Landscape of somatic alterations in large-scale solid tumors from an Asian population

Kai Wang (wangk@origimed.com)
OrigiMed., Shanghai
Qun Wu
Department of Hepatobiliary Surgery, Affiliated Hospital of Qingdao University, Qingdao
Herui Yao
Sun Yat-sen Memorial Hospital of Sun Yat-sen University https://orcid.org/0000-0001-5520-6469

Article

Keywords: next-generation DNA sequencing, somatically altered genes, tumor mutational burden

Posted Date: November 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-916644/v1

Version of Record: A version of this preprint was published at Nature Communications on July 23rd, 2022. See the published version at https://doi.org/10.1038/s41467-022-31780-9.

Abstract

Extending the benefits of tumor molecular profiling for all cancer patients will require comprehensive analysis of tumor genomes across distinct patient populations world-wide. In this study, we performed deep next-generation DNA sequencing (NGS) from tumor tissues and matched blood specimens from over 10,000 patients in China by using a 450-gene comprehensive assay, developed and implemented under international clinical regulations. We performed a comprehensive comparison of somatically altered genes, the distribution of tumor mutational burden (TMB), gene fusion patterns and the spectrum of various somatic alterations between Chinese and American patient populations. In total, 64% of cancers from Chinese patients in this study were found to have clinically actionable genomic alterations, which may affect clinical decisions related to targeted therapy or immunotherapy. These findings describe the similarities and differences between tumors from Chinese and American patients, providing valuable information for personalized medicine.

Main Text

Cancer morbidity and mortality remain a major challenge to public health in China, with over two million cancer deaths per year in China[1,2]. In recent years, precision oncology has enabled individual diagnosis, prognosis and treatment based on increasingly accurate and high-resolution molecular stratification of cancers, largely focused on genomically targeted therapies[3,4]. Notably, patient ethnicity can also be a factor in cancer diagnostics and treatment, since differences in cancer mutations exist between tumor patient populations of various ethnicities[5-7].

To explore the genomic landscape of Chinese patients with solid tumors as encountered in clinical practice, we collected a total of 11,553 tumor specimens and matched peripheral blood specimens from 11,553 individuals encompassing 25 principal tumor types and >100 tumor subtypes. After excluding samples (n = 1359) with insufficient tumor content or DNA yield or subsequent technical failure (**Supplementary Fig. 1a**), we successfully sequenced 10,194 (88%) tumor samples. Summaries of the clinical characteristics of the patients' specimens included in the study, and the median sequencing target coverage of samples are presented in **Supplementary Table 1-2** and **Supplementary Fig. 2-3**. A total of 31 ethnicities were presented in our cohort with Han being the most frequent (92%, 9382/10,194). The majority of patients in this study were from eastern and southern provinces in China ("East China" and "South China" in Wikipedia) (41% and 29%, respectively). In terms of tumor stage, 55% (5,652/10,194) of patients had advanced-stage cancers (stage III/IV) while 35% (3,579/10,194) had early-stage cancers (pre-cancers or stage I/II). In our cohort, majorities (76%) of patients were treatment-naïve and the ratio of patients who had received treatment was 16% (**Supplementary Fig. 1b**). The major tumor types were non-small cell lung cancer (NSCLC; 20%), colorectal carcinoma (CRC; 12%), liver hepatocellular carcinoma (LIHC; 11%), gastric cancer (GC; 8%), esophageal carcinoma (ESCA; 6%), soft tissue sarcoma (STS; 6%), intrahepatic cholangiocarcinoma (ICC; 5%), pancreatic cancer (PAC; 5%), extrahepatic cholangiocarcinoma (ECC; 3%), and breast carcinoma (BRCA; 3%) (**Fig. 1a**). In general, the distribution of

these predominant tumor types such as liver cancer (LIHC, ICC, and ECC) and lung cancer (NSCLC and SCLC) represented the distribution of tumors encountered in clinical practice in China[1].

Based on an NGS-based assay with a validated 450-gene panel[8], we detected 80,703 single nucleotide variants (SNVs) and insertions and deletions (indels), 19,192 truncations, 17,779 gene amplifications, 1,688 gene homozygous deletions, and 3,111 gene fusions/ rearrangements in the 10,194 cases. We only focused on somatic alterations within tumors in this study, and no germline genetic data. Analysis of significantly mutated cancer related genes in solid tumors found the most frequently altered genes to be *TP53* (58% of cases), *KRAS* (18%), *TERT* (14%), *EGFR* (13%), *APC* (13%), *CDKN2A* (12%), and *PIK3CA* (11%). The most common mutations were $KRAS^{G12}$, $EGFR^{L858}$, and $TP53^{R273}$ (**Fig. 1b; Supplementary Table 3**); of note, *EGFR* and *KRAS* represented obviously pairwise co-occurring alterations with SNV/indel and copy number variation (CNV) (**Supplementary Fig. 4).** Subsequent analysis of CNV showed high frequencies of *CDKN2A/B* deletion, *SMAD4* deletion, *ERBB2* amplification, *EGFR* amplification and *MYC* amplification in metastatic samples, and chromosome 11q13.3 (*CCND1/ FGF3/ FGF4/ FGF19*) amplification in primary samples at the pan-cancer level; meanwhile *ERBB2* amplification and chromosome 11q13 amplification were respectively enriched in breast cancer (BRCA) (24% vs. 2%; FDR = 7.645E-105) and ESCA (43% vs. 4%; FDR = 3.553E-301), compared to other tumor types (**Supplementary Fig. 5**; **Supplementary Table 4-5**). In addition, we sought to investigate the features of gene fusions in solid tumor and identified a total of 513 fusion events, including 31 driver genes in our cohort. As shown in **Fig. 1c**, fusion events in genes such as *ALK* (n = 139), *ROS1* (n = 51), *RET* (n = 50), *FGFR2/ 3* (n = 50), *NTRK1/ 3* (n = 30) and *BRAF* (n = 12) occurred widely across tumor types, while others such as *EWSR1* and *TFE3* were enriched in certain tumor types (sarcomas [soft tissue sarcoma or STS, and bone sarcoma] and KIRC, respectively); *PRKACA* fusions were only detected in a specific tumor type (LIHC, subsequent diagnosis as fibrolamellar hepatocellular carcinoma [FL-HCC]). Moreover, new fusion partners for driver genes were identified; multiple novel fusions were seen in kinase genes, including *GRIK2-ROS1*, *PARP12-BRAF*, *KIF13B-MET*, and *LRRC28-NTRK3*; novel fused exons were seen in *KIF5B-ALK* and *EML4-ALK* compared to the Quiver database (http://quiver.archerdx.com/) (**Fig. 1d**, **Supplementary Table 6**, and **Supplementary Fig. 6-7**).

