Supplementary Materials

1. EXTENDED METHODS

Patient cohort, sample collection, DNA and RNA extraction. A total of 43 GAC patients with malignant ascites (PC) were enrolled in this study. The detailed clinical and histopathological characteristics of this cohort are described in the **Supplementary Table 1**. GACs were staged according to the American Joint Committee on Cancer Cancer Staging Manual (8th edition)[1, 2]. PCs were confirmed by cytologic examination. PC specimens were prospectively collected at The University of Texas MD Anderson Cancer Center (Houston, USA) from February 2016 under an approved protocol that requires written informed consent from each participant. The proportion of malignant cells was evaluated by a gastrointestinal pathologist (Y.W.) and were measured by morphological and immunofluorescence staining of EpCAM and YAP1. Among 43 patients, 39 had enough PC cells to obtain high quality genomic DNA (gDNA) and RNA. This project was approved by the institutional review committee and is in accordance with the policy advanced by the Helsinki Declaration of 1964 and later versions.

PC specimens were spun down 20 minutes at 2000g and pelleted cells were used for gDNA and RNA extraction. The pellets were washed with red blood cell lysis buffer consisted of 0.079g Ammonium Bicarbonate (NH4HCO3) and 6.1g Ammonium Chloride (NH4Cl) in 1 L distilled H2O if they contain blood cells. The gDNAs were isolated using the QIAamp DNA Mini Kit (Qiagen) and the total RNAs were isolated using the miRNeasy Mini Kit (Qiagen) following the manufacturer's instructions. Only samples passed sample intake quality check (gDNA: >200ng; total RNA: RNA integrity number-RIN >7) were further processed for DNA and RNA sequencing.

Whole-exome sequencing (WES). Whole-exome sequencing (WES) was performed on a total of 34 PC specimens, with 15 cases having matched gDNAs extracted from blood. Extracted gDNAs were submitted to the Sequencing and Microarray Core Facility (SMF) at UT MD Anderson Cancer Center for WES. Illumina compatible exome libraries were prepared from 200-500ng of Biorupter Ultrasonicator (Diagenode) sheared RNase-treated gDNA using the Agilent SureSelectXT Reagent Kit (Agilent Technologies). Libraries were prepared for capture with 6-10 cycles of PCR amplification, then assessed for size distribution on Fragment Analyzer using the High Sensitivity NGS Fragment Analysis Kit (Advanced Analyticals) and quantity using the Qubit dsDNA HS Assay Kit (ThermoFisher). Exon target capture was performed using the Agilent SureSelectXT Human All Exon V6 kit.

Following capture, index tags were added to the exon enriched libraries using seven cycles of PCR. The exon-enriched indexed libraries were then assessed for size distribution using the Agilent TapesStation and quantified using the Qubit dsDNA HS Assay Kit respectively. Libraries were multiplexed 8-9 samples per pool and the pools were quantified by qPCR using the KAPA Library Quantification Kit (KAPABiosystems). Each pool was sequenced at 76-bp paired end mode on the Illumina HiSeq4000 platform.

Whole-exome sequencing data processing and genotyping quality check. Raw output of Illumina exome sequencing data was processed using Illumina's Consensus Assessment of sequence And Variation (CASAVA) tool (v1.8.2) (see **URLs**) for demultiplexing and conversion to FASTQ format. The FASTQ files were aligned to the human reference genome (hg19) using BWA (v0.7.5)[3] with 3 mismatches (2 mismatches must be in the first 40 seed regions) for a 76-base sequencing run. The aligned BAM files were then subjected to mark duplication, realignment and base recalibration using Picard (v1.112) and GATK (v3.1-1) software tools[4]. The generated BAM files were then used for downstream analysis. Genotyping quality check was performed to rule out any possible sample swapping or contamination. Briefly, germline SNPs were called using Platypus (v0.8.1)[5]. Samples from the same patient were confirmed/identified by the percentage of genotyping-identity between them, which is defined by the fraction of identical germline alleles among the overlapping SNPs between the two samples. All samples in this study passed quality check and no sample swapping or contamination was detected.

Somatic mutation calling, filtering, and functional annotation. MuTect (v1.1.4)[6] was applied to identify somatic point mutations, and Pindel (v0.2.4)[7] was applied to identify small insertion and deletions (Indels). The MuTect and Pindel outputs were then run through our pipeline for filtering and annotation. Briefly, only MuTect calls marked as "KEEP" were selected and taken into the next step. For both substitutions and Indels, mutations with a low variant allelic fraction (VAF<0.02) or had a low total read coverage (<20 reads for tumor samples; <10 reads for germline sample) were removed. In addition, Indels that had an immediate repeat region within 25 base pairs downstream towards its 3' region were also removed. After that, common variants reported by ExAc (the Exome Aggregation Consortium), Phase-3 1000 Genome Project, or the NHLBI GO Exome Sequencing Project (ESP6500) (see **URLs**) with the minor allele frequency greater than 0.5% were further removed. The intronic mutations, mutations at 3' or 5' UTR or UTR flanking regions, silent mutations, in-frame small insertions and deletions were also removed. To evaluate the probability of a missense mutation being functionally deleterious, dbNSFP (v3.0)[8]

was applied to add prediction scores for all missense mutations from twelve commonly used functional prediction algorithms. A missense mutation that was called as “deleterious” or “damaging” by five or more algorithms were defined as “deleterious”.

