

Supplementary data

Methods

Analysis of TCGA samples

Analysis of non-TCGA datasets

Statistical analyses

Data and code availability

Supplementary figures

Figure S1. Characteristics of the eight tumor genomic subtypes derived from TCGA solid tumors (N=9794)

Figure S2. Features of SMOKING (SMK) subtype (n=1072)

Figure S3. Features of ULTRAVIOLET LIGHT (UVL) subtype (n=401)

Figure S4. Features of APOBEC (APB) subtype (n=1036)

Figure S5. Features of MISMATCH REPAIR DEFICIENCY (MRD) subtype (n=339)

Figure S6. Features of *POLE* (POL) subtype (n=81)

Figure S7. Features of HOMOLOGOUS RECOMBINATION DEFICIENCY (HRD) subtype (n=1956)

Figure S8. Features of GENOMICALLY STABLE (GNS) subtype (n=909)

Figure S9. Correlation of representative etiological signatures with immune-related gene expression scores

Figure S10. Consistency of clustering results from different variant callers

Figure S11. Developing classifiers through machine learning algorithms

Figure S12. Relationship between tumor genomic subtype, response rate and cancer type in the whole cohort

Figure S13. Studies examined by type of drug

Figure S14. Study in the dataset from the KEYNOTE trials, which were prospective cohort studies of patients treated with solely pembrolizumab (n=311)

Figure S15. Study using the cohorts' optimal cutoff of TMB high

Figure S16. Association between TMB and ICI response divided by irGS status per dataset

Figure S17. Determination of MSI-high cases using MSIsensor

Figure S18. Comparison of the number of missense mutations or non-synonymous SNVs from our WES pipeline and previously published data

References

Methods

Analysis of TCGA samples

• Sample selection and collection of clinical information

Clinical and somatic gene mutation profiles of all tumors except diffuse large B-cell lymphoma, acute myeloid leukemia, and thymoma, PanCancer Atlas datasets were downloaded from cBioPortal¹ (29 studies, 10075 cases). Information on smoking habit and HPV infection status was obtained from GDAC². Among these tumors, 9794 cases for which the somatic mutation profiles analyzed by Mutect2³ in the MAF format were available from the GDC portal⁴ were selected for analysis.

The annotation of cancer types with FDA approval for ICI monotherapy was based on a previous report⁵. The response rates for ICI monotherapy for each tumor type were obtained from previous reports^{6,7}. The response rate data for endometrial cancer with mismatch repair deficiency or with mismatch repair proficiency were calculated from another previous report⁸.

• Identification of genomic subtypes based on mutational signatures

Using MutationalPatterns⁹, the contribution values of each sample to COSMIC v2 30 mutational signatures¹⁰ were calculated from the four different somatic mutation profiles, which were pre-computed using Mutect2³, Varscan2¹¹, MuSE¹², and Somatic Sniper¹³ and were available in the MAF format on the GDC portal⁴. Using the log10 transformed values of these contribution values, unsupervised hierarchical clustering with Ward's method was performed.

• Annotation of gene alterations

Somatic gene mutations annotated as significant in the PanCancer Atlas studies of cBioPortal¹ were included, while those marked as mutations of unknown significance were excluded. For germline mutations, we obtained the annotations from a previous report¹⁴, where those annotated as “likely pathogenic” or “pathogenic” in “Overall Classification” column were retained. For gene mutations in *BRCA1* and *BRCA2*, we extracted those with locus-specific LOH or homozygous deletions as we previously reported¹⁵. For gene promoter methylations in *MLH1* and *BRCA1*, we obtained annotations from a previous report¹⁶.

• Insertions and deletions-based mutational signatures (ID signatures)

The annotated somatic mutations of each sample called by Mutect2 were obtained from the GDC portal in the VCF format, and the insertions and deletions (indels) with "PASS" annotations were extracted. The contribution values of each sample to the COSMIC reference small indels signatures¹⁷ were calculated using YAPSA¹⁸. The ratio of contribution values of indel signature 3 and indel signature 6 to the number of all detected mutations were calculated as indel signature 3 ratio and indel signature 6 ratio, respectively.

• MSI score and MSI-high annotation

We calculated MSI scores of all samples using MSI sensor¹⁹ with the default parameters from normal-tumor paired WES sequencing data. For UCEC, CRC, STAD, and ESCA, MSI status was obtained from the clinical information in cBioPortal. Within these samples, the optimal cutoff value of MSI score for the annotated MSI-high cases was calculated using the ROC curve and the Youden index (Figure S17A,B). Then, for the other cancer types, samples with the score above this cutoff were determined to be MSI-high (Figure S17C).

• Other genomic alterations scores

The following scores were calculated from the somatic mutation profiles calculated by Mutect2. Tumor mutational burden: the number of missense mutations. Total indel count: the total number of frameshift insertions, inframe insertions, frameshift deletions, and inframe deletions. Indel ratio: the ratio of the total indel count to the total number of detected mutations, including synonymous mutations. Insertion to indel ratio: the ratio of the total number of frameshift insertions and inframe insertions to the total indel count.

For predicted neoantigens counts based on netMHCpan²⁰, we obtained the pre-computed data for SNVs and indels from a previous report²¹.