To reveal further somatic alterations associated with clinical characteristics in Chinese cancer patients, we implemented an integrative analysis across the tumor-type distribution of genomic profile and six clinical features including age, gender, tumor stage, smoking history (only in NSCLC, SCLC and HNC), treatment and sample type (primary vs. metastatic/recurrent) (**Supplementary Table 7; Fig. 2a-b**). In general, clinical feature-associated genomic differences were observed distributed in CRC and NSCLC. In CRC, the numbers of different mutated genes were respectively 270 and 100 in younger and early-stage patients as compared to older and advanced-stage patients, which could be consistent with the significantly high ratio of hypermutated subtypes, such as microsatellite instability-high (MSI-H) and *POLE*-associated CRC (with microsatellite stability [MSS], high mutation burden and an inactive *POLE* mutation) in younger and early-stage CRC (**Supplementary Fig. 8**, $FDR < 0.05$). In NSCLC, the mutated frequency of genes was markedly affected by gender and smoking history; of note, gender and smoking

history were not independent factors in our cohort, because the majority of nonsmokers were female. It was found that female nonsmokers with early-stage NSCLC harbored more mutations in *EGFR*; while male smokers with advanced-stage NSCLC were characterized by more mutations in *TP53*, *CDKN2A*, *PIK3CA* and *KRAS*, consistent with recent reports[9]. Moreover, younger female gastric cancer patients had more *CDH1* mutations; in contrast, older gastric cancer patients tended to have more mutations in *TP53*, *NOTCH1* and *FAT4*. In addition, younger LIHC, KIRC and bone sarcoma patients harbored respectively *TP53*, *TFE3* and *VEGFA* mutations, while older LIHC, HNC and STS patients had respectively *CTNNB1*, *TERT* and *TP53* mutations (**Fig. 2b**, FDR < 0.05).

To assess the characteristics of cancer genomes from Chinese patients in a global context, we made a comparison of genomic alterations with the largest published cancer genomic study: the Memorial Sloan Kettering Cancer Center (MSK) IMPACT study[10], including 10,366 cases, mostly advanced cancer specimens. A total of 266 correspond genes were compared in 15 comparable advanced-stage tumor types between OM cohort (aOM, n = 4699) and MSK cohort (n = 6161). To limit the bias of comparisons, we subdivided NSCLC of the two cohorts into lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Overall, only 20 tumor type: gene pairs presented significant differences in the frequency of gene variants between the aOM cohort and the MSK cohort (FDR < 0.05) (**Fig. 2c-d** and **Supplementary Table 8**), suggesting frequencies of most significantly mutated genes and the tumor-type distribution in aOM cohort were highly consistent with in the MSK cohort, such as CRC:*APC* (72% vs. 77%, FDR = 0.27) and SCLC:*RB1* (81% vs. 72%, FDR = 1). The significant differences between two cohorts were mainly found in lung adenocarcinoma and hepatobiliary tumors, such as LUAD: *EGFR*, ICC: *KRAS* and ECC: *IDH1*. At the same time variants of *TP53* gene exhibited widely differences in multiple tumor types. Furthermore, cohorts were independently and significantly related to most of different tumor type: gene pairs in logistic regression analysis of several clinical features, including primary/metastasis/recurrent tumor specimens, sampling method, gender and smoke (**Supplementary Table 8**). Moreover, several gene fusions and CNVs also showed differences between the aOM cohort and the MSK cohort. For example, *ALK* fusions in LUAD (**Supplementary Table 6**) occurred more frequently in the aOM cohort than that in the MSK cohort (9% vs. 3%, P = 3.913E-10); while the incidences of *ROS1* fusions (3% vs. 2%, P = 0.38) and *RET* fusions (2% vs. 2%, P = 0.65) in LUAD were similar between the two cohorts.

To further confirm the similarities and differences between the OM and MSK studies observed in advanced cancers, we also compared the aOM data with genomic data of advanced-stage cases from The Cancer Genome Atlas studies (aTCGA). Because of heterogeneous methodologies (including detecting platform, algorithm and report criteria of variants), substitutions, indels and truncations mutations were considered in the comparison. In 9 comparable tumor types and 266 genes, we identified a total of 26 tumor type: gene pairs with significant differences between the aOM cohort (n = 3505) and the aTCGA cohort (n = 2065) (FDR < 0.05) (**Fig. 2e-f** and **Supplementary Table 9),** of which 7 different tumor type: gene pairs presented consistently changed trends with those in the comparison between the aOM cohort and the MSK cohort, including higher frequencies of LIHC: *TP53*, BRCA: *TP53*, LUAD: *EGFR* and lower frequencies of LUAD: *KEAP1*, LUAD: *KRAS*, HNC: *TP53*, HNC: *PIK3CA* in the aOM cohort,

compared with other two cohorts. Altogether, these multiple comparisons revealed at the greatest extent the similarity and distinctive of genomic alteration across these cohorts.

In addition to targeted therapy, the recent clinical success of immune checkpoint blockade[11-14] makes the comparison of immunotherapy related mutations and signatures across cancers from patients in different countries another important question. Hence, we analyzed the distribution of tumor mutational burden (TMB) within tumor types. Even though an algorithm to evaluate TMB in routine clinical practice has not yet reached a consensus[15], an individual TMB has been shown to predict patient outcome after immunotherapy[11-14]. Here, we identified as TMB high (TMB-H) and TMB low (TMB-L) according to the TMB-high status definition from KEYNOTE-158 study (the value ≥ 10 or not)[14]. As shown in **Fig. 3a** and **Supplementary Table 10**, median TMB values in most tumor types in the aOM cohort were different compared with the MSK cohort (**Supplementary Fig. 9**). Overall, the whole pattern of TMB distribution in the aOM cohort was similar to that in the MSK cohort, characterized by a "tail" that includes 119 samples with TMB ≥ 40 (**Fig. 3b**). We further analyzed the distribution of 186 samples harboring MSI-H in our cohort and found that the overall proportion of patients with MSI-H was 2% and was mainly in CRC (55%, 102/186) (**Fig. 3c**). Furthermore, because MSI-H and TMB-H have recently been recognized as biomarkers for response to immune checkpoint blockades (anti-PD-1/PD-L1)[11,14], we evaluated the combined association of TMB and MSI with PD-L1 expression evaluated by immunohistochemical (IHC) staining in 2,723 tumors of OM cohort. The overall proportion of samples with at least one of MSI-H, TMB-H, or PD-L1 positive was 30.3% (891/2,723) and the proportion of such samples is highest in SCLC (48%; 24/50), followed by NSCLC (46%; 298/648), and ESCA (34%; 80/235) (**Fig. 3d, Supplementary Fig. 10**), suggesting the possibility of a high proportion of Chinese patients with lung cancer benefitting from immunotherapy.

