Data S1

Supplementary methods

Target gene sequencing and mutational analysis

Sample preparation, DNA isolation and sequencing (1). Genomic DNA from formalin-fixed paraffin-embedded (FFPE) sections was extracted with the QIAamp DNA FFPE Tissue kit (cat. no. 56404; Qiagen Inc.). FFPE tissue size was >0.5x0.5 cm, thickness was >0.6 μm and three FFPE sections were used. Sequencing libraries were prepared using the KAPA Hyper Prep Kit (cat. no. KK8505/07962371001; KAPA Biosystems; Roche Diagnostics GmbH) according to the manufacturer's instructions for different sample types. Customized xGen lockdown probes (Integrated DNA Technologies, Inc.) targeting 475 tumor-related genes were used for hybridization enrichment. DNA was quantified by dsDNA HS Assay Kit (cat. no. Q32854; Thermo Fisher Scientific, Inc.) according to the manufacturer's recommendations (Table SI). The capture reaction was performed with Dynabeads M-270 (cat. no. 35354D; Thermo Fisher Scientific, Inc.) and the IDT 96 rxn xGen Lockdown Reagents (Hybridization and Wash kit) (cat. no. 1072281; Integrated DNA Technologies, Inc.) according to manufacturers' protocols. Captured libraries were on-bead PCR amplified with Illumina p5 (5'-AATGAT ACGGCGACCACCGA-3') and p7 (5'-CAAGCAGAAGAC GGCATACGAGAT-3') primers in KAPA HiFi HotStart ReadyMix (KAPA Biosystems; Roche Diagnostics GmbH), followed by purification using Agencourt AMPure XP beads (cat. no. A63987; Beckman Coulter, Inc.). Libraries were quantified by qPCR using the KAPA Library Quantification kit (KAPA Biosystems; Roche Diagnostics GmbH). Library fragment size was determined by Bioanalyzer 2,100 (Agilent Technologies, Inc.). The target-enriched library was then sequenced on the HiSeq4000 NGS platform (Illumina, Inc.) according to the manufacturer's instructions. The mean coverage sequencing depth was 1,000X for FFPE tissues. The aforementioned sequencing was performed by Nanjing Geneseeq Technology, Inc.

Analysis of DNA sequences (2-4). Sequencing data was processed as previously described (1). In brief, the data was first demultiplexed and subjected to FASTQ file quality control to remove low quality data or N bases. Qualified reads were mapped to the reference human genome GRCh37/hg19 using the BWA Aligner V.0.7.12 with BWA-MEM algorithm3 and default parameters to create SAM files4. Picard V.1.119 (Picard toolkit) was used to convert SAM files to compressed BAM files, which were then sorted according to chromosome co-ordinates. The Genome Analysis Toolkit5 (V.3.4-0; Broad Institute) was used to locally realign the BAM files at intervals with insertions/deletion (indels) mismatches and recalibrate base quality scores of reads in BAM files. VarScan26 was employed for the detection of single-nucleotide variations (SNVs) and insertion/deletion mutations. The resulting mutation lists were further filtered through an internally collected list (1,000 normal samples) of recurrent artifacts on the same sequencing platform. SNVs and indels were further filtered with the following parameters: Minimum read depth, 20; minimum base quality, 15; minimum variant supporting reads, 5; variant supporting reads mapped to both strands; strand bias ≤10%; if present in >1% population frequency in the 1,000 g or ExAC database; and through an internally collected list of recurrent sequencing errors using a normal pool of 100 samples. Copy number variations (CNVs) were analyzed with CNVkit7 Depth ratios of >2 and <0.6 were considered as CNV gain and CNV loss, respectively.

Bulk RNA sequencing. Frozen tissues (-80˚C) were weighed and homogenized in RLT lysis buffer, and nucleic acids were extracted using the AllPrep DNA/RNA Mini Kit (cat. no. 80204; Qiagen Inc.) according to the manufacturer's instructions. RNA was eluted in nuclease-free water and DNA in 0.5X Buffer ethidium bromide. Approximately 100 ng of fresh frozen RNA per sample was used for RNA library construction using the KAPA RNA Hyper library prep kit (Roche Diagnostics GmbH) according to the manufacturer's instructions. Customized adapters with unique molecular indexes (Integrated DNA Technologies, Inc.) and sample-specific dual-indexes primers (Integrated DNA Technologies, Inc.) were added to each library. The loading concentration of the libraries was measured with Qubit (Thermo Fisher Scientific, Inc.) and quality measured by TapStation Genomic DNA Assay (Agilent Technologies, Inc.). Equal amounts of each RNA library (~500 ng) were pooled for hybridization capture with IDT Whole Exome Panel V1 (Integrated DNA Technologies, Inc.) using a customized capture protocol modified from NimbleGen SeqCap Target Enrichment system (Roche Diagnostics GmbH). The captured DNA libraries were then sequenced on an Illumina HiSeq4000 with paired end reads (2Å~100 bp), at 50 million reads/sample.