For chromosomal changes, we obtained HRD scores and CNV burden scores from GDC portal⁴.

• Gene expression scores

We obtained batch-corrected gene expression values from a previous report²¹. The CYT score was calculated from the geometric mean of the expression levels of *GZMA* and *PRF1* according to the previous report²². We obtained a gene set "HALLMARK_PI3K_AKT_PATHWAY" from MSigDB²³ and calculated its enrichment score using ssGSEA²⁴. In addition, using the literature²⁵ as a reference, the GEP score was calculated as the sum of each gene expression multiplied by the following coefficients: *CL5*=0.008346; *CD27*=0.072293; *CD274*=0.042853; *CD276*=-0.0239; *CD8A*=0.031021; *CMKLR1*=0.151253; *CXCL9*=0.074135; *CXCR6*=0.004313; *HLA.DQA1*=0.020091; *HLA.DRB1*=0.058806; *HLA.E*=0.07175; *IDO1*=0.060679; *LAG3*=0.123895; *NKG7*=0.075524; *PDCD1LG2*=0.003734; *PSMB10*=0.032999; *STAT1*=0.250229; *TIGIT*=0.084767.

• Development of Tumor Genomic Subtype Analyzer (TGSA)

Using the 30 signature contribution values as features and the genomic subtypes as labels in the selected 7181 samples (Figure S10C), we built four independent classifiers from four different algorithms, namely, k-nearest neighbor, support vector machine, logistic regression, and random forest using the Scikit-learn module in Python. For the former three classifiers, the main hyperparameters were optimized by double cross-validation (Figure S11A). First, the above selected TCGA samples were divided into two parts, X1 and X2. Second, parameters were calculated using X1 by two-fold cross-validation, and those parameters were evaluated using X2 as test data. Third, X1 and X2 were swapped, and the same calculations were performed. These processes were repeated 100-1000 times to determine the optimal parameters (Figure S11A). For the random forest model, since parameter adjusting hardly changed the prediction accuracy, the default settings were used.

After calculating the contribution values of the 30 mutational signatures from external somatic mutation profiles in VCF or MAF format using MutationalPatterns⁹, the four classifiers independently make predictions using these values as input, and then integrate the results to output the final classification (Figure 2A). In the classification into eight subtypes, if the results from three or more classifiers matched, the matched result was determined to be the subtype, otherwise it was determined to be undeterminable (UND). In parallel, if the results from three or more classifiers belong to immune-related genomic subtype (irGS), namely, SMK, UVL, APB, POL, or MRD, it was classified as irGS, otherwise non-irGS. This tool is available in the GitHub page (<https://github.com/shirotak/TGSA>).

Analysis of non-TCGA datasets

• Clinical Proteomic Tumor Analysis Consortium (CPTAC) datasets

Clinical information including smoking status and somatic mutation profiles in the MAF format were obtained from GDC portal⁴. Somatic mutations in MMR genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*) were retained only for truncating mutations or for missense mutations with COSMIC annotations. For gene expression analysis, we obtained the fragments per kilobase of exon per million mapped reads (FPKM) value from the GDC portal⁴. CYT score and GEP score were calculated in the same way as TCGA data analysis. For indel mutational signature, we downloaded annotated somatic mutations data in the VCF format. After counting the number of all somatic mutations annotated with "PASS", insertions and deletions were extracted and the contribution of indel signatures was calculated using YAPSA¹⁸. ID3 ratio and ID6 ratio were calculated as the ratio of each contribution values to the number of all somatic mutations.

• cBioPortal datasets

Somatic mutation profiles released in MAF format were obtained from the websites (Table S2).

- **National Bioscience Database Center (NBDC) datasets**

We obtained raw whole exome datasets (Table S2) from the NBDC Human Database²⁶, calculated somatic mutation profiles in our WES analysis pipeline (see below).

- **Sample selection and definition of response in ICI-treated cohorts**

Collected samples were derived from pairs of the primary tumor and normal blood or tissue, and those collected from metastatic sites different from the primary tumor (lymph nodes, bones, distant internal organs, etc.) were excluded. In addition, collected samples were taken before or during ICI administration, and those with a history of ICI treatment at the time of sample collection were excluded. Most of the cases were evaluated using the RECIST criteria for radiological response or equivalent, where CR/PR was defined as a responder and SD/PD/NE as a non-responder. In some datasets, we could not find a response assessment by such criteria from the articles, and we determined responder/non-responder using alternative clinical data available. For example, in the datasets from Snyder et al²⁷ and from Anagnostou et al²⁸, where survival outcome was available, after excluding cases who survived less than six months of follow-up, cases with a PFS of 12 months or more were determined to be responders, and others non-responders. For the dataset from Cristescu et al²⁵, we distinguished between responders and non-responders based on the values listed in the figures and tables in the paper, and confirmed that the annotations were consistent with the results of the other figures. As a result, a total of 938 patients from 13 datasets that met the above criteria and had available response information were included in the analysis (Table S3).