In addition, recent evidence has suggested somatic amplification in the gene for programmed cell death ligand 1 (PD-L1/*CD274*) as a response biomarker to immunotherapy in solid tumors, even in the absence of MSI-H, PD-L1 overexpression or TMB-H[16]. Herein, we identified a total of 85 (1%) tumors with *CD274* amplification (copy number ≥ 6) in the OM cohort, a proportion consistent with a previous study[17](**Supplementary Fig. 11a**). Furthermore, in 30 evaluable tumors with *CD274* amplification tested for PD-L1 expression, the PD-L1 positive rate was 70% (**Supplementary Fig. 11b**). Subsequently, we also examined the mutational landscape of the 85 samples with *CD274* amplification and found the co-occurrence of *CD274* amplification with adjacent *PDCD1LG2* and *JAK2* amplification (89% and 82% respectively), which are nearby genes in chromosome 9p24.3-9p22.2, associated with advanced stage and poorer outcome[17]. A high frequency of *TP53* mutations (78%) was also observed in these tumors (**Supplementary Fig. 11c**).

Finally, to assess the potential clinical impact of the somatic alterations found in our cohort, we used the MSK criteria[10,18] to systematically evaluate actionable variants identified in solid tumors from Chinese patients in our cohort, using the OncoKB (http://oncokb.org/, v3.6) knowledge base. Patients who harbored potential actionable variants in their tumors were classified into different evidence levels of

predictive biomarkers. As shown in **Fig. 4a** and **Supplementary Fig. 12a**, 64% of patients (n = 6, 498 harbored at least one genomic variants with a variable highest level of clinical evidence (Level 1, 32%; Level 2, 1%; Level 3A, 1%; Level 3B, 13%; Level 4, 16%), including TMB-H as a predictive biomarker of response to immunotherapy[14]. To further investigate whether the remaining 3, 696 patients without OncoKB Level 1-4 variants in the OM cohort had an actionable biomarker, we analyzed PD-L1 expression. We found that 4% of these patients exhibited at least PD-L1 positive (**Supplementary Fig. 12b**), suggesting those patients could be candidates for treatment with immune checkpoint inhibitors even if their tumors did not meet the criteria for Level 1-4. A higher ratio of Level 1 was observed mainly in NSCLC, BRCA, SCLC, and UC, compared to that in other cancer types (**Fig. 4b**). Level 1 was predominantly represented by TMB-H and EGFR mutations in NSCLC, including EGFR L858R (20%; the ratio of variant to the total number of samples with the tumor type), exon 19 deletion (19%) and G719 (3%) mutations. Others included ALK (7%) fusions in NSCLC, PIK3CA mutations (31%), and ERBB2 amplification (24%) in BRCA and MSI-H in CRC (8%) (**Fig. 4c**). In terms of population-level mutation of actionable variants, *KRAS*, *EGFR* and *PIK3CA* substitution/indels, *ERBB2* amplification, and *ALK* fusions were most common, which was consistent with reports in the MSK cohort (**Supplementary Fig. 12a**). Interesting, in NSCLC, TMB-H is negatively associated with fusion positive (3% vs. 13% fusion frequency in TMB-H cohort and TMB-L cohort, respectively, P = 1.31E-11), mostly from *ALK* gene; in contrast, MSI-H shows positive association with fusion positive (6% vs. 1% fusion frequency in MSI-H cohort and MSS cohort, respectively, P = 0.04), mostly from *NTRK* gene, which hint that clinical benefit of patients from the combination of fusion-based targeted therapy and immunotherapy is different in different types of cancers and the finding requires more studies to confirm in the future. All these findings suggested the relevance of treatment to the mutational landscape of Chinese tumor patients.

In conclusion, we report herein the somatic mutation landscape of over 10,000 solid tumors in Chinese patients. To our knowledge, this is the largest and most comprehensive mutational landscape analysis of solid tumors in an Asian population. This report provides a highly reliable dataset and resource for cancer medicine. More importantly, this population-level comparative analysis has comprehensively revealed similarities and differences between somatic alterations and actionable variants between Chinese and other ethnic populations with solid tumors, and has an important implication for the selection of patients to clinical trials with molecularly targeted therapies.

Online content

Data summary in this study is accessible on cbioportal (https://www.cbioportal.org/study/summary?id=pan_origimed_2020), and software in our FTP (http://ftp.origimed.com).

References

1. Chen, W., *et al.* Cancer statistics in China, 2015. *CA Cancer J Clin* **66**, 115–132 (2016).
2. Bray, F., *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394–424 (2018).

3. Senft, D., Leiserson, M.D.M., Ruppin, E. & Ronai, Z.A. Precision Oncology: The Road Ahead. *Trends Mol Med* **23**, 874–898 (2017).
4. Berger, M.F. & Mardis, E.R. The emerging clinical relevance of genomics in cancer medicine. *Nat Rev Clin Oncol* **15**, 353–365 (2018).
5. D'Angelo, S.P., *et al.* Incidence of EGFR exon 19 deletions and L858R in tumor specimens from men and cigarette smokers with lung adenocarcinomas. *J Clin Oncol* **29**, 2066–2070 (2011).
6. De Roock, W., De Vriendt, V., Normanno, N., Ciardiello, F. & Tejpar, S. KRAS, BRAF, PIK3CA, and PTEN mutations: implications for targeted therapies in metastatic colorectal cancer. *Lancet Oncol* **12**, 594–603 (2011).
7. Grenade, C., Phelps, M.A. & Villalona-Calero, M.A. Race and ethnicity in cancer therapy: what have we learned? *Clin Pharmacol Ther* **95**, 403–412 (2014).
8. Cao, J., *et al.* An Accurate and Comprehensive Clinical Sequencing Assay for Cancer Targeted and Immunotherapies. *Oncologist* **24**, e1294-e1302 (2019).
9. Chen, J., *et al.* Genomic landscape of lung adenocarcinoma in East Asians. *Nat Genet* **52**, 177–186 (2020).
10. Zehir, A., *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* **23**, 703–713 (2017).
11. Samstein, R.M., *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet* **51**, 202–206 (2019).
12. Reck, M., *et al.* Nivolumab plus ipilimumab versus chemotherapy as first-line treatment in advanced non-small-cell lung cancer with high tumour mutational burden: patient-reported outcomes results from the randomised, open-label, phase III CheckMate 227 trial. *Eur J Cancer* **116**, 137–147 (2019).
13. Wang, F., *et al.* Safety, efficacy and tumor mutational burden as a biomarker of overall survival benefit in chemo-refractory gastric cancer treated with toripalimab, a PD-1 antibody in phase Ib/II clinical trial NCT02915432. *Ann Oncol* **30**, 1479–1486 (2019).
14. Marabelle, A., *et al.* Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol* **21**, 1353–1365 (2020).
15. Truesdell, J., Miller, V.A. & Fabrizio, D. Approach to evaluating tumor mutational burden in routine clinical practice. *Transl Lung Cancer Res* **7**, 678–681 (2018).
16. Roemer, M.G., *et al.* PD-L1 and PD-L2 Genetic Alterations Define Classical Hodgkin Lymphoma and Predict Outcome. *J Clin Oncol* **34**, 2690–2697 (2016).
17. Goodman, A.M., *et al.* Prevalence of PDL1 Amplification and Preliminary Response to Immune Checkpoint Blockade in Solid Tumors. *JAMA Oncol* **4**, 1237–1244 (2018).
18. Chakravarty, D., *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**(2017).