Mutation load. The mutation load was calculated by first counting the number of nonsynonymous somatic mutations (nonsense, missense, splicing, stop gain, and stop loss substitutions and frameshift insertions and deletions) called in a sample and then normalized the numbers by the sample’s 20x target base coverage.

Mutation signature analysis. The Mutalisk tookit[9] was applied to quality-filtered somatic mutations (both the synonymous and non-synonymous substitutions), to statistically quantify the relative contribution of each of the 30 characterized COSMIC mutational signatures[10, 11]. Mutalisk was run on mutations called from each individual ascites sample and also on aggregated mutations from each histological subtype. The profiles of mutational signatures were then compared across histological subtypes.

Analysis of clonal and subclonal architecture of somatic mutations. The R package SciClone[12] was used to infer the clonal and subclonal architecture of mutations by clustering variants with similar mutant allele frequency together in an individual sample using the Bayesian binomial mixture model. Only mutations with a depth of 20x or greater coverage were selected. The allele-specific copy number profiles inferred by Sequenza were used to determine whether a segment has a LOH state, which was subsequently excluded as suggested[12].

DNA copy number analysis. DNA copy number analysis was conducted using an in-houses application ExomeLyzer[13] followed by CBS segmentation[14]. The copy number segmentation files were loaded to IGV[15] for visualization. R package “CNTools” (v1.24.0) was used to identify copy number gains (log2 copy ratios > 0.3) or losses (log2 copy ratios < −0.3) at the gene level. The burden of copy number gain or loss was then calculated as the total number of genes with copy number gains or losses per sample. The fraction of genome altered was calculated as the proportion of the genome with copy number gains or losses against the total length of genome with copy number profiled. R package Sequenza (Version 2.1.2)[16] was used to derive ploidy of the genome and allele-specific copy number alterations. Sequenza was only applied to matched tumor-normal pairs (n=15) in this study. Presence of whole genome doubling event was inferred using a modified version of ABSOLUTE algorithm[17] as previously described[18].

Whole transcriptome sequencing (RNA-seq). Whole transcriptome sequencing (RNA-seq) was performed on 39 PC specimens. Among them, 21 specimens also had WES data. Extracted total RNA were submitted to the SMF Core at UT MD Anderson Cancer Center for RNA-seq. Stranded mRNA libraries were prepared using the KAPA Stranded mRNA-Seq Kit (Kapa Biosystems). Briefly, 250ng of total RNA was captured using magnetic Oligo-dT beads. After bead elution and clean-up, the resultant PolyA RNA was fragmented using heat and magnesium. First strand synthesis was performed using random priming followed by second strand synthesis with the incorporation of dUTP into the second strand. The ends of the resulting double stranded cDNA fragments were repaired, 5′-phosphorylated, 3'-A tailed and Illumina-specific indexed adapters were ligated. The products were purified and enriched for full length library with 8 –9 cycles of PCR. The libraries were quantified using the Qubit dsDNA HS Assay Kit and assessed for size distribution using the 4200 Agilent Tapestation (Agilent Technologies), then multiplexed and sequenced with the HiSeq4000 sequencer using the 75-bp paired end mode.

RNA-seq data processing and quality check. RNA-seq FASTQ files were processed through FastQC (v0.11.5) (see **URLs**), a quality control tool to evaluate the quality of sequencing reads at both the base and read levels and RNA-SeQC (v1.1.8)[19] to generate a series of RNA-seq related quality control metrics. All samples passed quality check for this study. STAR 2-pass alignment (v2.5.3)[20] was performed with default parameters to generate RNA-seq BAM files.

Identification of differentially expressed genes and enriched signaling pathways. HTSeq-count (v0.9.1)[21] tool was applied to aligned RNA-seq BAM files to count for each gene how many aligned reads overlap with its exons. The HTSeq raw count data were then processed by Deseq2 (v3.6)[22] software to identify differentially expressed genes (DEGs) across histopathological subtypes. A cut-off of gene expression fold change of >=2 or <=-0.5 and a FDR q-value of <=0.01 was applied to select the most differentially expressed genes. A ranked list of genes was generated based on DeSeq2 FDR q-values for all coding genes and processed by Gene Set Enrichment Analysis (GSEA)[23] against the curated gene sets from Molecular Signature Database (MSigDB)[24] to identify significantly enriched signaling pathways. A cut-off of FDR q-value of <=0.05 was applied to select the most significantly enriched signaling pathways.