Differentially expressed genes were analyzed using edgeR. Differentially expressed genes were defined by an adjusted P-value of <0.05 and an absolute fold-change of >1. Gene set enrichment analysis was performed to calculate the normalized enrichment score. Immune deconvolution was performed with CIBERSORT (immunedeconv v2.1.0) to estimate the relative fraction of 22 immune subsets (5).

Clonal evolution analysis (6,7). Variant allele frequency and the copy number of the genomic region containing the mutation and an estimate of tumor content were used as the input to PyClone. The proportion of cell clones with specific mutation (cellular prevalence) was then yielded. The optimal tree solutions were obtained with the iterative version of Citup (http://sourceforge.net/projects/citup/). The fish plot and phylogenetic tree were then depicted with Timescape.

Phylogenetic trees of clones (subclonal hierarchies) were reconstructed by SubClonal Hierarchy Inference from Somatic Mutations (SCHISM) (v.1.1.2) with estimated somatic mutation clusters with cancer cell fractions (CCFs) across tumor samples from PyClone as inputs. SCHISM was run in the Sequential Mode, with the K-means algorithm. Default settings were used for other parameters as follows: Generation count, 50; generation size, 1,000; random object fraction, 0.2; mutation probability, 0.9; and fitness coefficient, 5.0. Mutations falling in the truncal clone were classified as clonal. If the truncal clone consisted of only one single mutation and the CCF of its direct descendent clone was >80% of its CCF, mutations falling in this descendent clone was also considered clonal. All the other mutations were classified as subclonal.

Immunohistochemistry. Human tissue samples were obtained from patients diagnosed with colorectal cancer, following approval by the Institutional Review Board and informed patient consent. Tissues were fixed in 10% neutral buffered formalin for 24 h, embedded in paraffin and sectioned at a 4-μm thickness. For hematoxylin and eosin staining, the tissue sections were deparaffinized, rehydrated and stained with hematoxylin for 5 min, followed by eosin for 2 min. For immunohistochemistry (IHC), the sections were incubated with primary antibodies against Ki-67 (1:200 dilution), CD20 (1:200 dilution), MYC (1:200 dilution), p53 (1:200 dilution), granzyme B (1:200 dilution), CD56 (1:200 dilution), CD3 (1:200 dilution), BCL2 (1:200 dilution) and CD10 (1:200 dilution) overnight at 4˚C, followed by incubation with HRP-conjugated secondary antibodies for 1 h at room temperature. All stained sections were examined under a light microscope (Eclipse 80i; Nikon Corporation) at x200 magnification. Images were captured using a digital camera (DS-Fi3; Nikon Corporation) and analyzed using ImageJ software (version 1.53; National Institutes of Health).

EBER in situ hybridization. EBER gene expression was detected using a digoxin-labeled RNA probe. Tissue samples were fixed with 4% paraformaldehyde, paraffin-embedded and cut into 5-μm sections. Sections were dewaxed, hydrated and treated with proteinase K for 10 min. Hybridization was performed at 55˚C for 16 h with a probe concentration of 200 ng/ml. After hybridization, the sections were washed with 2X Saline Sodium Citrate buffer (SSC) and 0.1X SSC. Signals were detected by alkaline phosphatase-coupled anti-digoxigenin antibody with NBT/BCIP color development. Positive controls used tissues known to express the target gene, and negative controls used sections hybridized without probes. Images were acquired under a light microscope (Eclipse 80i) and analyzed with ImageJ software.

References

1. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ and Prins P: Sambamba: Fast processing of NGS alignment formats. Bioinformatics 31: 2032-2034, 2015.
2. Wang X, Gao Y, Shan C, Lai M, He H, Bai B, Ping L, Rong Q, Ai R, Wen L, *et al*: Association of circulating tumor DNA from the cerebrospinal fluid with high-risk CNS involvement in patients with diffuse large B-cell lymphoma. Clin Transl Med 11: e236, 2021.
3. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA: The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303, 2010.
4. Talevich E, Shain AH, Botton T and Bastian BC: CNVkit: Genome-Wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol 12: e1004873, 2016.
5. Chen B, Khodadoust MS, Liu CL, Newman AM and Alizadeh AA: Profifiling tumor infifiltrating immune cells with CIBERSORT. Methods Mol Biol 1711: 243-259, 2018.
6. Li HW: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013: 1303.3997, 2013.
7. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L and Wilson RK: VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568-576, 2012.