Declarations

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 81872492, 81871886, 81572355), Shanghai Pujiang Program (No. 15PJD007), Natural Science Foundation of Guangdong Province (No. 2017A030313474) and Guangdong Science and Technology Department (No. 2017A050501015, 2017B030314026) and Key Task Project of Tianjin Health and Family Planning Commission (No. 16KG128). The study was supported in part by the Wu Jie Ping Medical foundation program (No. 320. 6750. 17).

Author contributions

L.W. participated in the writing of study documentation, analyzed data and wrote the manuscript. H.Y. designed and, coordinated the acquisition, distribution and quality evaluation of clinical samples, and wrote the manuscript. H.C. was responsible for data collection, provided statistical analysis and wrote the manuscript. A.W., K.G., W.G., Y.Y., X.L., S.Y. and M.Y. provided statistical analysis and contributed to study design. Jinwei H., L.C., B.L. and S.Z. performed next generation sequencing. X.D. and W.W. contributed to statistical analysis and participated in the design of experimental work. Jing H., Qi L., S.D., Yan W., Qiang L., W.C., S.W., Y.D., F.F., G.Z., J.Z., L.H., J.X., W.Y., Z.T., D.J., T.J., Qiao L., L.X., H.H., L.S., Jin L., Kefeng W., D.W., J.S., Y.L., T.Z., C.L., Yusheng W., Y.S., J.G., S.X., Junfeng L. and G.L. contributed to data collection for study samples. Kai W. and M.W. designed the experimental work, wrote study documentation, analyzed data, wrote the manuscript and were chief investigators of the study.

Competing interests

The authors declare no competing interests.

Methods

Samples and patients

A total of 11,553 patients across 25 tumor types were submitted for a next-generation sequencing (NGS) based cancer assay (CSYS) in a Clinical Laboratory Improvement Amendments (CLIA)-certified and College of American Pathologists (CAP)-accredited laboratory (OrigiMed). Tumor types were annotated according to an institutional classification system, OncoTree (http://www.cbioportal.org/oncotree/). This study was approved by the Research Ethics Committees of hospitals. All patients gave informed consent to participate in the study and gave permission for the use of samples.

Unique tumor samples and matched normal blood samples of each patient were collected by standardized protocols. All tumor samples were formalin fixed and paraffin embedded (FFPE). Hematoxylin and eosin (H&E)-staining sections of tumor samples were reviewed by senior pathologists for the estimation of tumor cellularity. For each tumor sample, 15 to 25 eligible unstained sections were collected for DNA extraction. According to multiple quality control metrics, 825 (7%) samples with insufficient tumor content (<10%), 321 (3%) samples with inadequate extracted DNA yield (<50 ng) and

213 (2%) samples with a sequencing technical failure (unique mean coverage lower than 300×, biased coverage distribution or sample contamination) were excluded. In total, 10,194 (88%) samples were successfully included in the final analysis (**Supplementary Fig**. 1).

Sequencing workflow.

The laboratory and bioinformatics protocols of CSYS have been described and validated in previous study (**Supplementary Fig. 13**) [8]. DNA extracted from tumor tissues and matched normal peripheral blood was fragmented to ~250 bp and subjected to library construction using KAPA HyperPrep Kits (KAPA Biosystems), followed by hybridization capture using custom xGen Lockdown Probes and Reagents (Integrated DNA Technologies). As a main component, the custom hybridization capture panel targets ~2.6 Mb of the human genome containing all coding exons of 450 genes (**Supplementary Table 11**), as well as promoter of *TERT* and select introns of 39 genes frequently rearranged in cancer. Post-capture libraries were mixed, denatured and diluted to 1.5-1.8 pM (NextSeq 500) or 200-230 pM (NovaSeq 6000) and subsequently sequenced on NextSeq 500 or NovaSeq 6000 sequencers (Illumina). Paired-end sequencing was done following the manufacturer's protocols. Tumor samples were sequenced to a median unique coverage of 1202× (**Supplementary Fig. 3**) and matched normal blood samples were sequenced to a mean unique coverage 300×. Data quality was inspected and controlled, followed by a suite of customized bioinformatics pipelines for variant calls. SNVs, indels and CNVs were identified using MuTect, Pindel and EXCAVATOR, respectively. Gene rearrangements were detected using an algorithm developed in house. At least 5 unique supporting reads were necessary for a SNV/indel. All variants were manually reviewed in the Integrative Genomics Viewer (IGV) and a custom visual software to avoid false positives. Test results, including somatic variants and inherited pathogenic variants, were returned to patients and their physicians based on their needs.

Microsatellite instability (MSI) and tumor mutational burden (TMB)

MSI status and TMB of tumor samples is according to bioinformatics approaches developed in house [8]. Microsatellite instability-high (MSI-H) is defined as more than 15% of selected microsatellite loci show unstable in tumor compared to matched peripheral blood. TMB score of each tumor sample is calculated by counting the number of somatic SNVs and indels per megabase (Mb) in targeted coding region of genome. Noncoding mutations, hotspot mutations and known germline polymorphisms in the U.S. National Center for Biotechnology Information's Single Nucleotide Polymorphism Database (dbSNP) are not counted. In this study, 10 was adopted as the threshold value for differentiating TMB high (TMB-H) from TMB low (TMB-L).