- **WES analysis pipeline**

For the raw paired WES data obtained from NBDC, dbGaP and EGA (Table S2, Table S3), somatic mutations were analyzed in the following steps according to the Best Practice Workflows of somatic short variant (SNVs + Indels) discovery published by the Broad institute GATK team²⁹, where Genome Analysis Toolkit (GATK) v4.0.12 and Picard v2.20.2 were used. (1) BAM files were converted to FASTQ files using Picard SamToFastq. (2) Low-quality reads and presumed adapter sequences were removed using TrimGalore³⁰. (3) The trimmed reads were aligned to the human reference genome hg38 obtained from the GATK resource bundle using BWA mem (v0.7.17-r1188)³¹. (4) Data preprocessing including marking duplicates, making the panel of normal samples, and estimating contamination was performed using the Picard and GATK tools. (5) Somatic mutation calling including orientation bias filtering was performed using Mutect2. (6) The following parameters were used to filter out variants with low reliability: annotated as "PASS"; "TLOD" greater than 6.3 and 9 for single nucleotide variants and insertions and deletions, respectively; coverage of altered alleles greater than or equal to 5; altered allele frequency greater than or equal to 2.5%. (7) Variant annotation was performed using Funcotator with default setting using data source of v1.7.20200521s to generate MAF format files.

For the dataset from Hellmann et al³², only the VCF files after somatic mutation calling were available, so we filtered variants using the same criteria as described above and converted to MAF files for subsequent analysis.

- **Validation of the WES analysis pipeline**

As reported by Litchfield et al³³, TMB does not differ significantly depending on the exome capture kits used. For compatibility between TCGA and other datasets, we adopted the Agilent SureSelect Human All Exon V5 capture kit as the common exome regions to be analyzed in our WES analysis pipeline. Comparing the number of missense mutations calculated from cBioPortal¹ with those from our pipeline among the same samples, the Pearson correlation tended to exceed 0.9 in most of the datasets (Figure S18). Similarly, when we compared the number of non-synonymous SNVs in the Cristescu et al paper²⁵ with those from our pipeline, the Pearson correlation was 0.945 (Figure S18).

Statistical analyses

Unless otherwise noted, statistical analyses were performed in Python (3.7.4). The Mann–Whitney U test, chi-square test, and Spearman's rank correlation coefficient test were performed using SciPy (1.6.1). Survival analyses including

the Kaplan–Meier curve, log-rank test, and Cox proportional hazard regression model were performed using Lifelines (0.25.10) and StatsModels (0.12.2). Machine learning analyses were performed using Scikit-learn (0.24.1). We considered a p-value < 0.05 as being statistically significant. Venn diagrams were depicted using “*VennDiagram*” (1.6.20) package in R. The Jonckheere-Terpstra test was performed using “*clinfun*” (1.0.15) package in R. The Passing-Bablok regression was performed using “*mcr*” (1.2.2) package in R.

Data and code availability

Controlled access data used in this study were obtained from dbGaP, EGA, and NBDC with access permissions according to the respective required procedures (Table S2 and S3). The processed data and codes to reproduce the results of this work are available on the GitHub page (https://github.com/shirotak/pancancer_MutSig_ICI). Other codes for preprocessing or restricted-access data are available from the corresponding author upon reasonable request.

