Xianxian Wu[8] · Tian Luo[2,4,9] · Na Gao[7] · Huijuan Luo[2,4,9] · Zhilin Sui[8] · Zhentao Yu[8] · Hongjing Jiang[6] · Xiaobin Shang[6] · Chuangui Chen[6] · Jie Yue[6] · Fianbiao Meng[1] · Xiaofeng Duan[6] · Bo Xu[1,3]

Received: 6 September 2024 / Accepted: 10 January 2025

Abstract

Background Gastroesophageal junction adenocarcinoma (GEJAC) exhibits distinct molecular characteristics due to its unique anatomical location. We sought to investigate effective and reliable molecular classification of GEJAC to guide personalized treatment.

Methods We analyzed the whole genomic, transcriptomic, T-cell receptor repertoires, and immunohistochemical data in 92 GEJAC patients and delineated the landscape of genetic and immune alterations. In addition to COSMIC nomenclature, the de novo nomenclature was also utilized to define signatures and investigate their correlation with survival. A novel molecular subtype was developed and validated in other cohorts.

Results We found 30 mutated driver genes, 7 novel genomic signatures, 3 copy-number variations, and 2 V-J gene usages related to prognosis that were not identified in previous study. A high frequency of COSMIC-SBS-384–1 and De novo-SV-32-A was associated with more neoantigen generation and a better survival. Using 19 molecular features, we identified three immune-related subtypes (immune inflamed, intermediate, and deserted) with discrete profiles of genomic signatures, immune status, and clinical outcome. The immune deserted subtype (27.2%) was characterized by an earlier KRAS mutation, worse immune reaction, and prognosis than the other two subtypes. The immune inflamed subtypes exhibited the highest levels of neoantigens, TCR/pMHC-binding strength, CD8 + T-cell infiltration, IFN-α/γ response pathways, and survival rate.

Zhao Ma, Mengting Li, Fuqiang Li and Kui Wu are co-first authors and have contributed equally to this work.

✉ Bo Xu
xubo731@cqu.edu.cn

1 Department of Biochemistry and Molecular Biology, Key Laboratory of Breast Cancer Prevention and Therapy, Ministry of Education, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin 300060, China

2 HIM-BGI Omics Center, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences (CAS), BGI Research, Hangzhou 310000, China

3 Chongqing Key Laboratory of Intelligent Oncology for Breast Cancer, Chongqing University Cancer Hospital and Chongqing University School of Medicine, 181 Hanyu Rd., Shapin District, Chongqing 400030, China

4 Guangdong Provincial Key Laboratory of Human Disease Genomics, Shenzhen Key Laboratory of Genomics, BGI Research, Shenzhen 518083, China

5 College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

6 Department of Minimally Invasive Esophageal Surgery, Key Laboratory of Prevention and Therapy of Tianjin, Tianjin's Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Tianjin 300060, China

7 Department of Pathology, Key Laboratory of Prevention and Therapy of Tianjin, Tianjin's Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Tianjin 300060, China

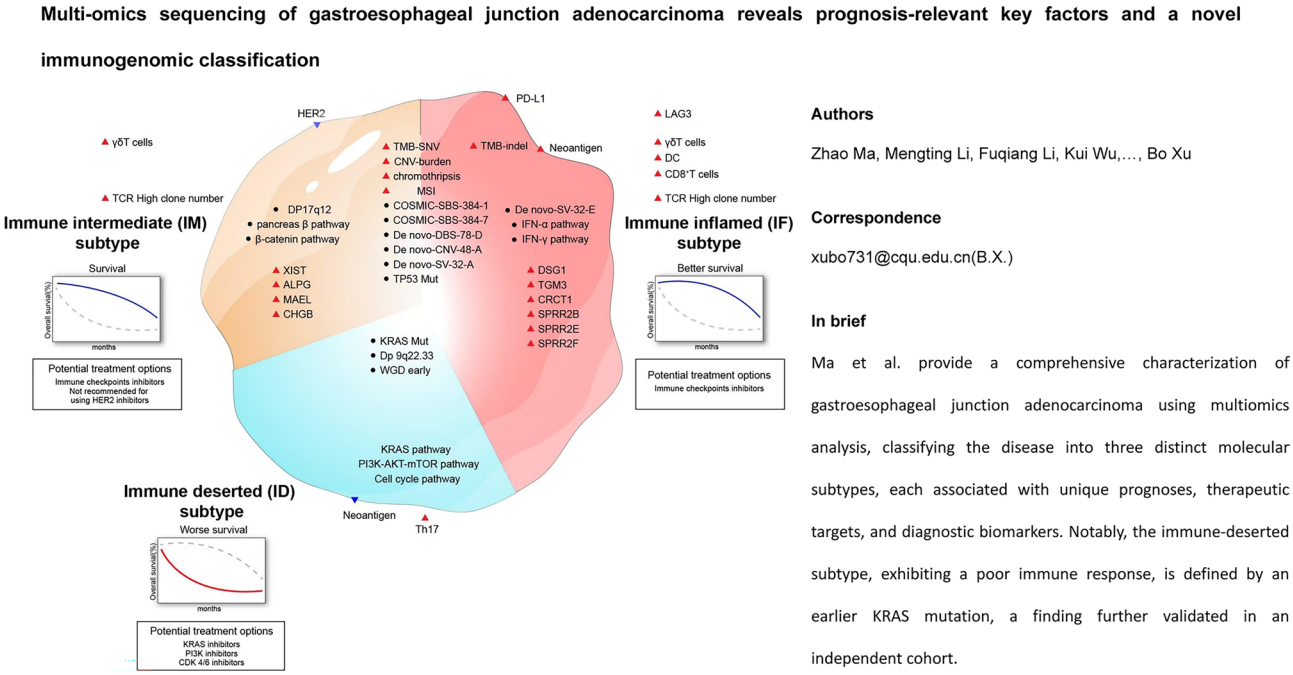
8 Department of Thoracic Surgery, National Cancer Center, National Clinical Research Center for Cancer, Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, China

9 BGI Genomics, Shenzhen 518083, China

Published online: 30 January 2025

Conclusions These results emphasize the immune reaction and prognostic value of novel molecular classifications based on multi-omics data and provide a solid basis for better management of GEJAC.

Graphical abstract



Authors

Zhao Ma, Mengting Li, Fuqiang Li, Kui Wu,…, Bo Xu

Correspondence

xubo731@cqu.edu.cn(B.X.)

In brief

Ma et al. provide a comprehensive characterization of gastroesophageal junction adenocarcinoma using multiomics analysis, classifying the disease into three distinct molecular subtypes, each associated with unique prognoses, therapeutic targets, and diagnostic biomarkers. Notably, the immune-deserted subtype, exhibiting a poor immune response, is defined by an earlier KRAS mutation, a finding further validated in an independent cohort.

Keywords Gastroesophageal junction adenocarcinoma · Molecular classification · Prognosis · Immune reaction · Sequencing

Background

The incidence of gastroesophageal junction adenocarcinoma (GEJAC) in Asian countries has increased alarmingly [1], with highly aggressive malignancy and poor outcomes [2]. Although there are many similarities among oesophageal, gastroesophageal junction, and gastric adenocarcinoma [3], GEJAC exhibits distinct signatures due to its unique anatomical location and complex biological origins. Its features cannot be defined by a single molecular profile [4]. Currently, no uniform and reliable molecular typing of GEJAC exists to guide personalized clinical treatment. Thus, it is urgent to establish a robust and feasible genomic method to classify GEJAC for optimal treatment.

Lin et al. [5]. reported COSMIC signature 17 as a prognostic marker for GEJAC patients. In another small set of GEJAC whole exome sequencing analyses, Hao et al. [6] demonstrated that APOBEC mutational signatures and intact chromosome 4 were correlated with longer survival. Previous studies on GEJAC focused on single base substitution (SBS) signatures without investigating genomic structure variations (SV), which may affect more of the cancer genome than any other type of somatic genetic alteration [7]. We analyzed the prognostic impact of all kinds of genomic alteration signatures, including SBS, double-based substitution (DBS), small insertion and deletion (ID), copy-number variation (CNV), and SV signatures in GEJAC patients.

To develop GEJAC molecular classification, we performed a comprehensive multi-omics molecular analysis, including whole genomic, transcriptomic, TCR repertoires, and immunohistochemical analysis, on 92 patients of GEJAC tissues and paired normal adjacent tissues. Our study may improve current knowledge about the molecular features of GEJAC and provide feasible molecular subtyping to predict effective therapeutic regimens and prognosis.

