**Tumor microenvironment characterization in gastric cancer identifies prognostic and immunotherapeutically relevant gene signatures**

Dongqiang Zeng M.D.1\*, Meiyi Li M.D.1\*, Rui Zhou, M.D.1, Jingwen Zhang, M.D.1, Huiying Sun, M.D.1, Min Shi, M.D., Ph.D.1, Jianping Bin, M.D., Ph.D.2, Yulin Liao, M.D., Ph.D.2, Jinjun Rao M.D., Ph.D.3, Wangjun Liao, M.D., Ph.D.1**#**

1Department of Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, P. R. China

2Department of Cardiology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, P. R. China

3Key laboratory of new drug screening of Guangdong province. School of Pharmaceutical Sciences, Southern Medical University, Guangzhou, Guangdong, P. R. China

\*These authors contributed equally to this work

**#Corresponding author**

Wangjun Liao, Department of Oncology, Nanfang Hospital, Southern Medical University, 1838 North Guangzhou Avenue, Guangzhou 510515, China. Tel: +86-20-62787731; Fax: +86-20-61360657; E-mail: nfyyliaowj@163.com.

**Supplementary Methods**

**1. Gastric cancer dataset search strategy**

The following search parameters were used: (((survival OR prognosis OR prognostic OR outcome OR death OR relapse OR recurrence))) AND ((gastric cancer[MeSH Terms]) OR ((((((((((gastric cancer[Title]) OR gastric adenocarcinoma[Title]) OR gastric neoplasm[Title]) OR gastric tumor[Title]) OR gastric carcinoma[Title]) OR stomach cancer[Title]) OR stomach adenocarcinoma[Title]) OR stomach neoplasm[Title]) OR stomach tumor[Title]) OR stomach carcinoma[Title])). In the initial search, 656 items were recognized, and only the first 100 were independent chip series. Among the 100 series, 32 contained mRNA expression profiles of cancer tissues from patients with gastric cancer; these were: GSE17154, GSE57176, GSE36398, GSE13861, GSE21983, GSE22377, GSE17187, GSE15460, GSE15459, GSE15456, GSE15081, GSE13911, GSE63288, GSE87666, GSE80389, GSE80388, GSE54129, GSE26899, GSE26901, GSE79973, GSE66229, GSE62254, GSE49051, GSE30727, GSE51105, GSE26253, GSE51575, GSE2998, GSE38749, GSE37023, GSE38024, GSE84437, and GSE28541. Moreover, two additional series were subsequently identified from the subseries list of the corresponding super series (GSE34942 and GSE57303), and one additional series was identified from the related literature (GSE29272). Among the 35 items, three were repeated: GSE15460, GSE15456, and GSE66229; nine items included fewer than 40 patients: GSE57176, GSE17154, GSE17187, GSE87666, GSE79973, GSE49051, GSE30727, GSE51575, and GSE38749; and four were derived from high-throughput sequencing data: GSE80389, GSE80388, GSE63288, and GSE36398. Among the remaining, we only obtained survival data for the following six items: GSE62254/ACRG, GSE15459, GSE29272, GSE84437, GSE26253, and GSE57303. The patients related to these six items were included in the subsequent analysis.

**2. Data sources and preprocessing**

**①** Raw data from the microarray datasets generated using Affymetrix® and Illumina® were downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). The raw data for the dataset from Affymetrix® were processed using the RMA algorithm for background adjustment using the “*Affy”* package (1). RMA was used to perform background adjustment, quantile normalization, and final summarization of oligonucleotides per transcript using the median polish algorithm. The raw data for the dataset from Illumina® were processed using the *“lumi”* package.The *“ComBat”* algorithm (2) was applied to reduce the likelihood of batch effects from non-biological technical biases.

② Level 3 RNA-Seq data (FPKM normalized) for genes from 407 treatment naïve TCGA-STAD samples processed on 2017-09-14 were downloaded from the UCSC Xena browser (<http://xena.ucsc.edu/>) GDC hub (<https://xenabrowser.net/datapages/?dataset=TCGA-STAD%2FXena_Matrices%2FTCGA-STAD.htseq_fpkm.tsv&host=https%3A%2F%2Fgdc.xenahubs.net> ) and can be downloaded at (<https://gdc.xenahubs.net/download/TCGA-STAD/Xena_Matrices/TCGA-STAD.htseq_fpkm.tsv.gz>)

③ Gene-level mutation data (non-silent somatic mutations; wustl\_hiseq automated) for 395 samples were obtained from <https://xenabrowser.net/datapages/?dataset=TCGA.STAD.sampleMap%2Fmutation_wustl_hiseq_gene&host=https%3A%2F%2Ftcga.xenahubs.net>.

④ Updated clinical data and sample information for TCGA-STAD samples were obtained from GDC using the R package “*TCGAbiolinks” (3)*.

⑤ The numbers of predicted neo-epitopes based on tumor-specific HLA typing, total mutations, and CYT for each patient were obtained for 263 STAD samples from Supplementary Table S4 of Rooney et al. (4) and are available at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4856474/bin/NIHMS717941-supplement-Table_S4.xlsx>.