Supplementary Figures

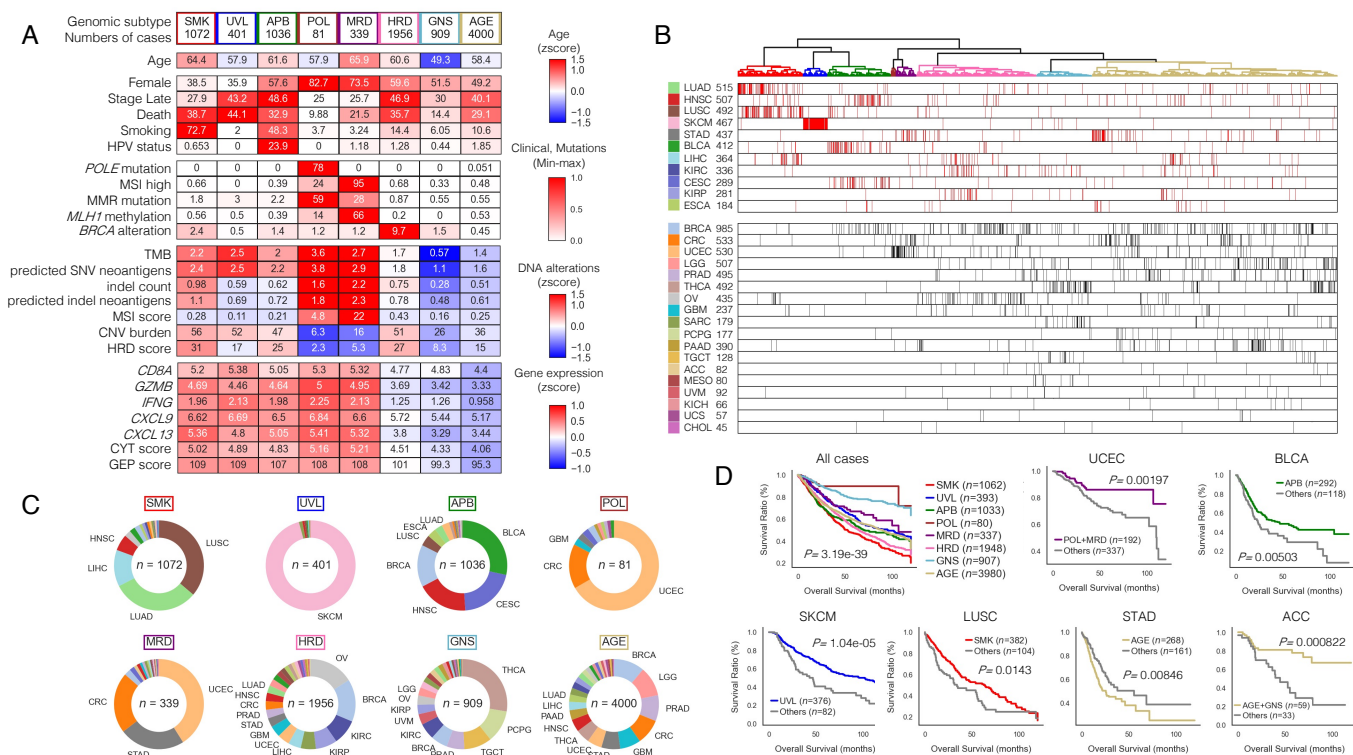


Figure S1. Characteristics of the eight tumor genomic subtypes derived from TCGA solid tumors (N=9794)

- A) Clinical and genomic information for each subtype, related to Figure 1A. The values are either the average of the Z-scores or the frequency for each subtype, and the values are displayed as a heat map. Predicted SNV and indel neoantigens counts by netMHCpan were significantly correlated with TMB and indel counts, respectively. CNV burdens and HRD scores were also correlated with each other.
- B) Relationship between the distribution of cancer types based on the hierarchical clustering and the FDA approval for ICI therapy. Each case is represented by a red bar (with FDA approval for ICI) or a black bar (without FDA approval for ICI).
- C) Differences in survival outcomes between subtypes. When all cases were compared among the eight subtypes, the prognosis was different. Furthermore, for the cancer types shown here, the prognosis was significantly different by subtype in the analysis of each cancer type. P-value is based on log-rank test.
- D) Type of cancer and number of samples per subtype. The color of each cancer type matches the color shown in B.

ACC, Adrenocortical carcinoma; BLCA, Bladder urothelial carcinoma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, Cholangiocarcinoma; CRC, Colorectal adenocarcinoma; ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and neck squamous cell carcinoma; KICH, Kidney chromophobe; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LGG, Brain lower grade glioma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; MESO, Mesothelioma; OV, Ovarian serous cystadenocarcinoma; PAAD, Pancreatic adenocarcinoma; PCPG, Pheochromocytoma and paraganglioma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; SKCM, Skin cutaneous melanoma; STAD, Stomach adenocarcinoma; TGCT, Testicular germ cell tumors; THCA, Thyroid carcinoma; UCEC, Uterine corpus endometrial carcinoma; UCS, Uterine carcinosarcoma; UVM, Uveal melanoma.