Results

Prognostic mutational driver genes, signatures, and structural variations in GEJAC

We collected 92 fresh frozen tumor and paired normal samples and performed WGS analysis with a minimum read coverage of ≥ 100 ×. The key clinical characteristics of the patients are summarized in Fig. 1A and Table S1. The most mutated gene was TP53 (72%) (Figure S1A). We used dNdScv, MutSigCV and MutSig2CV software

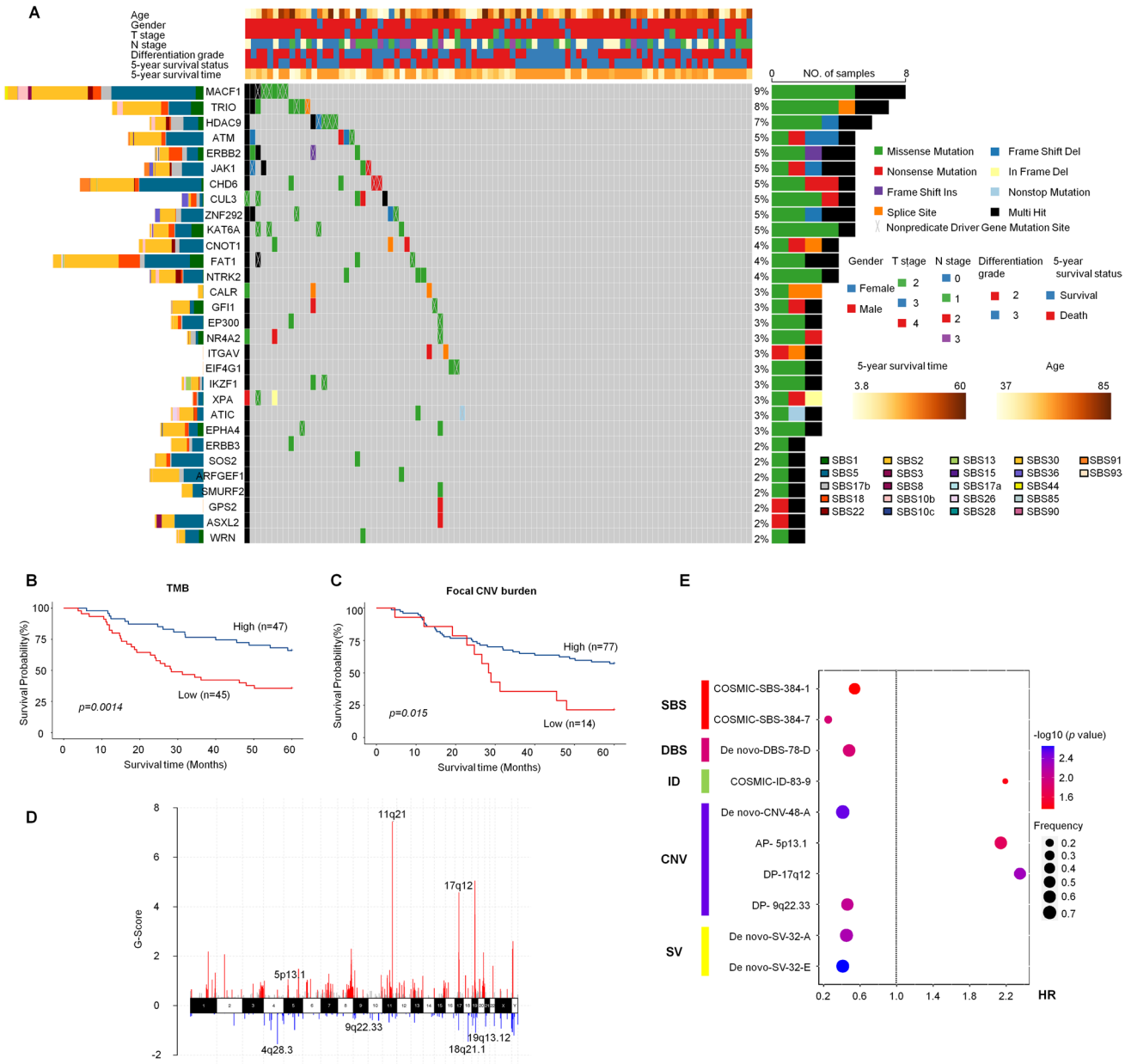


Fig. 1 The featured genomic changes of gastroesophageal junction adenocarcinoma (GEJAC). **A** Summary of significantly mutated driver genes and clinical features of the 92 GEJAC patients. Tiling bars above the heatmap show the distribution of different clinicopathological characteristics. The histogram on the left shows the normalized cumulative mutational contribution of COSMIC signatures for significantly mutated driver genes. The histogram on the right shows the contribution of the mutation type of each driver gene. **B-C** Kaplan—Meier curves of 92 GEJAC patients' 5-year survival according to the level of TMB (b) and focal CNV burden (c). *p* values were calculated using log-rank tests. **D** The frequency of somatic copy-number variations (CNVs) across 22 chromosomes in 92 GEJAC patients. **E** Pooled hazard ratio (HR) of the genomic signatures for 5-year survival rate. The size of bubbles indicates the frequency of signature occurrence in 92 GEJAC patients. The color of bubbles represents the *p* value for the impact of each signature on the 5-year survival. The black dashed line indicates HR = 1

to predict mutated driver genes and identified 30 mutated driver genes with significant impacts on survival in our cohort (all $p < 0.05$, Fig. 1A, S2). Four driver genes (*EPHA4, ZNF292*, *NTRK2, and CNOT1*) were validated for their survival influence value (all $p < 0.05$) in other cohorts [3, 8] (Figure S3A-D). Patients with higher tumor mutation burden (TMB) (≥ 2.92 mutations) had significantly better survival than those with lower (5-year survival, 66.0% *vs.* 35.6%, $p = 0.0014$) (Fig. 1B). We also calculated SNV and InDel mutation burden, naming them TMB-SNV and TMB-InDel [9]. Similar to the total TMB, both of these indicators demonstrated a positive prognostic impact (all $p < 0.01$) (Figure S1B-C).

We mapped the somatic mutation data to the known COSMIC nomenclature. The signatures that contributed the most to the mutations for SBS were SBS2 (2,795,835/6,735,743, 41.5%), SBS5 (1,145,956/6,735,743, 17%), and SBS 17b (473,882/6,735,743, 7.0%), but there was no correlation with prognosis (Figure S4C-E). To identify etiological mutational processes underlying driver mutations, we investigated the mutational signature contribution to 30 predicted driver genes. The prominent features in most driver genes were COSMIC SBS-2 and SBS-5 (Fig. 1A), which are associated with apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like (APOBEC) activities and probably aging or tobacco smoking [10].

For doublet base substitution (DBS), COSMIC DBS-78–1 was the dominant signature (56,063/75,480, 74.3%). After consideration of the transcriptional strand bias [11, 12], the SBS-96 matrix could be further elaborated into SBS-384. We found that the more patients carrying the SBS-384–1 and SBS-384–7 signatures, the longer the survival time were (5-year survival, 63% *vs.* 41%, $p = 0.045$; 82% *vs.* 44%, $p = 0.013$, separately) (Fig. 1E, S5A-B).

To investigate more GEJAC signatures affecting prognosis, we used the de novo method to generate SBS and DBS matrix signatures again. Among 78 DBS signatures, De novo-DBS-78-D (3,658/75,480, 4.8%), which was dominated by thymine nucleotide mutations, was the only signature significantly related to prognosis ($p = 0.013$) (Fig. 1E, S5C). We also detected small InDel features and found that the occurrence of COSMIC-ID-83–9 (5,843/535,941, 1.1%) resulted in worse survival outcome ($p = 0.040$) (Fig. 1E, S5D).

For CNVs, we calculated CNV burden, and high level of focal CNV burden was associated with better prognosis (Fig. 1C). Tumors of longer or shorter survivors exhibited similar CNVs across all 23 chromosomes, except for amplification of 5p13.1 and deletion of 17q12 and 9q22.33 (all $p < 0.05$) (Fig. 1D, 1E, S6A-D). Only the amplification of *MDM2* (5.4%) was closely related to a worse prognosis (median survival time, 60 months *vs.* 18.6 months, $p = 0.03$) (Figure S6E).

We next examined the CNV and SV signatures with de novo methods. Tumors of longer or shorter survivors exhibited similar CNV signatures, except for De novo-CNV-48-A, which occurred more frequently in longer survivors than in shorter survivors (5-year survival, 60.3% *vs.*25.0%) (Fig. 1E, S7A). Human genomes differ more as a consequence of structural variation than of single-base-pair differences [7, 13]. We explored all SV signatures and found that the frequency of De novo-SV-A and De novo SV-32-E were significantly higher in the tumor of longer survivors (all $p < 0.01$, Fig. 1E, S7B-C). Both of the two features were characterized by inversion of less than 1 M long oligonucleotides.

Prognosis relevant immunogenomic subtypes of GEJAC based on 19 features.

To illuminate cooccurring and mutually exclusive genetic lesions, we clustered 19 genomic features that significantly impacted the 5-year survival rate and classified 92 GEJAC patients into 3 distinct molecular subtypes with significant differences in survival time by multi-omics integration and visualization methods (Fig. 2A, S8A). Based on biological features, we termed them the immune inflamed subtype (IF), immune intermediate subtype (IM), and immune deserted subtype (ID). IF subtype patients were associated with the best survival rate, and patients with ID subtype showed the worst survival time (5-year survival 72% *vs.* 54% *vs.* 16%, $p < 0.0001$) (Fig. 2B).