Overall comparative analysis pipeline

The available full data (mutation results and clinical information) of the MSK-IMPACT and TCGA (PanCancer Atlas and ovarian cancer, Nature 2011) studies were downloaded from cBioPortal (https://www.cbioportal.org/). The corresponding tumor types with more than 60 patients in each cohort were comparable. Somatic variants of OrigiMed (OM) and MSK datasets were comparatively analyzed,

including somatic SNVs, indels, deletions of tumor suppressor genes, amplifications of oncogenes and functional fusions/rearrangements. Considering the differences of detecting methods between OM and TCGA studies, only comparative analysis of somatic SNVs and indels of the TCGA dataset was performed. These variants were in coding regions, exon-intron flanks, 5'flanks (*TERT* gene) of 266 comparable cancer related genes. All variants were divided into several subtypes, including substitution/indel, truncation, amplification, deletion and fusion/rearrangement. Chi-squared test (χ^2) and Fisher's exact test were performed to the comparison of the frequencies of gene variants between two cohorts, and then *P* values were corrected with Benjamini-Hochberg (BH) method. Genes whose cohort-level altered frequency difference with statistic false discovery rate (FDR < 0.05) were reported as significant. We then controlled for clinical factors in the multivariate logistic regression analysis when compared the frequencies of mutated genes between aOM and MSK cohorts (gender, smoking status [for LUAD, LUSC and HNC only], sampling method and primary/metastasis/recurrent tumor specimen), or between aOM and aTCGA cohorts (gender, age and primary/metastasis/recurrent tumor specimen).

Programmed death-ligand 1 (PD-L1) immunohistochemistry staining assay

We performed immunohistochemistry (IHC) staining of FFPE tissue sections for PD-L1 protein using an anti-PD-L1 antibody (clone 28-8; Cat#ab205921; Abcam). Briefly, slides were incubated at 60°C, deparaffinized in xylenes, and rehydrated with graded ethanol. Antigen retrieval was performed using the Universal HIER antigen retrieval reagent (Cat#ab208572; Abcam) in a steamer. Non-specific binding was blocked with the Dako EnVision FLEX Peroxidase-Blocking Reagent. Dilutions (1:300) of the primary antibodies were used for PD-L1 antigen detection. All other staining was performed primarily with Dako series reagents (Cat#K8002; Dako). All slides were counterstained with haematoxylin. Specimens were scored as positive by the pathologist using Tumor Proportion Score (TPS), which is the percentage of viable tumor cells with partial or complete membrane staining at any intensity. PD-L1 positivity in the study was defined as TPS ≥1%, and the specimens with 1-50% TPS and ≥50% TPS were respectively scored as weak and strong positive, respectively.

Clinical utility evaluation

We used previously reported criteria in the MSK study to assess the clinical actionability of variants. OncoKB (August 31, 2021, http://oncokb.org/) knowledge base was used to annotate and classify variants into different levels: Food and Drug Administration (FDA)-recognized biomarkers (Level 1), variants that predict response to standard-of-care therapies (Level 2), variants that predict response to investigational agents in clinical trials (Level 3), or variants that predict to investigational agents in preliminary, preclinical studies (Level 4). These levels were also subdivided according to evidence within or between tumor types: Level A (1, 2A, 3A, 4) for the same tumor type, and Level B (2B, 3B) for different tumor types. Although wildtype *KRAS* was defined as a level-related factor, we excluded wildtype *KRAS* in CRC in this study when establishing the subset of Level 1. A high level of MSI-H was considered as an independent predictive biomarker with evidence Level 1, regardless of the tumor type. If a variant was involved in different levels, the highest level was chosen for further analysis according to the rank: Level 1

> 2A > 3A > 2B > 3B > 4. The final level of each patient was defined as the highest evidence level of all variants detected in the patient. Information about drugs was from the U.S. Food and Drug Administration (FDA) (http://www.fda.gov) and National Medical Products Administration (NMPA) of China (http://www.nmpa.gov.cn).

Statistical analysis

Chi-squared test (χ^2), Fisher's exact test, and Benjamini–Hochberg (BH) method were used in the comparison of the gene alteration frequency between two cohorts, and they were also used to evaluate the association between clinical characteristics and significantly altered genes mutations. Multivariate logistic regression models were applied to predict the factors that might influence alteration of gene in specific cancer. Cohort factor, as an independent variable, and several clinical features were added to each regression model and the multivariate modeling was reassessed. Corrections were also performed using BH method. A Wilcoxon-test could be done within each tumor type to compare the tumor mutational burden (TMB) between the MSK and the OM. The significant differences in this study were based on *P* values or FDR < 0.05.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Figures