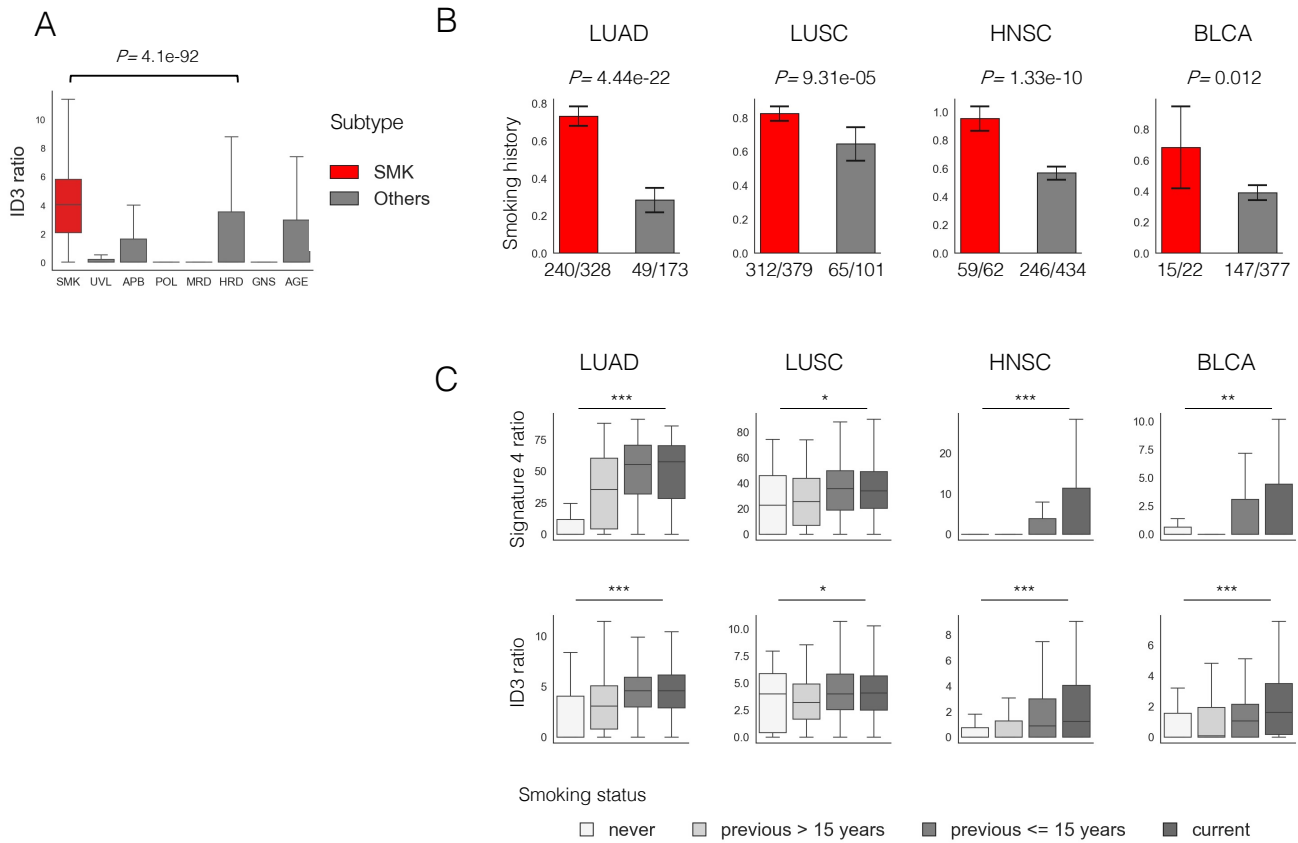


Figure S2. Features of SMOKING (SMK) subtype (n=1072)

The association with smoking habits was investigated in LUAD, LUSC, HNSC and BLCA where clinical information on smoking was available.

- A) The ratio of indel signature 3, which is related to tobacco smoking, was higher in SMK subtype than other subtypes.
- B) SMK cases had a higher frequency of smoking history (current smoking and smoking within 15 years) than the other groups.
- C) The ratio of signature 4 (upper) and indel signature 3 (lower), which are known to be related to smoking, were positively correlated with smoking habit. (Jonckheere-Terpstra test, *** $P < 1e-4$, ** $P < 0.01$, * $P < 0.05$)

BLCA, Bladder urothelial carcinoma; HNSC, Head and neck squamous cell carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma

A

Cancer type		Numbers
SKCM		384
Non-SKCM	SARC	5
	HNSC	3
	LUSC	3
	BLCA	2
	BRCA	1
	CRC	1
	GBM	1
	THCA	1

B

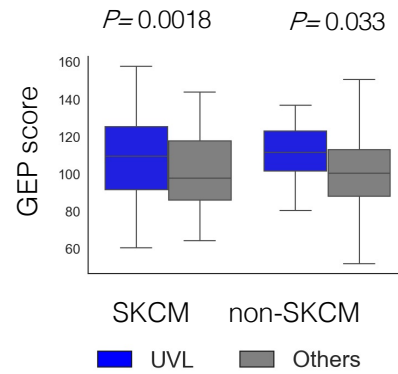


Figure S3. Features of ULTRAVIOLET LIGHT (UVL) subtype (n=401)

A) Cancer type of UVL subtype. 95.8% were SKCM.

B) Excluding SKCM, UVL tumors showed higher GEP score than the other subtypes ($P=0.033$, Mann-Whitney U test).

These results suggest that, although rare in cancers other than SKCM, there are some tumors that exhibit UVL subtype and may have a high tumor immunogenicity.

BLCA, Bladder urothelial carcinoma; BRCA, Breast invasive carcinoma; CRC, Colorectal adenocarcinoma; HNSC, Head and neck squamous cell carcinoma; LUSC, Lung squamous cell carcinoma; SARC, Sarcoma; THCA, Thyroid carcinoma