In terms of the genomic alterations, the IM and IF subtypes showed some similarities. Compared with the ID subtype, both of them were characterized by high levels of TMB (total and SNV), focal CNV burden, chromothripsis, microsatellite instability, and enrichment with genomic alteration signatures (COSMIC-SBS-384–1, COSMIC-SBS-384–7, De novo-DBS-78-D, De novo-CNV-48-A, and De novo-SV-32-A) (ID *vs.* IF, all $p < 0.01$;ID *vs.* IM, all $p < 0.05$) (Fig. 2C, S8B-C). However, compared with the IM subtype, the levels of TMB (total, SNV and InDel) and neoantigen in the IF subtype were higher (all $p < 0.05$) (Fig. 2D-E, S8B-C). And the IF subtype was characterized by a positive signature of De novo-SV-32-E (frequency 82.1%, IF *vs*. IM or ID, all $p < 0.05$) (Fig. 2C). In the IM subtype, we detected the highest frequency of 17q12 deletion (frequency 78.6%, IM *vs.* IF or ID, all $p < 0.05$) (Fig. 2C), whose genomic regions harbored the oncogene *ERBB2*. Accordingly, the RNA expression levels of *ERBB2* in the IM subtype were significantly lower than in the IF subtype ($p = 0.028$) and showed a decreasing trend compared to the ID subtype ($p = 0.113$). (Fig. 2F).

The ID subtype exhibited lower levels of several genomic alterations mentioned previously and the survival time of this subtype patients was lowest compared to the IF and IM subtypes respectively (5-year survival, 16% *vs.*

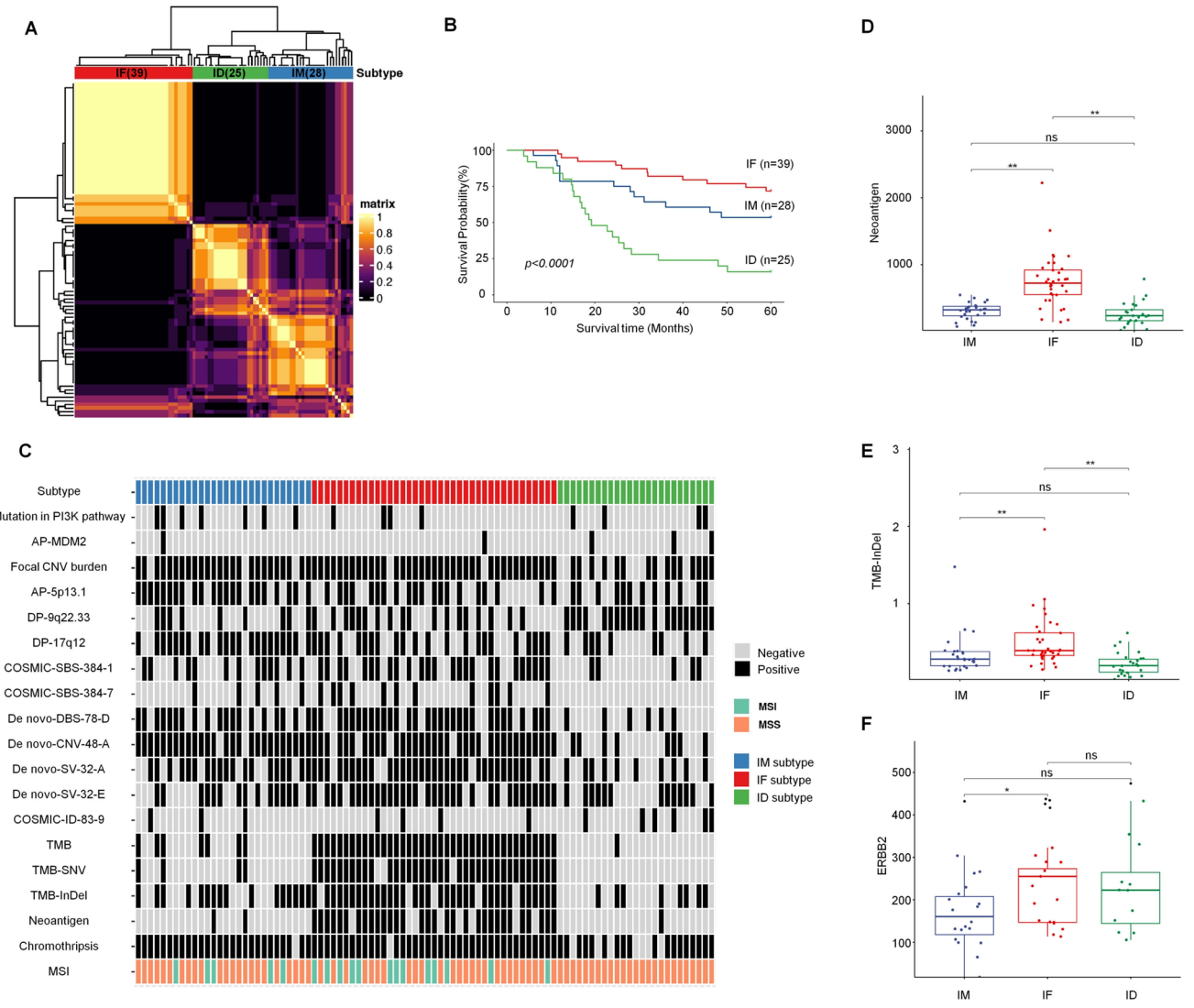


Fig. 2 Genomic subtypes of GEJAC with prognostic significance. **A** Clustered heatmap of 19 genomic features across 92 GEJAC patients. The three subtypes are represented on the top bar, with the number of samples in each subtype indicated in brackets. **B** Kaplan—Meier curves of the 5-year survival of 92 GEJAC patients among the three subtypes. **C** Summary of 19 key genomic features of 92 GEJAC patients grouped by subtypes. The three subtypes are shown at the top bar. All feature data were converted to dichotomous data based on their optimal cut-off value. The black box represents a high occurrence level of genomic features. **D-F** Box plots comparing the level of neoantigen (d), TMB-InDel (e), and expression of ERBB2 (f) among three subtypes. *$p < 0.05$, **$p < 0.01$

72%, $p < 0.000$; 16% *vs.* 54%, $p = 0.003$) (Fig. 2B-E, S8B-C). The differentiation grade of the ID subtype appears to be worse than that of the IM and IF subtypes ($p = 0.132$, $p = 0.138$, separately) (Table S2). The level of T > G substitution frequency of the ID subtype was also lower than that of the IF subtype ($p < 0.01$) (Figure S8D). However, the ID subtype was significantly enriched with deletion genomic regions 9q22.33 (frequency 80%, ID *vs.* IF or IM, all $p < 0.01$) compared with the IM and IF subtypes (Fig. 2C). This feature may play a crucial role in the tumor development of the ID subtypes.

The three subtypes of GEJAC exhibited distinct TCR repertoire features.

Ninety-two GEJAC and paired normal tissues and 64 matched peripheral blood mononuclear cell (PBMC) samples were subjected to TCRβ repertoire sequencing. The most frequent CDR3 nucleotide length was 45 bp (Figure S9A). The number of unique CDR3 sequences in the tumor was lower than that in normal tissues and PBMCs (Fig. 3A). However, the percentage of the top 1000 unique CDR3 amino acids in gastroesophageal junction tissues was