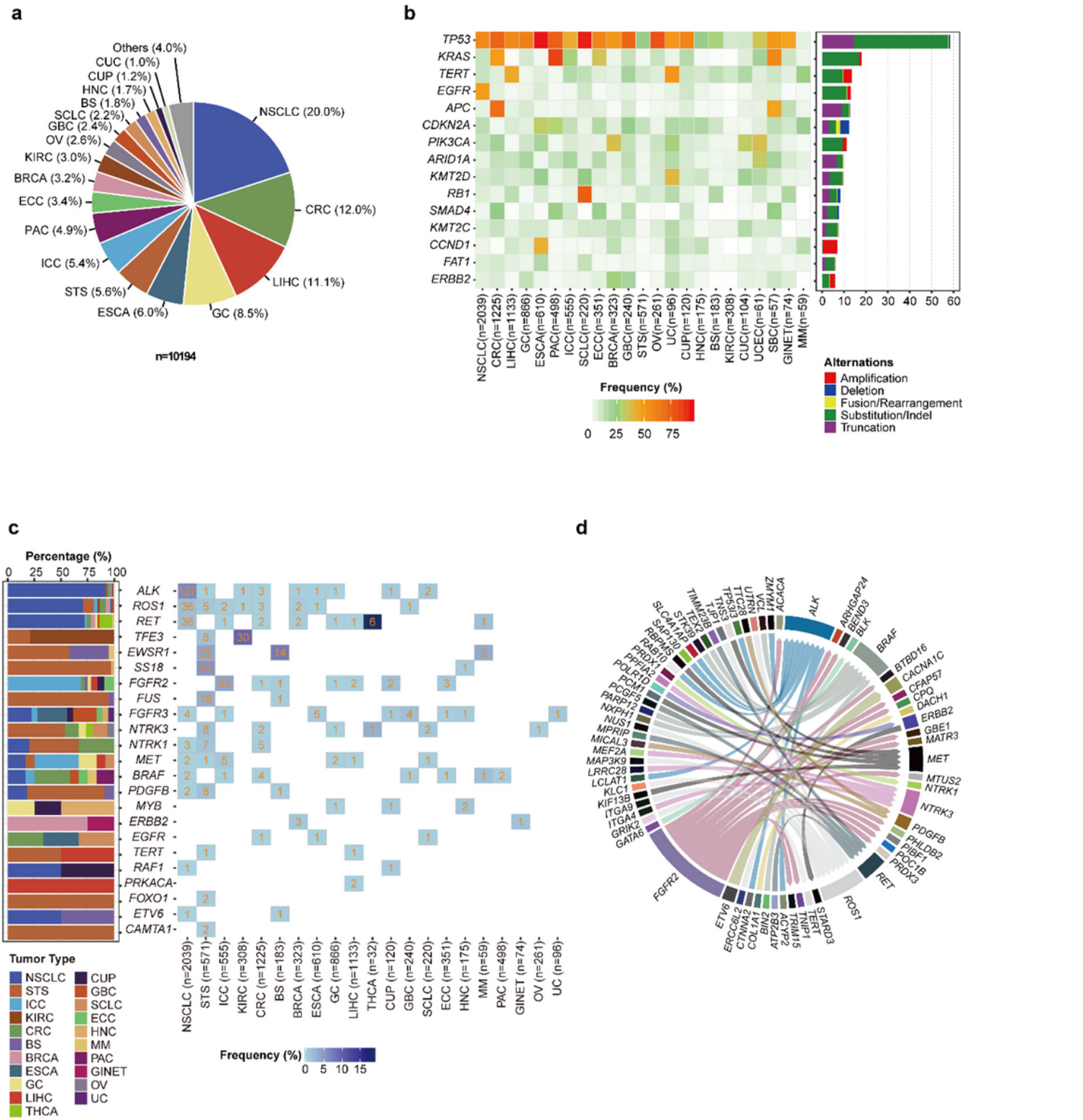


Figure 1

Overview of the mutational landscape of solid tumors from Chinese patients. (a) Distribution of tumor types among cases successfully sequenced from 10,194 patients. (b) Recurrent somatic alterations across common tumor types. The top 15 Tier 1 Cancer Gene Census oncogenes and tumor suppressor genes with a cohort-level variation frequency of ≥ 6% are shown, in descending order. Bars represent the percent of cases within each tumor type having at least 1 of 5 different classes of genomic alterations.

(c) Genes recurrently rearranged to form putative gene fusions are displayed across principal tumor types. The tumor type-specific distribution of these genes is presented on the left side (various colors represent different tumor types). The number of corresponding gene fusions in each tumor type is shown in the right boxes, and the frequency is shown in gradient blue. (d) Novel gene fusions across multiple tumor types. A total of 57 novel driver-partner relationships were detected spanning 71 genes. The thickness of the line between two genes implies the relative count.