**⑥** Somatic copy number alteration (SCNA) levels, immune signature scores, and cell cycle signature scores were obtained for 269 STAD samples from Supplementary Table S7 of Davoli et al. (5), available at
<http://science.sciencemag.org/highwire/filestream/689461/field_highwire_adjunct_files/7/aaf8399-Davoli-SM-table-S7.xlsx>.

**⑦** Five genomic and transcriptomic datasets from patients with metastatic urothelial cancer (6) treated with anti-PD-L1 agent (atezolizumab), patients with metastatic melanoma (7) treated with anti-PD-1 (pembrolizumab), patients with advanced melanoma treated with various types of immunotherapy from TCGA-SKCM cohort (8), patients with advanced melanoma treated with MAGE-3 antigen-based immunotherapy (9) and mouse model treated with anti-CTLA4 antibody (10) were downloaded and analyzed to determine the predictive value of the TMEscore. For the urothelial cancer dataset, a fully documented software and data package is freely available under the Creative Commons 3.0 license and can be downloaded from http://research-pub.gene.com/IMvigor210CoreBiologies. After quality control using the R package “*arrayQualityMetrics”* (Bioconductor), count data were normalized using the trimmed mean of M-values (TMM) and transformed with *voom* to log2-counts per million with associated precision weights (11). For the melanoma dataset (GSE78220, N = 28), expression profiles (FPKM normalized) and phenotypes have been deposited into the Gene Expression Omnibus under the accession code GSE78220. The expression profiles (FPKM normalized) of GSE78220 were transformed into TPM (transcripts per kilobase million), converting FPKM data to values more comparable between samples (12). For TCGA-SKCM cohort, the expression profiles (FPKM normalized) download from UCSC Xena browser were transformed into TPM which used to calculate TMEscore. For the melanoma cohort (GSE35640, N = 55) treat with MAGE-3 antigen-based immunotherapy, the raw data was downloaded and processed using the RMA algorithm for background adjustment using the “*Affy”* package (1). For mouse model treated with CTLA-4 blockade (accession number GSE63557, N = 20), the normalized data was obtained from GEO and annotated with GPL19103 file.

⑧ Level 3 RNA-Seq data (FPKM normalized) were used to determine gene expression profiles of 7,453 treatment naïve TCGA samples with 14 other cancer types; these data were downloaded from the UCSC Xena browser (http://xena.ucsc.edu/) GDC hub. Corresponding survival data were accessed from supplementary data of recently published TCGA Pan-Cancer research (13).

**3. Other gene signatures enrolled in this study**Because KEGG pathways often include large numbers of genes with only vaguely related functions, we used a gene set curated by Mariathasan et al. (6) (**Supplementary Tables S10**), including:

**A**: CD8 T-effector signature (14)

**B:** Antigen processing machinery (15)

**C:** Immune checkpoint

**D:** Epithelial-mesenchymal transition (EMT) markers previously reported (16)

**E:** Pan-fibroblast TGF-β response signature (Pan-F-TBRS) (6)

**F:** DNA replication-dependent histones (6)

**G:** Select members of the DDR-relevant gene set (17)

**H:** Angiogenesis signature previously reported (18)

**I:** Cell cycle genes (KEGG)

**J:** WNT targets (19)

**K:** Cell cycle regulators (20)

**L:** Mismatch repair (KEGG)

**M:** Nucleotide excision repair (KEGG)

**N:** Homologous recombination (KEGG)

**4. Generation of TME gene signatures and signature score computation**

The construction of TME signature gene sets was performed as follows. First, each differentially expressed gene (DEG) among the three TME phenotypes (A, B, and C) was standardized so that its mean expression was zero and standard deviation was 1 across all gastric cancer samples from the ACRG cohort. And unsupervised clustering method (K-means) (21) for analysis of DEGs was used to classify patients into three groups for further analysis. Then, the Boruta algorithm (22) (a wrapper built around the random forest classification algorithm) was used to performed dimension reduction in order to reduce the noise or redundant genes. Next, the “*clusterProfiler”* R package (23) was adopted to annotate the cluster of genes. Then, a consensus clustering algorithm (24) was applied to define the gene patterns and to calculate the signature score. For gene expression (normalized by RMA or TPM methods) analysis, the expression of each gene in a signature was first transformed into a z-score. Then, a principal component analysis (PCA) was performed, and principal component 1 was extracted to serve as the gene signature score. This approach has the advantage of focusing the score on the set with the largest block of well-correlated (or anti-correlated) genes in the set, while down-weighting contributions from genes that do not track with other set members. After obtaining the prognostic value of the signature score for each gene pattern, we applied a method similar to GGI (25) to calculate the TMEscore of each patient: TMEscore= ∑ PC1i *–* ∑PC1j., where *i* is the signature score of clusters whose Cox coefficient is positive, while *j* is the expression level of genes whose Cox coefficient is negative.