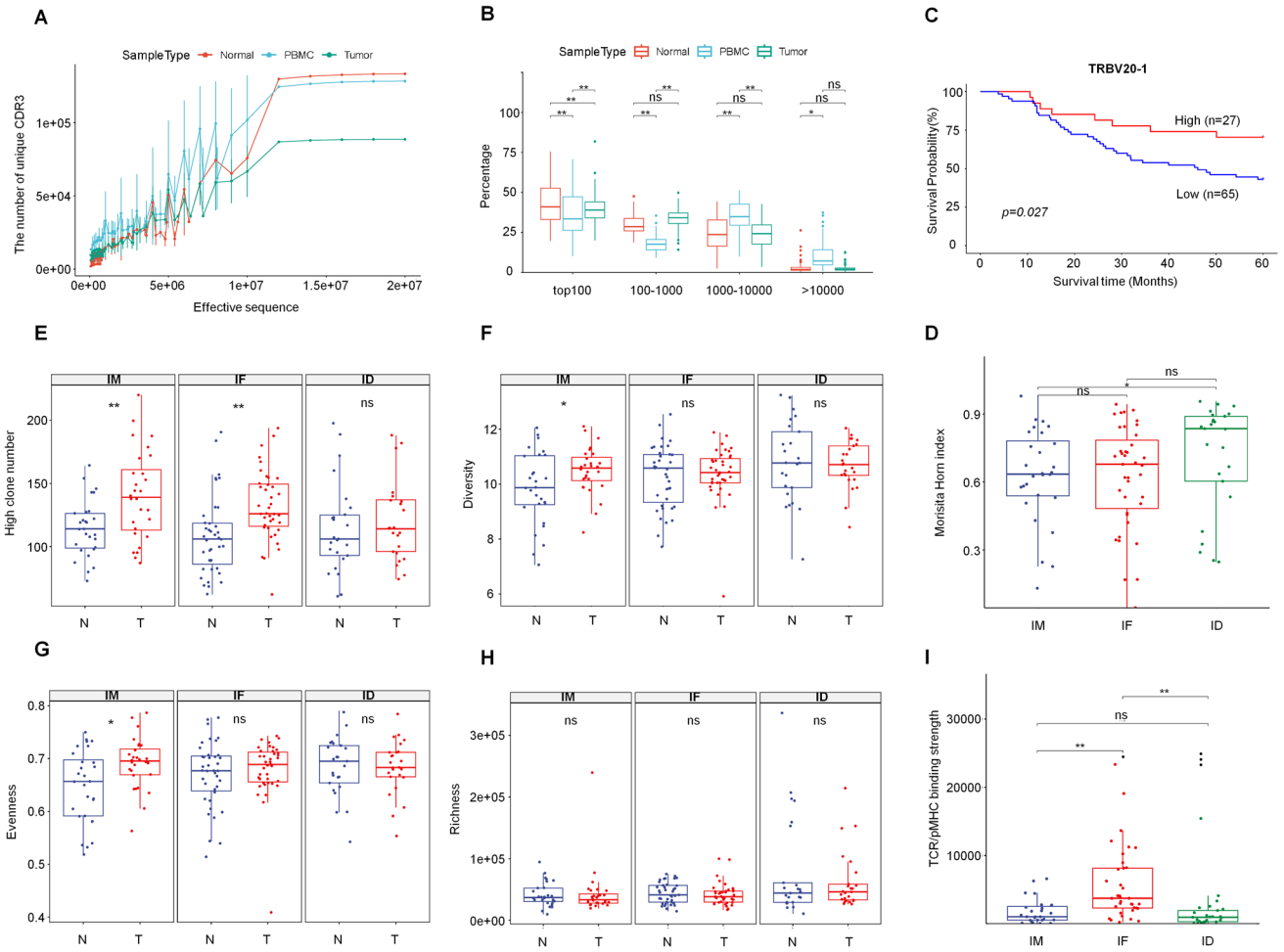


Fig. 3 TCR repertoire analysis in the 3 subtypes. **A** The number of CDR3 sequences in GEJAC and paired normal tissues or PBMCs. **B** Differences in the distribution of CDR3 amino acid sequences between GEJAC and paired normal tissues or PBMCs. **C** Kaplan—Meier curves of patient 5-year survival according to the frequency of TRBV20-1. **D** Box plots comparing the Morisita–Horn index levels among the 3 subtypes. **E–H** Changes in high clone number (e), diversity (f), evenness (g), and richness (h) between tumor and normal tissues among the 3 subtypes. **I** Box plots comparing level of the TCR/pMHC-binding strength among the 3 subtypes. $*p<0.05$, $**p<0.01$

higher than that in PBMCs (all $p<0.01$, Fig. 3B), indicating a higher specificity and concentration of CDR3 fragments in tissues. We estimated the V-J gene utilization profiles of TRB. The segments TRBV20-1 and TRBJ2-1 were the most frequent for intratumoral T cells (Figure S9B-C). Patients with higher TRBV20-1 and TRBJ2-1 in the tumor had better OS (all $p<0.05$) (Fig. 3C, S9D). We utilized the PanPep tool[14] to predict the antigen sequence most likely recognized by TRBV20-1, which was identified as 'LVQRHRSGIR' (Table S3). 29.8% and 27.9% of TRBV20-1 corresponded to HLA-B01 and HLA-A03, respectively.

The Morisita–Horn index of the IM and IF subtypes were lower than that of the ID subtype (IM *vs.* ID, $p=0.042$; IF *vs.* ID, $p=0.068$), indicating that it was easy to activate TCR reaction changes in the IM and IF subtypes (Fig. 3D). Compared to the paired normal tissues, we found that the TCR high clone number was obviously increased in the IM and IF subtypes, but not in the ID subtype (all $p<0.01$) (Fig. 3E). Although there were no significant differences in diversity and clonality among the 3 subtypes in tumors, we found elevated TCR diversity and lower TCR clonality in tumor of the IM subtype than in paired normal tissues (all $p<0.05$, Fig. 3F, S9E). We dissected the single parameters of diversity and found that the increased diversity in tumor of the IM subtype appeared to be most explained by a gain of evenness ($p=0.017$) (Fig. 3G-H). We also calculated the Gini coefficient index difference between IM subtype tumor and normal tissues and the TCR repertoires in tumor of the IM subtype showed more even features ($p=0.051$) (Figure S9F).

To synthesize the analysis of the relationship between neoantigens and TCRs, we adopted TCR/pMHC-binding strength [15], which represents a potentially stronger ability to activate T cells. Compared to the IM and ID subtype, the

IF subtype showed significantly strongest immune activation ability (all $p < 0.001$) (Fig. 3I).

Neoantigen may be generated from certain genomic alterations and are enriched in the immune inflamed subtype.

We used NetMHC (v4.0) and NetMHCpan (v4.1) software to calculate the neoantigens in GEJAC. The high neoantigen load (> 547) improved the prognosis significantly (5-year survival 68% *vs.* 41%, $p = 0.005$) (Fig. 4A). The highest load of neoantigen was enriched in the IF subtypes compared to the IM and ID subtypes (median, 747 *vs.* 329 *vs.* 237 respectively, all $p < 0.01$) (Fig. 2D). Considering the potential bridging role of neoantigens in the genomic and TCR repertoire alterations, we calculated the correlation of neoantigens with the genomic and TCR indices. There were no correlations between neoantigen with the common indices of TCR (all $p > 0.05$). However, an increased number of neoantigens may result in stronger TCR/pMHC binding ($r = 0.92$, $p < 0.01$) (Fig. 4B).

All aforementioned significant genomic signatures were also investigated. High TMB and TMB-SNV were significant factors leading to neoantigens (both $r = 0.99$, $p < 0.001$) (Fig. 4B). COSMIC-SBS-384–1 ($p = 0.047$) and De novo-SV-32-A ($p = 0.001$) were positively associated

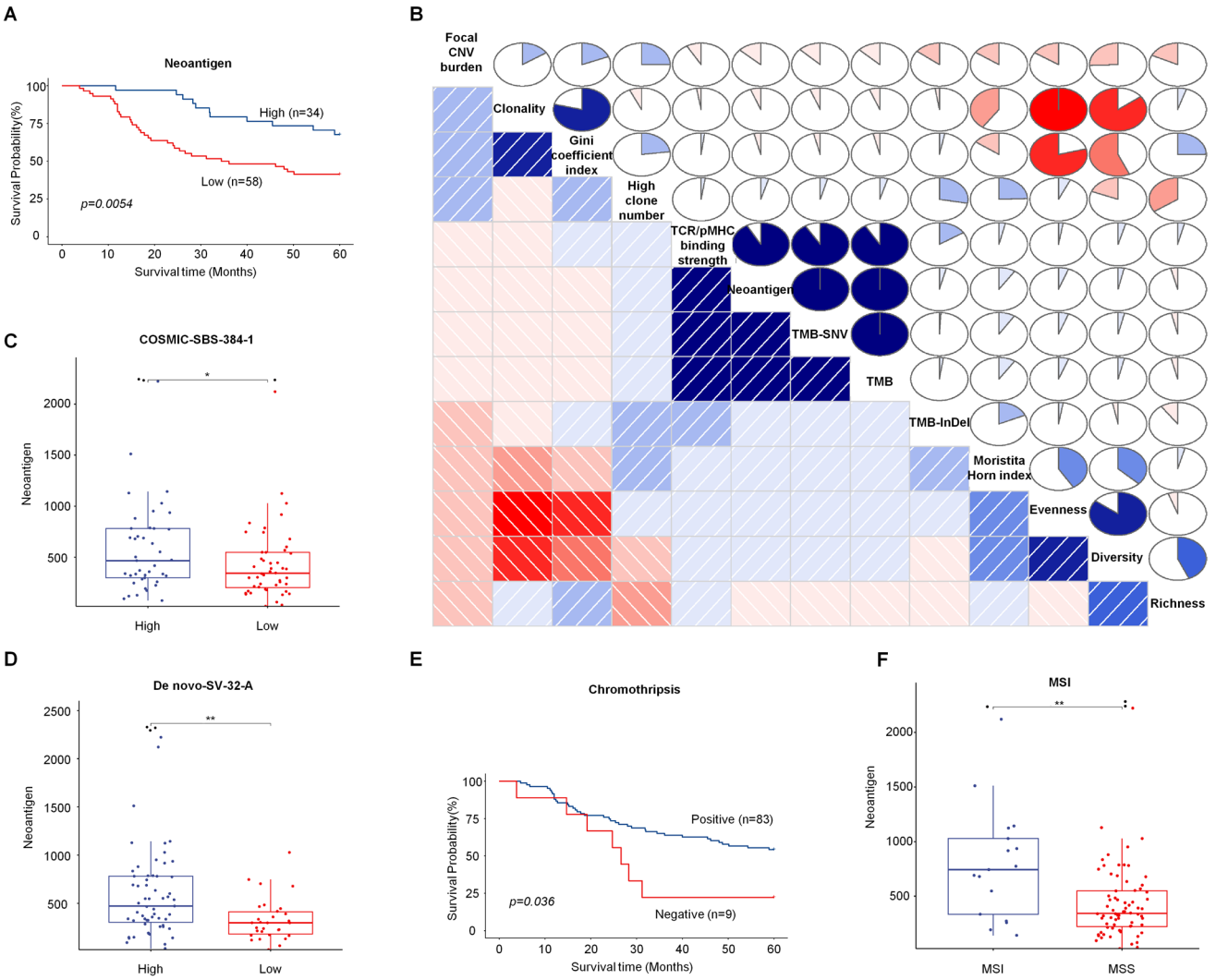


Fig. 4 Genomic and TCR repertoire's features associated with neoantigens. **A** Kaplan—Meier curves of 92 GEJAC patients' 5-year survival according to the expression of predicted neoantigens. **B** A correlation matrix in which the type and depth of coloring were used to highlight the positive (blue) or negative (red) Spearman correlation coefficient between genomic changes and TCR repertoire's features. **C-D** Box plots comparing neoantigen levels in GEJAC samples with and without COSMIC-SBS-384–1 (n = 43 and 49, respectively) (c) and De novo-SV-32-A (n = 62 and 30, respectively) (d). **E** Kaplan–Meier curves of 92 GEJAC patients' 5-year survival according to the expression of chromothripsis. **F** Box plots comparing neoantigen expression in GEJAC samples with and without MSI ($n = 18$ and 74, respectively). $^{*}p < 0.05$, $^{**}p < 0.01$