Gene expression normalization for hierarchical clustering. RNA-Seq gene expression raw read counts generated from HTSeq-count (v0.9.1)[21] were normalized into fragments per kilobase of transcript per million

mapped reads (FPKM) using the RNA-seq quantification approach suggested by the bioinformatics team of NCI Genomic Data Commons (GDC) (see **URLs**). Briefly, FPKM normalizes read count by dividing it by the gene length and the total number of reads mapped to protein-coding genes using a calculation described below:

$$FPKM = \frac{RC_g * 10^9}{RC_{pc} * L}$$

RC_g, number of reads mapped to the gene; RC_{pc}: number of reads mapped to all protein-coding genes; L, length of the gene in base pairs (calculated as the sum of all exons in a gene). The FPKM values were then log2-transformed for further downstream processes.

Deconvolution of the cellular composition of ascites. Two deconvolution approaches were applied–the CIBERSORT algorithm[25] to estimate the relative cellular fraction of 22 immune cell types, and the R package software MCP-counter[26] to produce the absolute abundance scores for 8 major immune cell types ($CD3^+$ T cells, $CD8^+$ T cells, cytotoxic lymphocytes, NK cells, B lymphocytes, monocytic lineage cells, myeloid dendritic cells, and neutrophils), endothelial cells, and fibroblasts. The log2-transformed FPKM expression matrix of ascites samples were used as the input data source for both algorithms, and the LM22 leukocyte gene signature was used as the input gene signature for CIBERSORT. The deconvolution profiles were then hierarchically clustered and compared across histology types.

Statistical analysis. In addition to the algorithms described above, all other basic statistical analysis was performed in the R statistical environment. All statistical tests performed in this study were two-sided. Statistical significance of differences observed between histological subtypes was determined by non-parametric Mann-Whitney U test when comparing continuous variables, and the Fisher's Exact test when comparing frequencies. The Spearman's rank correlation coefficient was calculated to assess the association between two continuous variables. To control the false discovery rate (FDR) and correct p-values for multiple testing, we apply Benjamini-Hochberg method[27] and a FDR adjusted p-value (or q-value) < 0.05 is considered as statistically significant.

2. URLs.

FastQC, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/; ExAc, http://exac.broadinstitute.org; Phase-3 1000 Genome Project, http://phase3browser.1000genomes.org/Homo_sapiens/Info/Index; ESP6500,

http://evs.gs.washington.edu/EVS/; CASAVA, http://support.illumina.com/sequencing/sequencing_software/casava.html; RNA-seq quantification approach by GDC, https://gdc.cancer.gov/about-data/data-harmonization-and-generation/genomic-data-harmonization/high-level-data-generation/rna-seq-quantification.

3. Supplemental table legend:

Supplemental Table 1. Characteristics of 44 Ascites samples with gastric adenocarcinoma for whole exome sequencing and RNA sequencing. All clinical information was included-Sex, Age, Location of primary tumor, differentiation state, Lauren classification, Signet Ring cell state, treatment before collection and response for treatment.

4. Supplemental References:

1. Amin MB, Edge S, Greene F, et al. AJCC cancer staging manual. 8th ed. New York: Springer 2017.
2. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin 2017;67:93-99.
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.
4. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491-8.
5. Rimmer A, Phan H, Mathieson I, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet 2014;46:912-918.
6. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213-9.
7. Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25:2865-71.
8. Liu X, Wu C, Li C, et al. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat 2016;37:235-41.

9. Lee J, Lee AJ, Lee JK, et al. Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. Nucleic Acids Res 2018.
10. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415-21.
11. Alexandrov LB, Jones PH, Wedge DC, et al. Clock-like mutational processes in human somatic cells. Nat Genet 2015;47:1402-7.
12. Miller CA, White BS, Dees ND, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol 2014;10:e1003665.
13. Zhang J, Fujimoto J, Zhang J, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science 2014;346:256-9.
14. Olshen AB, Venkatraman ES, Lucito R, et al. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5:557-72.
15. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178-92.
16. Favero F, Joshi T, Marquard AM, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol 2015;26:64-70.
17. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 2012;30:413-21.
18. Dewhurst SM, McGranahan N, Burrell RA, et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discov 2014;4:175-185.
19. DeLuca DS, Levin JZ, Sivachenko A, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 2012;28:1530-2.
20. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.
21. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166-9.

22. Love MI, Anders S, Kim V, et al. RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000Res 2015;4:1070.

23. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.

24. Liberzon A, Birger C, Thorvaldsdottir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1:417-425.

25. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453-7.

26. Becht E, Giraldo NA, Lacroix L, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol 2016;17:218.

27. Benjamini Y, Drai D, Elmer G, et al. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001;125:279-84.