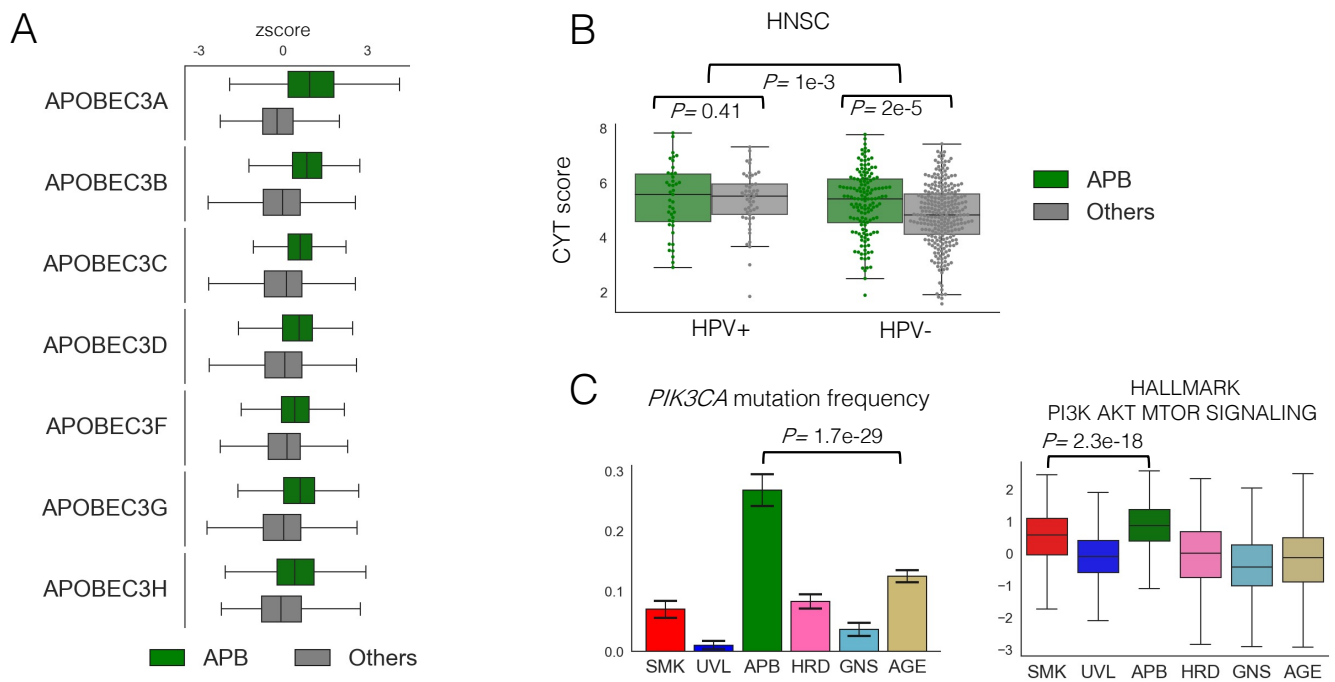


Figure S4. Features of APOBEC (APB) subtype (n=1036)

- A) The expression of APOBEC3 family genes was significantly higher in APB subtype (all P-values < 2.2e-19, Mann-Whitney test).
- B) HPV infection was associated with higher immune response in TCGA-HNSC (n=507). Besides, even in HPV-negative tumors, APOBEC subtype showed a higher immune response than the other groups.
- C) (Left) The *PIK3CA* mutation rate was the highest (26.4%) in the APB subtype excluding hyper mutator subtype (MRD and POL). Chi-square test $P=1.7e-29$. (Right), PIK3_AKT_MTOR pathway score was the highest in the APB subtype. Mann-Whitney $P=2.3e-18$. These results are consistent with the previously reported association between *PIK3CA* mutation and APOBEC-mediated cytosine deamination³⁴.

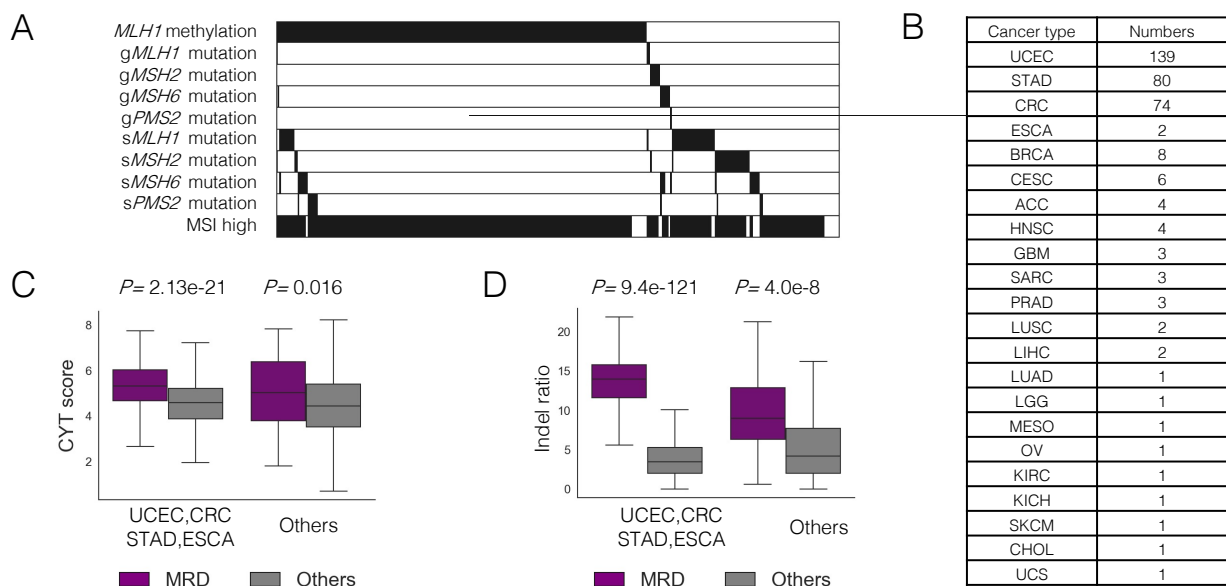


Figure S5. Features of MISMATCH REPAIR DEFICIENCY (MRD) subtype (n=339)

- A) 86.4% had MMR gene alterations (somatic and germline mutation in *MLH1*, *MSH2*, *MSH6*, and *PMS2*, and *MLH1* methylation).
- B) Cancer types of MRD subtype.
- C) Excluding cancer types with relatively high frequently of MSI-high tumors (UCEC, CRC, STAD, ESCA), immune-related gene expression scores were still higher in MRD subtype than the others ($P=0.016$, Man-Whitney U test)
- D) Excluding cancer types with relatively high frequently of MSI-high tumors (UCEC, CRC, STAD, ESCA), indel ratio was still higher in MRD subtype than the others ($P=4.0e-8$, Man-Whitney U test)