with neoantigens (Fig. 4C-D). There was a significant correlation between the 5-year survival rate and chromothripsis ($p = 0.032$) (Fig. 4E), but not with microsatellite instability (MSI) ($p = 0.12$) (Figure S9G). However, high MSI was closely related to high neoantigen levels ($p = 0.004$) (Fig. 4F). Chromothripsis tends to result in more predicted neoantigens (Figure S9H). These findings indicate that the high level of TMB, COSMIC-SBS-384–1, De novo-SV-32-A, and MSI lead to the generation of more neoantigens and stronger TCR/pMHC binding.

The immune deserted subtype exhibited an earlier occurrence of KRAS mutations in evolutionary history of GEJAC.

To our knowledge, this is the first study to estimate the order [16] of acquisition of recurrent genomic aberrations, including somatic copy-number alterations (SCNAs), whole-genome doubling (WGD), and common cancer driver genes within each of the subtypes of GEJAC. In IM and IF subtypes, mutations in the driver genes *TP53*, *PIK3CA*, and *ARID1A* were generally early and high frequency (> 10%) events, occurring before WGD (Fig. 5A-B). Compared to the IF and IM subtypes, the appearance of *TP53* mutations occurred later in the ID subtype, while the time of WGD occurrence was much earlier. In the ID subtype, the earliest mutation driver gene event was *KRAS*. (Fig. 5C). The copy-number drivers in the IM and IF subtypes were balanced between gains and loss of heterozygosity (LOH), whereas early events of the ID subtype were dominated by LOH. The loss of 9q, which was specific in the ID subtype, was a relatively early event and resulted in a worse survival outcome (Fig. 5C, S6C).

Tumor immune microenvironment and biological characteristics in the three subtypes.

To explore the association between the molecular classification and tumor biology process, we next deconvoluted the RNA sequencing data of 53 GEJAC patients to infer GSEA pathway analysis and the immune microenvironment, including 28 infiltrating immune cell types and the expression of 20 immune-related genes [17]. Compared with the ID subtype, both the IM and IF subtypes showed a higher abundance of γδT cells and lower abundance of Th17 cells (all $p < 0.05$) (Fig. 6A, S10A-B). Meanwhile, the IF subtype tended to be enriched with higher activated dendritic cells (IF *vs.* ID, $p = 0.030$) and $CD8^+$ T (IF *vs.* ID, $p = 0.090$) cells infiltration compared with the ID subtype, suggesting an active or "hot" immune microenvironment (Fig. 6A, S10C-D).

We then investigated the expression differences of 20 immune checkpoint genes [17]. The expression of *PD-L1* was significantly higher in the IF subtype than in the IM and ID subtypes (IF *vs.* IM, $p = 0.009$; IF *vs.* ID, $p = 0.00075$) (Fig. 6B). We also found the higher

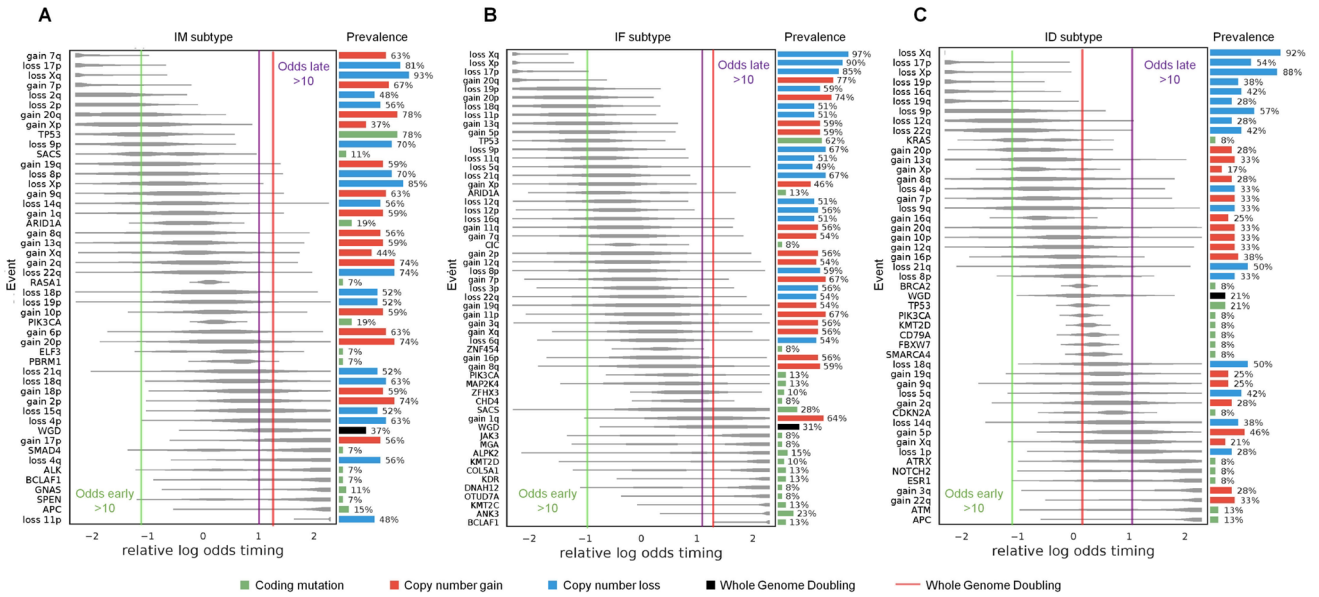


Fig. 5 Diagrams of estimated ordering of significant SCNAs (including chromosome gains/losses and mutations) relative to WGD in 3 subtypes. The size of violin plots indicates the uncertainty of timing for specific events across all samples. The short black solid lines denote the median time. The vertical solid red line represents the median time for WGD events. Events with odds greater than 10, occurring earlier or later, are depicted with vertical solid lines in green or purple. The histogram on the right displays the prevalence of each event in the cohort