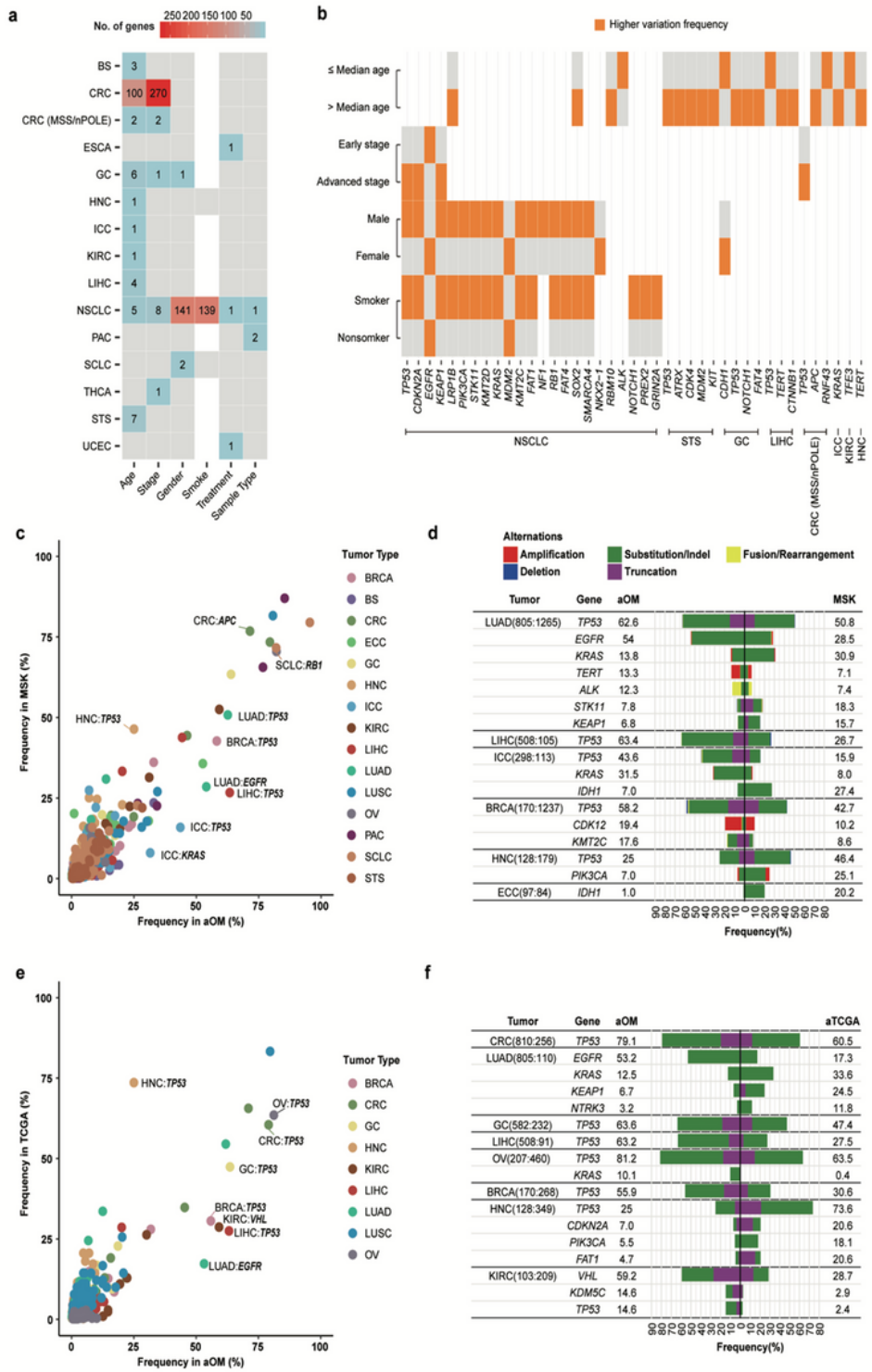


Figure 2

Analysis of somatic altered genes. (a) Numbers of correlated altered genes with six clinical features across tumor types. Only genes with significant differences (FDR < 0.05) between two groups of clinical feature are calculated. The “age” feature includes younger patient group and older patient group, separated by the median initial diagnosis age of patients of each tumor type. The “stage” feature includes early-stage cancer group and advanced-stage cancer group. The “smoke” feature, including smoker group (current smokers and former smokers) and nonsmoker group (never-smokers), is analyzed in lung cancers (NSCLC and SCLC) and head and neck cancers (HNC). The “treatment” feature includes treatment-naïve group and pretreated group. The “sample type” feature includes primary sample group and metastatic/recurrent sample group. (b) Correlation between Tier 1 Cancer Gene Census genes and clinical features. Genes with significant differences (FDR < 0.05, number of each group > 60, and sum of variation frequencies > 10%) between two feature groups are shown. The group with a higher variation frequency in each clinical feature is labeled in orange. (c) Frequency of altered gene in 15 comparable tumor types between the aOM cohort and MSK cohort. (d) Comparison of significantly different altered genes (FDR < 0.05) between the aOM cohort (left) and MSK cohort (right). Altered genes whose sum of frequencies in the two cohorts exceeds 10% are displayed. The alteration frequency (%) of specific genes are shown in the “aOM” and “MSK” columns. (e) Frequency of altered gene in 9 comparable tumor types between the aOM cohort and aTCGA cohort. (f) Comparison of significantly different altered genes (FDR < 0.05) between the aOM cohort (left) and MSK cohort (right). Altered genes whose sum of frequencies in the two cohorts exceeds 10% are displayed.

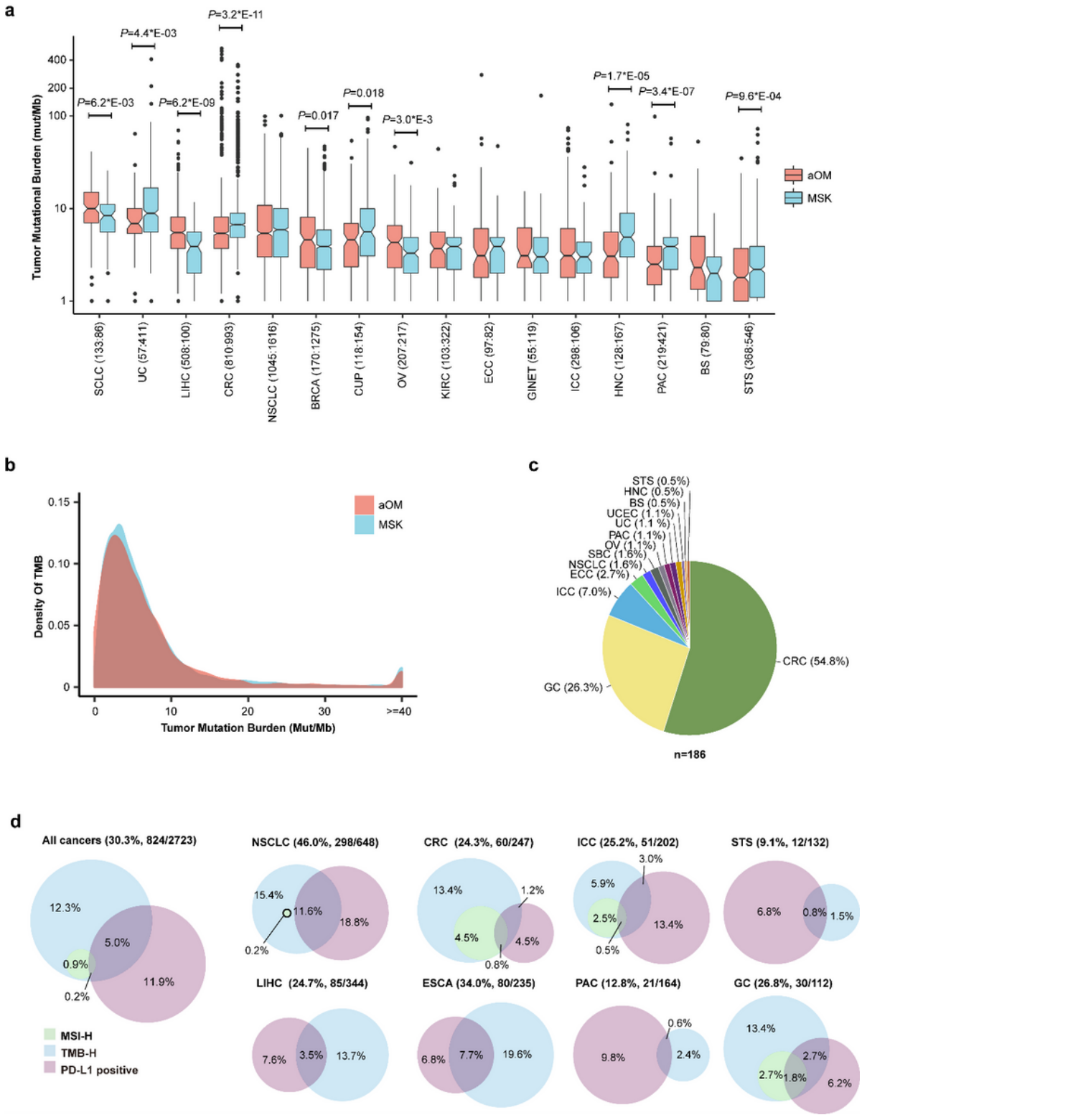
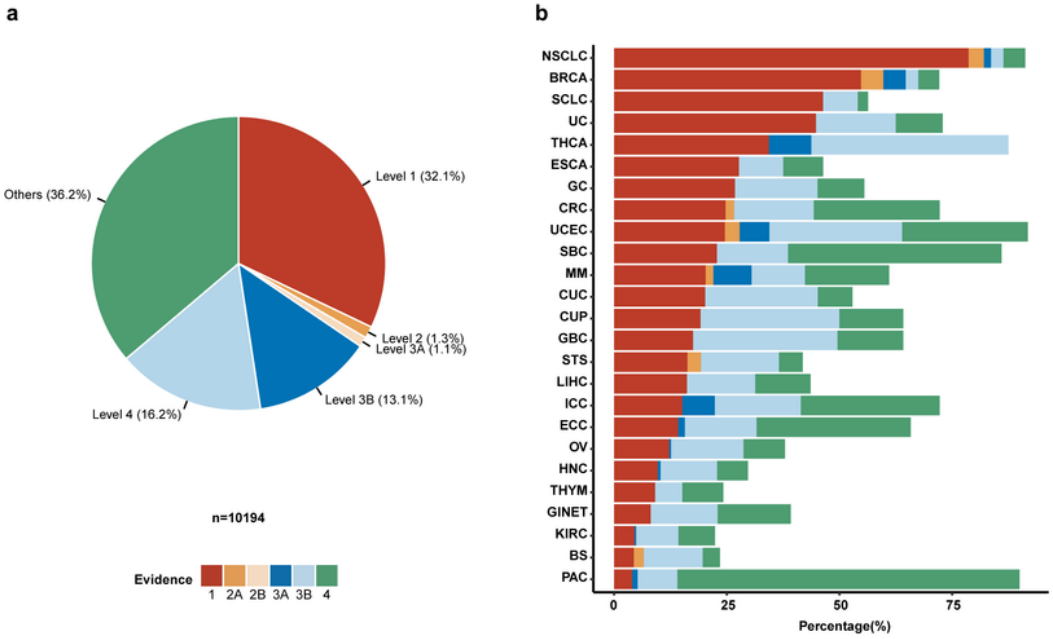