ACC, Adrenocortical carcinoma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, Cholangiocarcinoma; CRC, Colorectal adenocarcinoma; ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and neck squamous cell carcinoma; KICH, Kidney chromophobe; KIRC, Kidney renal clear cell carcinoma; LGG, Brain lower grade glioma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; MESO, Mesothelioma; OV, Ovarian serous cystadenocarcinoma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; SKCM, Skin cutaneous melanoma; STAD, Stomach adenocarcinoma; UCEC, Uterine corpus endometrial carcinoma; UCS, Uterine carcinosarcoma

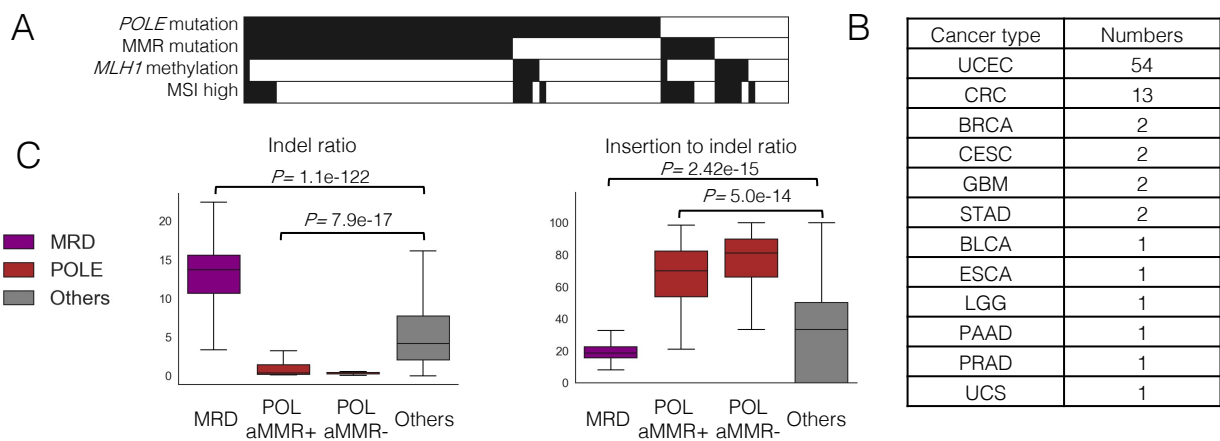


Figure S6. Features of *POLE* (POL) subtype (n=81)

- A) Somatic *POLE* mutation was observed in 76.5% of POL subtype, and MMR-gene mutation (somatic and germline mutation in *MLH1*, *MSH2*, *MSH6*, and *PMS2*) was observed in 59.3% of POL subtype.
- B) Cancer types of POL subtype
- C) POL subtype, both with and without MMR mutation, had low indel ratio (left) and high insertion to indel ratio (right) than other subtypes, in contrast to MRD subtype. These data indicate that most tumors with concurrent *POLE* and MMR-gene mutations have acquired *POLE* mutation first, as previously reported³⁵, and thus have the *POLE* mutation-dominant underlying mutational processes.

BLCA, Bladder urothelial carcinoma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CRC, Colorectal adenocarcinoma; ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; LGG, Brain lower grade glioma; PAAD, Pancreatic adenocarcinoma; PRAD, Prostate adenocarcinoma; UCEC, Uterine corpus endometrial carcinoma; UCS, Uterine carcinosarcoma