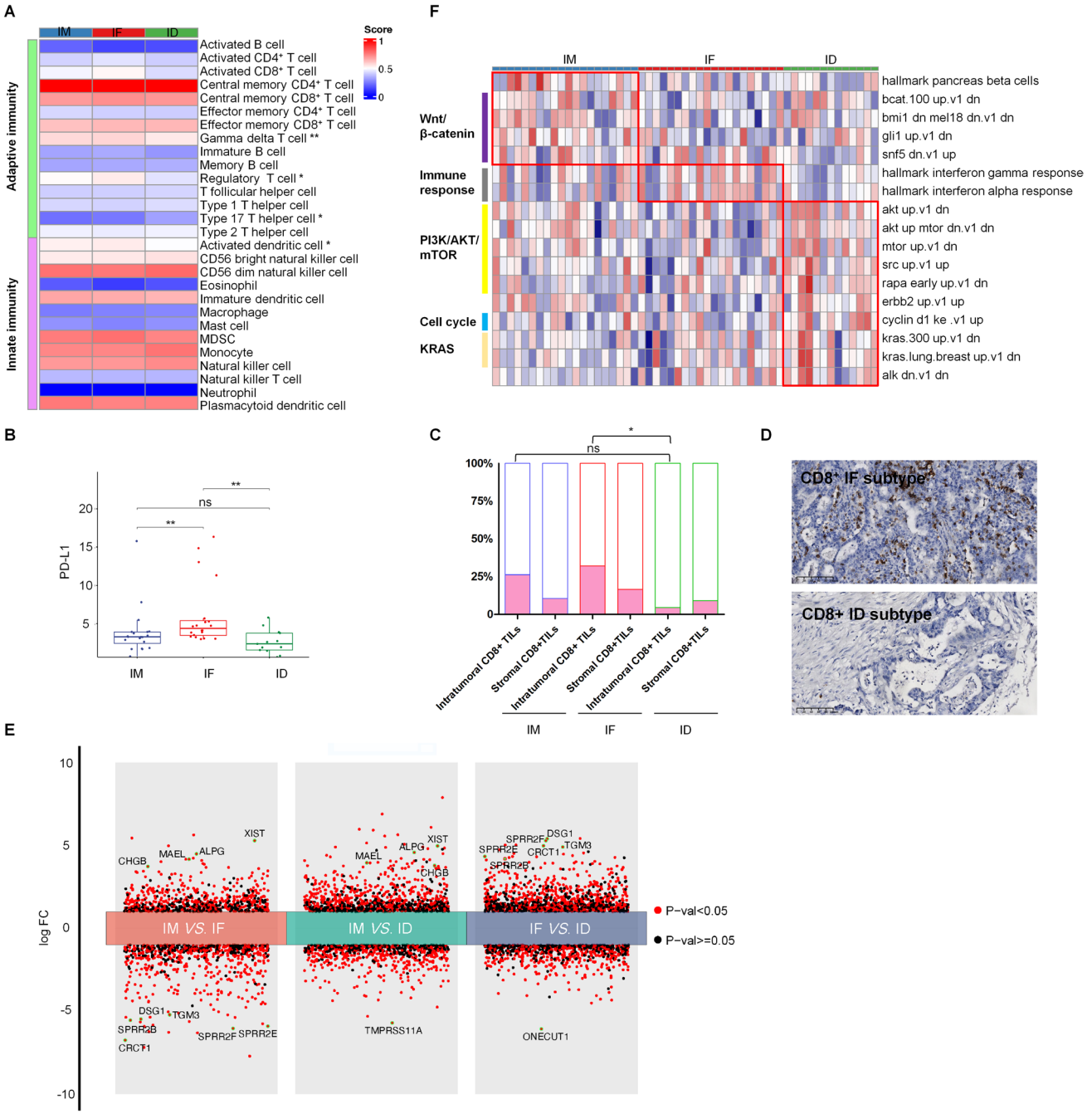


Fig. 6 Transcriptomics analysis of the 3 subtypes. **A** Heatmap showing the normalized enrichment score of tumor-infiltrating immune cell types across the three subgroups. On the right, significant differences in the quantity of infiltrating cell types among the 3 subgroups are indicated with an asterisk (* or **). **B** Box plots comparing the mRNA expression level of PD-L1 among the 3 subtypes. **C)** The proportions of $CD8^+$ T cells in intratumor or stromal tissues separately among the 3 subtypes. **D)** Positive or negative examples of anti-CD8 staining in the IF or ID group patients (X200, positive cells are displayed in brown). **E** Volcano plot showing differences in gene mRNA expression levels among the 3 subtypes. Dots above the horizontal line indicate upregulated genes (log2(fold change) > 0), and Dots under the horizontal line indicate downregulated genes (log2(fold change) < 0). **F** Gene set enrichment analysis was performed across the three subtypes. Differences for all pathways among the 3 subtypes were statistically significant. *$p < 0.05$.**$p < 0.001$

expression of *LAG3* in the IF subtype than in the IM subtype ($p = 0.024$) (Figure S10E). Compared to samples in the ID subtype, our immunohistochemistry analysis also revealed a significant increase in $CD8^+$ T cells abundance in the IM and IF subtypes, but not in $CD3^+$ or CD^{4+} T cells (Fig. 6C-D, S11A-D). This observation may partly explain why the patients with the ID subtype had worse immune reactions and prognosis.

We identified genes that were differentially expression among the 3 subtypes. Among the overlapping genes, the expression of genes (*XIST, ALPG, MAEL, CHGB*) in the IM subtype were more than tenfold higher than that in the IF and ID subtypes (Fig. 6E). In the IF subtype, we also identified tenfold highly expressed genes (*DSG1, CRCT1, TGM3, SPRR2B, SPRR2E, and SPRR2F*) compared to the other two subtypes (Fig. 6E). To further understand the biological features of each subtype, we compared gene expression with respect to a set of significant pathways. The pancreas beta cell pathway and Wnt/β-catenin related pathway were significantly upregulated in the IM subtype (all $p < 0.05$) (Fig. 6F). When compared with the IM and ID subtypes, the IF subtype was characterized by IFN-α and IFN-γ response pathways (all $p < 0.05$) (Fig. 6F). Meanwhile, the ID subtype displayed a significant enrichment of classical oncogenic pathways (KRAS, PI3K–AKT–mTOR, and cell cycle-related pathways) (all $p < 0.05$) (Fig. 6F). The characteristics of IF and ID subtypes were also validated in an additional 124 patients GEJAC cohort [5] (Figure S12A-F). The prognostic value of identified signatures was investigated in esophageal and gastric cancers using data from the TCGA database (Figure S13-14). More details were described in the supplementary data file.

Discussion

There is a lack of useful biomarkers for guiding treatment selection and predicting prognosis. We performed a multi-omics study on a large cohort of GEJAC patients. And this is the largest TCR sequencing analysis study of GEJAC. Utilizing COSMIC nomenclature and the de novo method, we first identified 7 novel genomic mutation and structural alteration signatures. We then classified GEJAC into three molecular subtypes. The three subtypes were distinguished by distinct dominant genomic alterations, tumor immune responses, and prognostic differences, highlighting their potential for personalized therapy. The mutational signatures influencing prognosis in GEJAC were either undetectable in esophageal or gastric cancers or had no impact on their survival. This highlights the distinct molecular characteristics of GEJAC, aligning with findings from previous studies [3, 4].

TMB has been widely explored as an effective biomarker for describing tumor status, response to ICIs and survival [18]. A higher TMB increases the probability of tumor neoantigen production and, therefore, the likelihood of immune recognition and tumor cell killing [19]. For the burden of mutational insertions and deletions (TMB-InDel), it has the potential to engender novel neoantigens that are more immunogenic [20]. The prognostic significance of TMB may vary by tumor type. In esophageal cancer, high TMB is linked to poor prognosis, while in GEJAC and gastric cancer, high TMB is associated with better outcomes. We found that 42.4% of patients (IF group) exhibited high TMB (both TMB-SNV and TMB-InDel), a high neoantigen load, as well as high CNV burden, MSI, and chromothripsis. Recent study indicated that overall outcomes with anti-PD-1-based therapies are favorable in MSI-high tumors [21, 22]. Frequent genomic alterations in the IF subtype exposed more neoantigens, promoted stronger immune recognition and response, and exhibited the best survival. We also found high expression of immune checkpoints, such as *PD-L1* and *LAG3,* in the IF subtype, which reveals that this IF might benefit most from checkpoint blockade therapy.

The IM subtype also exhibited a high frequency of CNV burden, MSI, and chromothripsis and γδT cells, along with a significantly increased high clone number and diversity of TCR clones, which might activate the immune response and enable the IM subtype to benefit from checkpoint blockade therapy [23–26]. However, possibly due to lower levels of TMB, TMB-SNV, and TMB-InDel, the IM subtype had a relatively low neoantigen count and survival time compared to the IF subtype. The IM subtype was characterized by deletion of DP17q12 (containing *ERBB2* gene) with low expression of *ERBB2*. It may not benefit from trastuzumab therapy, but this type was also suitable for immune therapy. Moreover, we identified the unique highly expressed genes in the IM subtype, such as *MEAL*, which promoted colon and hepatocellular cancer cell stemness and drug resistance [27, 28], indicating a potential therapeutic target.

The ID subtype in this study could be classified as a "cold tumor" [29], characterized by the lowest neoantigen exposure, weakest immune reaction and poorest prognosis. Due to low TMB [25], MSI [26], and immune checkpoint genes expression [30], these patients may not benefit from immunotherapy. However, *KRAS* mutations and WGD were earlier event in the ID subtype. Somatic activating or gain-of-function KRAS mutations are usually observed in many tumors and regulate downstream signaling cascades of pathways such as PI3K and MAPK [31, 32]. We observed upregulation of the KRAS and PI3K-AKT-mTOR pathway in the ID subtype in our and an additional cohort. Targeting the KRAS and PI3K-mTOR signaling pathways may be promising therapeutic options for this type. Due to our limited sample size and the occurrence of certain low-frequency mutational events, the generalizability of our findings on prognostic associations may be limited. Due to the unavailability of whole WGS data from the additional cohort, it is not feasible to completely validate our classification model, which may be also a limitation for our study.

Conclusions

Our T-cell receptor repertoire-based multi-omics analysis of a large cohort of GEJAC cases provides a better understanding of novel molecular signatures related to prognosis and

leads to molecular classification into three subtypes. The statuses of TMB-SNV and TMB-InDel are useful implications for stratification. In each subtype, we found potential therapeutic targets and specific molecular characteristics. Our results contribute to the understanding of the molecular landscape of GEJAC and provide a strong starting point for performing meaningful clinical trials.

Materials and methods

Patient samples

We prospectively collected 92 patients with gastroesophageal junction adenocarcinoma from 2010 to 2015 at the Tianjin Medical University Cancer Institute & Hospital. All patients did not receive preoperative treatment and underwent curative R0 surgery. After curative resection, GEJAC patients received systematic fluoropyrimidine-based chemotherapy and regular follow-up. Three public data sets were analyzed. We collected multi-omics data: somatic mutation data of Lin[5] and data from EMBL (PRJEB41070), and the expression data from NCBI (GSE159721). The somatic mutation data from cBioPortal were also used to verify our results.