Figure 3

Correlation of tumor mutational burden (TMB), microsatellite instability high (MSI-H), and PD-L1 expression in OM cohort. (a) The tumor type-specific distribution of TMB (excluding samples with TMB of 0) between the aOM cohort (light red) and the MSK cohort (light blue). Tumor types are sorted from left to right based on median TMB values (y-axis). The total number of samples is shown for each tumor type. P values are labeled on the top of corresponding tumor types in which TMB is significant different between

cohorts. (b) Distribution of TMB density between the aOM cohort (light red) and the MSK cohort (light blue). (c) The tumor type-specific distribution for 186 samples with MSI-H. (d) The analysis for the cohort-level or tumor type-specific correlation of TMB, MSI, and PD-L1 expression in 2,723 samples with available information on MSI, TMB, and PD-L1. The Venn diagram shows the proportion of TMB-H (light blue), PD-L1 positive (light purple), and MSI-H (light green). Total proportions, numbers of samples with at least either MSI-H, TMB-H, or PD-L1 positive, and total numbers of samples are shown in parentheses.



c

Count	Gene	Variation	Tumor	Proportion of Level	Proportion of Tumor	Drugs
544	TMB-H	TMB-H	NSCLC	16.6%	26.7%	Pembrolizumab
413	EGFR	L858R	NSCLC	12.6%	20.3%	Erlotinib, Erlotinib+Ramucirumab
388	EGFR	Exon19deletion	NSCLC	11.8%	19.0%	Erlotinib, Erlotinib+Ramucirumab
370	KRAS	GoF Mutations	PAC	22.4%	74.3%	Binimetinib, Cobimetinib, Trametinib
316	KRAS	GoF Mutations	CRC	19.2%	25.8%	Binimetinib, Cobimetinib, Trametinib
256	TMB-H	TMB-H	CRC	7.8%	20.9%	Pembrolizumab
222	TMB-H	TMB-H	GC	6.8%	25.6%	Pembrolizumab
179	TMB-H	TMB-H	LIHC	5.5%	15.8%	Pembrolizumab
150	TMB-H	TMB-H	ESCA	4.6%	24.6%	Pembrolizumab
137	ALK	Fusions	NSCLC	4.2%	6.7%	Crizotinib, Ceritinib, Alectinib
124	PIK3CA	GoF Mutations	CRC	9.3%	10.1%	Alpelisib
117	KRAS	GoF Mutations	ICC	7.1%	21.1%	Binimetinib, Cobimetinib, Trametinib
102	MSI-H	MSI-H	CRC	3.1%	8.3%	Nivolumab, Nivolumab+Ipilimumab, Pembrolizumab
102	TMB-H	TMB-H	SCLC	3.1%	46.4%	Pembrolizumab
100	PIK3CA	C420R, E542K, E545A, E545D, E545G, E545K, H1047L, H1047R, H1047Y, Q546E, Q546R	BRCA	3.1%	31.0%	Alpelisib
100	CDKN2A	LoF Mutations	PAC	6.1%	20.1%	Abemaciclib, Palbociclib, Ribociclib
91	PIK3CA	C420R, E542K, E545A, E545D, E545G, E545K, H1047L, H1047R, H1047Y, Q546E, Q546R	CRC	6.8%	7.4%	Alpelisib
91	KRAS	GoF Mutations	ECC	5.5%	25.9%	Binimetinib, Cobimetinib, Trametinib
89	ARID1A	TruncatingMutations	LIHC	5.4%	7.9%	PLX2853, Tazemetostat
88	KRAS	G12C	NSCLC	2.7%	4.3%	Sotorasib
77	ERBB2	Amplification	BRCA	2.3%	23.8%	Lapatinib, Trastuzumab+Pertuzumab, Neratinib, Margetuximab, Trastuzumab+Tucatinib
62	ARID1A	TruncatingMutations	GC	3.8%	7.2%	PLX2853, Tazemetostat
62	ERBB2	GoF Mutations	NSCLC	47.0%	3.0%	T-DM1, Trastuzumab Deruxtecan
62	CDK4	Amplification	STS	4.7%	10.9%	Abemaciclib, Palbociclib
57	EGFR	G719	NSCLC	1.7%	2.8%	Afatinib
56	TSC2	LoF Mutations	LIHC	4.2%	4.9%	Everolimus
56	KRAS	GoF Mutations	NSCLC	3.4%	2.7%	Binimetinib, Cobimetinib, Trametinib
55	ARID1A	TruncatingMutations	ICC	3.3%	9.9%	PLX2853, Tazemetostat
55	TMB-H	TMB-H	ICC	1.7%	9.9%	Pembrolizumab
54	BRAF	V600E	CRC	1.6%	4.4%	Encorafenib+Cetuximab
53	PIK3CA	GoF Mutations	GC	4.0%	6.1%	Alpelisib
49	MSI-H	MSI-H	GC	1.5%	5.7%	Pembrolizumab

Figure 4

Clinical actionability of somatic alterations in the OM cohort. (a) Variants are assigned to different levels of clinical actionability according to OncoKB. The distribution of the highest level of actionable variants across all patients is shown in the pie chart. The colors representing each level are used throughout the other panels in this figure. (b) Distribution of highest level of actionable variants across tumor types. (c) Details of the 30 most common actionable variants, proportions of levels in corresponding tumor types and their potential sensitive drugs. The numbers of patients in each level are shown in the bar graph. The right table shows genes, variants, and the tumor type for each level of clinical actionability, as well as the proportion of patients in corresponding tumor type and level. Potential sensitive drugs suggested by biomarkers are also shown in the table. Drugs approved by both U.S. Food and Drug Administration (FDA) of and National Medical Products Administration (NMPA) of China are labeled in green, drugs only approved by the FDA are labeled in blue, and drugs in development are labeled in black.

Supplementary Files

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

- SupplementaryFigures.docx
- supplTables.xlsx