**5. Other plot functions**

R package “*ggtree*” (26) was employed to visualize phylogenetic trees of TME signature genes (**Fig. S3B**). A developing R package “[*enrichplot*](http://bioconductor.org/packages/enrichplot)” , (<https://github.com/GuangchuangYu/enrichplot>) implements several visualization methods to help interpreting enrichment results and was adopted to visualize GSEA result of TME-gene-cluster-A and TME-gene-cluster-C (**Fig. 3E**). All heatmap (**Fig. 1C**, **Fig. 2A**, **Fig. S1C**, **Fig. S2F**, **Fig. S4A-B**) were generated by the function of “*pheatmap*”. OncoPlot (**Fig. 4E**)used to depict mutation landscape of TCGA-STAD cohort was constructed by “*ComplexHeatmap*” R package (27).

**References**

1. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics **2004**;20(3):307-15 doi 10.1093/bioinformatics/btg405.

2. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **2007**;8(1):118-27 doi 10.1093/biostatistics/kxj037.

3. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D*, et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research **2016**;44(8):e71-e doi 10.1093/nar/gkv1507.

4. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell **2015**;160(1-2):48-61 doi 10.1016/j.cell.2014.12.033.

5. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science **2017**;355(6322) doi 10.1126/science.aaf8399.

6. Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y*, et al.* TGFbeta attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature **2018**;554(7693):544-8 doi 10.1038/nature25501.

7. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S*, et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell **2016**;165(1):35-44 doi 10.1016/j.cell.2016.02.065.

8. Network CGA. Genomic Classification of Cutaneous Melanoma. CELL **2015**(1097-4172 (Electronic)) doi 10.1016/j.cell.2015.05.044.

9. Ulloa-Montoya F, Louahed J, Dizier B, Gruselle O, Spiessens B, Lehmann FF*, et al.* Predictive Gene Signature in MAGE-A3 Antigen-Specific Cancer Immunotherapy. Journal of Clinical Oncology **2013**;31(19):2388-95 doi 10.1200/jco.2012.44.3762.

10. Lesterhuis WJ, Rinaldi C, Jones A, Rozali EN, Dick IM, Khong A*, et al.* Network analysis of immunotherapy-induced regressing tumours identifies novel synergistic drug combinations. Scientific Reports **2015**;5(1) doi 10.1038/srep12298.

11. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W*, et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **2015**;43(7):e47-e doi 10.1093/nar/gkv007.

12. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences **2012**;131(4):281-5 doi 10.1007/s12064-012-0162-3.

13. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD*, et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell **2018**;173(2):400-16 e11 doi 10.1016/j.cell.2018.02.052.

14. Rosenberg JE, Hoffman-Censits J, Powles T, van der Heijden MS, Balar AV, Necchi A*, et al.* Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single arm, phase 2 trial. Lancet (London, England) **2016**;387(10031):1909-20 doi 10.1016/S0140-6736(16)00561-4.

15. Şenbabaoğlu Y, Gejman RS, Winer AG, Liu M, Van Allen EM, de Velasco G*, et al.* Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome Biology **2016**;17(1) doi 10.1186/s13059-016-1092-z.

16. Damrauer JS, Hoadley Ka Fau - Chism DD, Chism Dd Fau - Fan C, Fan C Fau - Tiganelli CJ, Tiganelli Cj Fau - Wobker SE, Wobker Se Fau - Yeh JJ*, et al.* Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. Proc Natl Acad Sci **2014**(1091-6490 (Electronic)).

17. Lange SS, Takata K-i, Wood RD. DNA polymerases and cancer. Nature Reviews Cancer **2011**;11:96 doi 10.1038/nrc2998.

18. Sjodahl G, Lauss M Fau - Lovgren K, Lovgren K Fau - Chebil G, Chebil G Fau - Gudjonsson S, Gudjonsson S Fau - Veerla S, Veerla S Fau - Patschan O*, et al.* A molecular taxonomy for urothelial carcinoma. Clin Cancer Res **2012**(1078-0432 (Print)).

19. Spranger S, Bao R, Gajewski TF. Melanoma-intrinsic beta-catenin signalling prevents anti-tumour immunity. Nature **2015**(1476-4687 (Electronic)).

20. Cancer Genome Atlas Research N. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature **2014**;507(7492):315-22 doi 10.1038/nature12965.

21. Hartigan JA WM. Algorithm AS 136: A k-means clustering algorithm. . Appl Stat **1979**;28(1).

22. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software **2010**;036(i11).

23. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS **2012**;16(5):284-7 doi 10.1089/omi.2011.0118.

24. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning **2003**;52(1):91-118 doi 10.1023/a:1023949509487.

25. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J*, et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst **2006**;98(4):262-72 doi 10.1093/jnci/djj052.

26. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution **2016**;8(1):28-36 doi 10.1111/2041-210X.12628.

27. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics **2016**;32(18):2847-9 doi 10.1093/bioinformatics/btw313.