Whole-genome sequencing and data processing

Whole-genome sequencing and data processing were performed as described previously[33]. The WGS libraries were constructed from 92 primary tumors and their paired normal samples according to the manufacturer's instructions for the MGIEasy FS DNA Library Prep Set (cat. 1,000,006,987; MGI, China). The libraries were sequenced on a DNBSEQ platform (BGI, Shenzhen) and 100-bp paired-end sequencing was performed to yield data of $\geq 100\times$ read coverage for all samples. During WGS data pre-processing, low-quality reads and adaptor sequences were removed by SOAPnuke (v2.0.7)[34]. Sentieon Genomics software was used to map and process high-quality reads for downstream analysis[35].

Somatic short variant calling was performed as described previously[33]. Putative somatic SNVs, MNVs, and/or InDels were identified in each tumor-normal pair using multiple accelerated tools (TNhaplotyper, corresponding to MuTect2[36] of GATK3; TNhaplotyper2, corresponding to MuTect2 of GATK4; TNsnv, corresponding to MuTect[37]) and TNscop[38] of Sentieon Genomics software (version: sentieon-genomics-202010). Somatic CNVs were detected using the Copy-Number Variant caller of Sentieon Genomics software (version: sentieon-genomics-202010), and ascatNgs[39] (version: v4.5). Somatic SVs were detected in each paired normal-tumor sample by TNscope.

Identification of potential driver genes and mutational signatures

We used dNdScv (v0.0.1.0) [40], MutSigCV(v1.41) (http://www.broadinstitute.org/cancer/cga/mutsig), MutSig2CV [41], and CGI(Cancer Genome Interpreter) [42] to identify genes with significantly recurrent coding-sequence SNVs/InDels/CNVs.

Analyses of mutational signatures were performed by SigProfilerExtraction [43] (version v1.1.4) with the parameters-reference_genome GRCh37-opportunity_genome GRCh37-minimum_signatures 1-maximum_signatures 40-nmf_replicates 500-cpu12-gpu True-cosmic_version 3.2. SigProfilerExtraction consists of two processes: de novo signature extraction and signature assignment [44, 45]. Hierarchical de novo extraction of SBS, DBS, and ID signatures from all samples was followed by estimation of the optimal solution (number of signatures) based on the stability and accuracy of all solutions. After identifying the signatures, their activities were estimated by calculating the number of mutations assigned to each sample. SigProfilerExtraction also decomposed de novo signatures to the COSMIC signature database (version 3.2) [46].

RNA sequencing and data processing

Pre-processing of RNA-seq data, including removal of low-quality reads and rRNA reads, was carried out using RiboDetector [47] and Cutadapt [48]. Clean sequencing data were mapped to human reference GRCh37 using STAR [49]. DESeq [50] was used to detect the differential expression genes with threshold of log2 fold changes ≥ 1 and P value < 0.05. We also employed a method called Gene Set Variation Analysis (GSVA) [51] to calculate gene set or pathway scores on a per-sample basis. GSVA transforms a gene by sample gene expression matrix into a gene set by sample pathway enrichment matrix. We made a heatmap of the enrichment matrix, and we used the GSVA scores for a number of other downstream analyses such as differential expression analysis.

TCR-sequence library preparation and sequencing

Genomic DNA from peripheral blood, tumor-adjacent tissues, and tumor tissues was extracted and analyzed for TCR-seq as previously described [52]. Briefly, we used the Multiplex PCR (MPCR) primers, which includes 30 forward V primers and 13 reverse J primers, to amplify the rearranged CDR3 regions of TCRs. Libraries with insert sizes of 200-–300 bp were analyzed using a Bioanalyzer, and 200 bp single-end sequencing was performed on a BGISEQ-500 platform (MGI Tech Co., Ltd.) following the manufacturer's protocol.

In our TCR data, IMonitor [53] was used to analyze TCR sequencing data as described previously [52]. We utilized the VDJtools [54] to calculate the diversity, richness, evenness, gini coefficient index, and Morisita-Horn index of T-cell receptors [24, 55, 56].

Prediction of tumor neoantigens

OptiType (v1.3.5) [57] and Polysolver (v1.0) [58] tools were used for HLA typing calculation and analysis in this study. The results obtained from the two tools' algorithm were merged. Neoantigens were predicted by NetMHC (v4.0) [59] and NetMHCpan (v4.1) [60].

TCR-peptide binding strength prediction

We used ERGO-II(Extended TCR-Peptide Binding Predictor) [15], which is a deep learning based method for predicting TCR and epitope peptide binding. Note that due to the model size, ERGO-II includes here only two models, one for the McPAS database and one for VDJdb. We used the VDJdb database, and we deemed ERGO score above 0.95 is reliable as positive binding in most cases.

Molecular typing

For the subtypes' analysis, we combined all identified genomic features which were significantly associated with 5-year survival. Tally 19 features are PI3K pathway mutation status, amplification of MDM2, focal CNV burden, amplification of 5p13.1, deletion of 17q12, deletion of 9q22.33, COSMIC-SBS-384 – 1, COSMIC-SBS-384 – 7, De novo-DBS-78 – D, De novo-CNV48 – A, De novo-SV-32 – 0, De novo-SV-32 – E, COSMIC-ID-83 – 9, TMB, TMB-SNV, TMB-InDel, Neoantigen, Chromothripsis, and MSI status. Considering the different variable type, we dichotomized PI3K pathway mutation status, amplification of MDM2, amplification of 5p13.1, deletion of 17q12, deletion of 9q22.33, COSMIC-SBS-384 – 1, COSMIC-SBS-384 – 7, De novo-DBS-78 – D, De novo-CNV48 – A, De novo-SV-32 – 0, De novo-SV-32 – E, and COSMIC-ID-83 – 9 into present/absent (the cut-off value of those 12 features were zero). We also dichotomized 5 features into high/low according to those best cut-off value for impacting survival. The thresholds of focal CNV burden, TMB, TMB-SNV, TMB-InDel, and neoantigen were 348, 2.92, 4.08, 0.28, and 547, respectively. Once all the 19 features were binarized, we constructed a matrix of samples using the R Package MOVICS [61] for multi-omics integration and visualization in cancer subtyping.

Inference of clonal structure and phylogenetic relationship

We demonstrate the use of PhylogicNDT [62] by applying it to whole-genome data of 92 samples. Single patient timing and the event timing in the cohort were inferred using PhylogicNDT LeagueModel. We identify significantly different progression trajectories across subtypes of gastroesophageal junction adenocarcinoma.

Statistical analysis

We used Chi-square or Fisher's exact test for any independence test between two categorical variables and Wilcoxon rank-sum test for any independence test between a continuous variable and a binary categorical variable, when there was no covariate to adjust for. Pearson's rank correlation coefficient was used to measure the correlation between two continuous variables. Survival curves were plotted by Kaplan–Meier method. A P value < 0.05 was considered to indicate statistical significance. Data analysis and plot generation were performed in R (version 4.2.3).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10120-025-01585-y.

Acknowledgements The authors would like to thank the National Human Genetic Resources Sharing Service Platform (2005DKA21300), The National Key Research and Development program of China: The Net construction of human genetic resource Bio-bank in North China (2016YFC1201703), Cancer Biobank of Tianjin Medical University Cancer Institute and Hospital for samples storage and processing, and the Guangdong Provincial Key Laboratory of Human Disease Genomics (2020B1212070028). The authors thank China National GeneBank (CNGB) and BGI-Henan for assistance with sequencing and computational resources. The authors also would like to thank Yining Zhang' work on retouching the graphical abstract figure.

Authors' contributions BX, ZTY, KW, and HJJ supervised the project. XXW, ZLS, XBS, CGC, JY, and XFD contributed to patient samples management. FQL and MTL were responsible for the methodology. ZM, LT, and HJL performed the experiments. ZM, MTL, NG, and MFB analyzed data. ZM, MTL, FQL, and XB drafted and revised the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the fund from the Fundamental Research Funds for Universities in Tianjin (2020KJ134) and the Central Universities (No. 2023CDJKYJH002), Chongqing Technology Innovation and Application Development Special Project (CSTB2023TIAD-KPX0050), National Natural Science Foundation (81974464, 82103236), National Natural Science Foundation Cultivation Project by Tianjin Cancer Hospital (230102), National Key Projects of Research and Development of China (2016YFC0904601), and Beijing-Tianjin-Hebei Basic Cooperation Research Project (20JCZXJC00050, 22JCZXJC00040, 19JCZDJC64500(Z)).

Data availability The datasets supporting the conclusions of this article are available in the CNGB Sequence Archive (CNSA) of China

National GeneBank DataBase (CNGBdb) with accession number CNP0004862.

Declarations

Conflict of interest None declared.

References

1. Liu K, Yang K, Zhang W, Chen X, Chen X, Zhang B, et al. Changes of Esophagogastric Junctional Adenocarcinoma and Gastroesophageal Reflux Disease Among Surgical Patients During 1988–2012: A Single-institution. High-volume Experience in China Ann Surg. 2016;263:88–95.
2. Moehler M, Högner A, Wagner AD, Obermannova R, Alsina M, Thuss-Patience P, et al. Recent progress and current challenges of immunotherapy in advanced/metastatic esophagogastric adenocarcinoma. Eur J Cancer. 2022;176:13–29.
3. Cancer Genome Atlas Research Network, Analysis Working Group: Asan University, BC Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, et al. (2017) Integrated genomic characterization of oesophageal carcinoma. Nature. 541:169–175.
4. Geng Q, Lao J, Zuo X, Chen S, Bei JX, Xu D. Identification of the distinct genomic features in gastroesophageal junction adenocarcinoma and its Siewert subtypes. J Pathol. 2020;252:263–73.
5. Lin Y, Luo Y, Sun Y, Guo W, Zhao X, Xi Y, et al. Genomic and transcriptomic alterations associated with drug vulnerabilities and prognosis in adenocarcinoma at the gastroesophageal junction. Nat Commun. 2020;11:6091.
6. Hao D, He S, Harada K, Pizzi MP, Lu Y, Guan P, et al. Integrated genomic profiling and modelling for risk stratification in patients with advanced oesophagogastric adenocarcinoma. Gut. 2021;70:2055–65.
7. Dubois F, Sidiropoulos N, Weischenfeldt J, Beroukhim R. Structural variations in cancer and the 3D genome. Nat Rev Cancer. 2022;22:533–46.
8. 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 U S A. 2015;112:1107–12.
9. Hellmann MD, Paz-Ares L, Bernabe Caro R, Zurawski B, Kim SW, Carcereny Costa E, et al. Nivolumab plus Ipilimumab in Advanced Non-Small-Cell Lung Cancer. N Engl J Med. 2019;381:2020–31.
10. 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.
11. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016;164:538–49.
12. Bergstrom EN, Huang MN, Mahto U, Barnes M, Stratton MR, Rozen SG, Alexandrov LB. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics. 2019;20:685.
13. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578:112–21.
14. Yicheng G, Yuli G, Yuxiao F, Chengyu Z, Zhiting W, Chi Z, et al. Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. Nat Mach Intell. 2023;5:236–49.
15. Springer I, Tickotsky N, Louzoun Y. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. Front Immunol. 2021;12:664514.
16. Zhang T, Joubert P, Ansari-Pour N, Zhao W, Hoang PH, Lokanga R, et al. Genomic and evolutionary classification of lung cancer in never smokers. Nat Genet. 2021;53:1348–59.
17. Zhao X, Subramanian S. Oncogenic pathways that affect antitumor immune response and immune checkpoint blockade therapy. Pharmacol Ther. 2018;181:76–84.
18. Shitara K, Özgüroğlu M, Bang YJ, Di Bartolomeo M, Mandalà M, Ryu MH, et al. Molecular determinants of clinical outcomes with pembrolizumab versus paclitaxel in a randomized, open-label, phase III trial in patients with gastroesophageal adenocarcinoma. Ann Oncol. 2021;32:1127–36.
19. Sholl LM, Hirsch FR, Hwang D, Botling J, Lopez-Rios F, Bubendorf L, et al. The promises and challenges of tumor mutation burden as an immunotherapy biomarker: a perspective from the international association for the study of lung cancer pathology committee. J Thorac Oncol. 2020;15:1409–24.
20. De Mattos-Arruda L, Vazquez M, Finotello F, Lepore R, Porta E, Hundal J, et al. Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the ESMO Precision Medicine Working Group. Ann Oncol. 2020;31:978–90.
21. Randon G, Aoki Y, Cohen R, Provenzano L, Nasca V, Klempner SJ, et al. Outcomes and a prognostic classifier in patients with microsatellite instability-high metastatic gastric cancer receiving PD-1 blockade. J Immunother Cancer. 2023;11(6):e007104.
22. Ooki A, Osumi H, Yoshino K, Yamaguchi K. Potent therapeutic strategy in gastric cancer with microsatellite instability-high and/or deficient mismatch repair. Gastric Cancer. 2024;27:907–31.
23. Porciello N, Franzese O, D'Ambrosio L, Palermo B, Nisticò P. T-cell repertoire diversity: friend or foe for protective antitumor response. J Exp Clin Cancer Res. 2022;41:356.
24. Yan C, Ma X, Guo Z, Wei X, Han D, Zhang T, et al. Time-spatial analysis of T cell receptor repertoire in esophageal squamous cell carcinoma patients treated with combined radiotherapy and PD-1 blockade. Oncoimmunology. 2022;11:2025668.
25. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015;348:124–8.
26. Marabelle A, Le DT, Ascierto PA, Di Giacomo AM, De Jesus-Acosta A, Delord JP, et al. Efficacy of Pembrolizumab in Patients With Noncolorectal High Microsatellite Instability/Mismatch Repair-Deficient Cancer: Results From the Phase II KEYNOTE-158 Study. J Clin Oncol. 2020;38:1–10.
27. Li Q, Wei P, Huang B, Xu Y, Li X, Li Y, et al. MAEL expression links epithelial-mesenchymal transition and stem cell properties in colorectal cancer. Int J Cancer. 2016;139:2502–11.
28. Shi C, Kwong DL, Li X, Wang X, Fang X, Sun L, et al. MAEL Augments Cancer Stemness Properties and Resistance to Sorafenib in Hepatocellular Carcinoma through the PTGS2/AKT/STAT3 Axis. Cancers (Basel). 2022;14(12):2880.
29. Galon J, Bruni D. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. Nat Rev Drug Discov. 2019;18:197–218.
30. Patel MA, Kratz JD, Lubner SJ, Loconte NK, Uboha NV. Esophagogastric cancers: integrating immunotherapy therapy into current practice. J Clin Oncol. 2022;40:2751–62.
31. Thein KZ, Biter AB, Hong DS. Therapeutics targeting mutant KRAS. Annu Rev Med. 2021;72:349–64.
32. Puliga E, De Bellis C, Vietti Michelina S, Capeloa T, Migliore C, Orrù C, et al. Biological and targeting differences between the rare KRAS A146T and canonical KRAS mutants in gastric cancer models. Gastric Cancer. 2024;27:473–83.

33. Luís Nunes FL, Meizhen Wu TL, Klara Hammarström EL, Ingrid Ljuslinder AM, Per-Henrik Edqvist AL, Carl Zingmark SE, Chatarina Larsson LM. Prognostic whole-genome and transcriptome signatures in colorectal cancers. Nature. 2023;633:137–46.
34. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience. 2018;7:1–6.
35. Donald Freed RA, Weber JA, JSE. The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. BioRxiv. 2017;89(6):700.
36. David Benjamin TS, Kristian Cibulskis GG, Chip Stewart LL. (2019) Calling Somatic SNVs and Indels with Mutect2. Preprint at http://biorxiv.org/lookup/doi/https://doi.org/10.1101/861054
37. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9.
38. Donald Freed RP, Aldana R. (2018) TNscope: Accurate Detection of Somatic Mutations with Haplotype-based Variant Candidate Detection and Machine Learning Filtering. Preprint at http://biorxiv.org/lookup/doi/https://doi.org/10.1101/250647
39. Raine KM, Van Loo P, Wedge DC, Jones D, Menzies A, Butler AP, et al. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. Curr Protoc Bioinformatics. 2016;56:15–9.
40. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal patterns of selection in cancer and somatic tissues. Cell. 2018;173:1823.
41. 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.
42. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10:25.
43. Islam S, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. Cell Genom. 2022;2:100179.
44. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
45. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534:47–54.
46. Nunes L, Li F, Wu M, Luo T, Hammarström K, Torell E, et al. Prognostic genome and transcriptome signatures in colorectal cancers. Nature. 2024;633:137–46.
47. Deng ZL, Münch PC, Mreches R, McHardy AC. Rapid and accurate identification of ribosomal RNA sequences via deep learning. Nucleic Acids Res. 2022;50: e60.
48. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
49. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
50. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
51. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
52. Lin Y, Peng L, Dong L, Liu D, Ma J, Lin J, et al. Geospatial Immune Heterogeneity Reflects the Diverse Tumor-Immune Interactions in Intrahepatic Cholangiocarcinoma. Cancer Discov. 2022;12:2350–71.
53. Zhang W, Du Y, Su Z, Wang C, Zeng X, Zhang R, et al. IMonitor: a robust pipeline for TCR and BCR repertoire analysis. Genetics. 2015;201:459–72.
54. Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, et al. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. PLoS Comput Biol. 2015;11:e1004503.
55. Magurran AE. Measuring biological diversity. Curr Biol. 2021;31:R1174–7.
56. Kansy BA, Shayan G, Jie HB, Gibson SP, Lei YL, Brandau S, et al. T cell receptor richness in peripheral blood increases after cetuximab therapy and correlates with therapeutic response. Oncoimmunology. 2018;7: e1494112.
57. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014;30:3310–6.
58. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins T, Wilson GA, et al. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell. 2017;171:1259-1271.e11.
59. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016;32:511–7.
60. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017;199:3360–8.
61. Lu X, Meng J, Zhou Y, Jiang L, Yan F. MOVICS: an R package for multi-omics integration and visualization in cancer subtyping. Bioinformatics. 2021;36:5539–41.
62. Leshchiner I, Mroz EA, Cha J, Rosebrock D, Spiro O, Bonilla-Velez J, et al. Inferring early genetic progression in cancers with unobtainable premalignant disease. Nat Cancer. 2023;4:550–63.