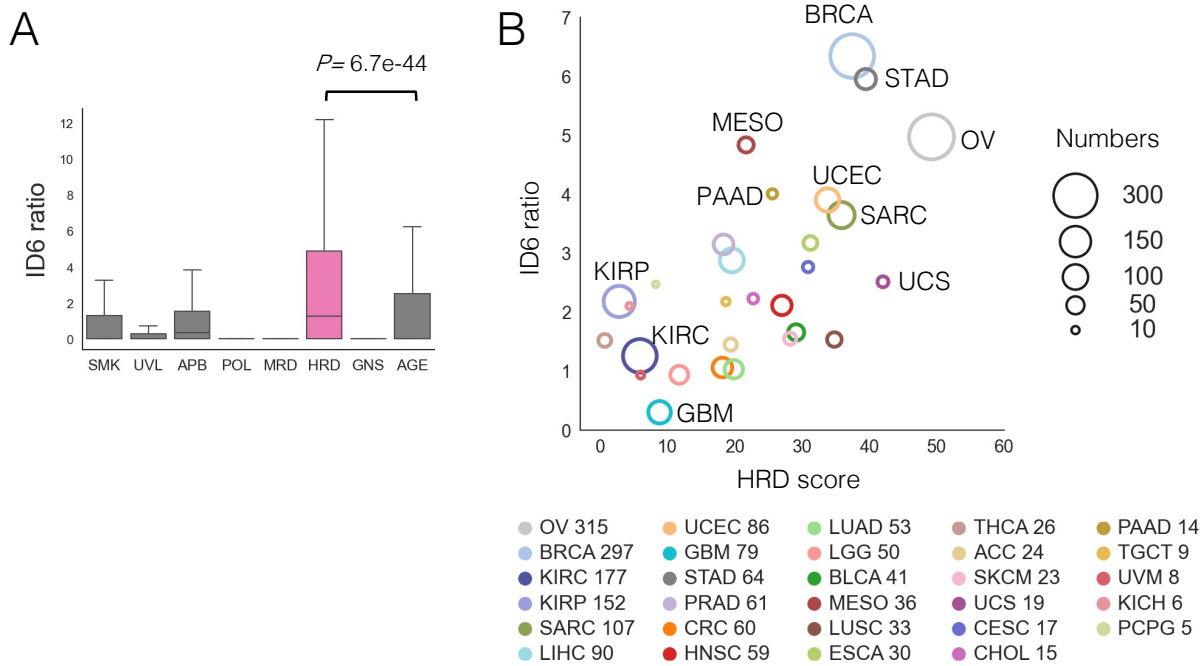


Figure S7. Features of HOMOLOGOUS RECOMBINATION DEFICIENCY (HRD) subtype (n=1956)

- A) The ratio of indel signature 6, which is related to homologous recombination deficiency, was higher in HRD subtype than in other subtypes.
- B) Association between mean HRD scores (x-axis) and indel signature 6 ratios (y-axis) per cancer type calculated in tumors classified into HRD subtype. These two values were simultaneously high in cancer types such as OV, BRCA, SARC, UCEC, and STAD, which are known to include a certain proportion of HRD phenotypes in previous reports ³⁶. On the other hand, cancer types such as KIRP, KIRC, and GBM showed low values in both, suggesting that the HRD subtype may include tumors which do not harbor HRD phenotype.

ACC, Adrenocortical carcinoma; BLCA, Bladder urothelial carcinoma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, Cholangiocarcinoma; CRC, Colorectal adenocarcinoma; ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and neck squamous cell carcinoma; KICH, Kidney chromophobe; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LGG, Brain lower grade glioma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; MESO, Mesothelioma; OV, Ovarian serous cystadenocarcinoma; PAAD, Pancreatic adenocarcinoma; PCPG, Pheochromocytoma and paraganglioma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; SKCM, Skin cutaneous melanoma; STAD, Stomach adenocarcinoma; TGCT, Testicular germ cell tumors; THCA, Thyroid carcinoma; UCEC, Uterine corpus endometrial carcinoma; UCS, Uterine carcinosarcoma; UVM, Uveal melanoma.

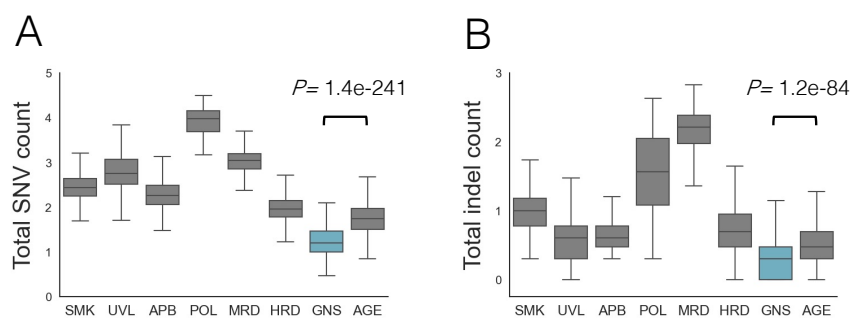


Figure S8. Features of GENOMICALLY STABLE (GNS) subtype (n=909)

Not only the total number of SNVs (A), but also that of insertions and deletions (B) were the lowest in GNS among all subtypes.

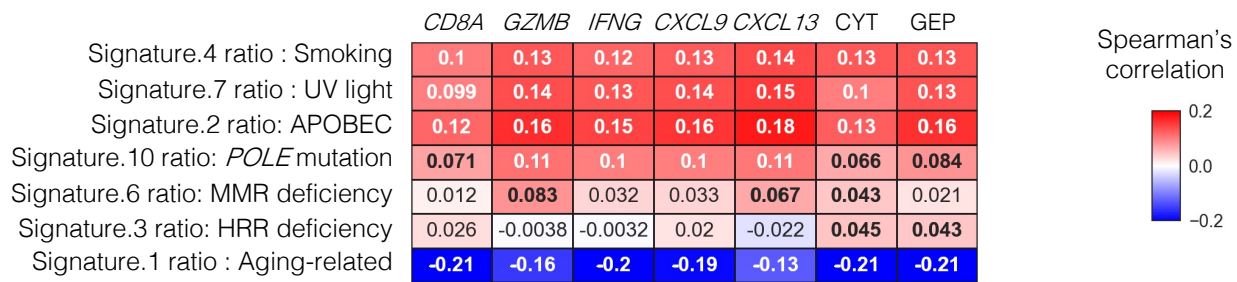


Figure S9. Correlation of representative etiological signatures with immune-related gene expression scores

Spearman's correlation coefficients between the ratios of representative etiological signatures and tumor immunity-related gene expression scores in all samples. Coefficients with p-values < 0.05 are marked in bold. The signature1 ratio was significantly negatively correlated with the immune-related scores, indicating age-related gene mutations may produce no or little immunogenicity.

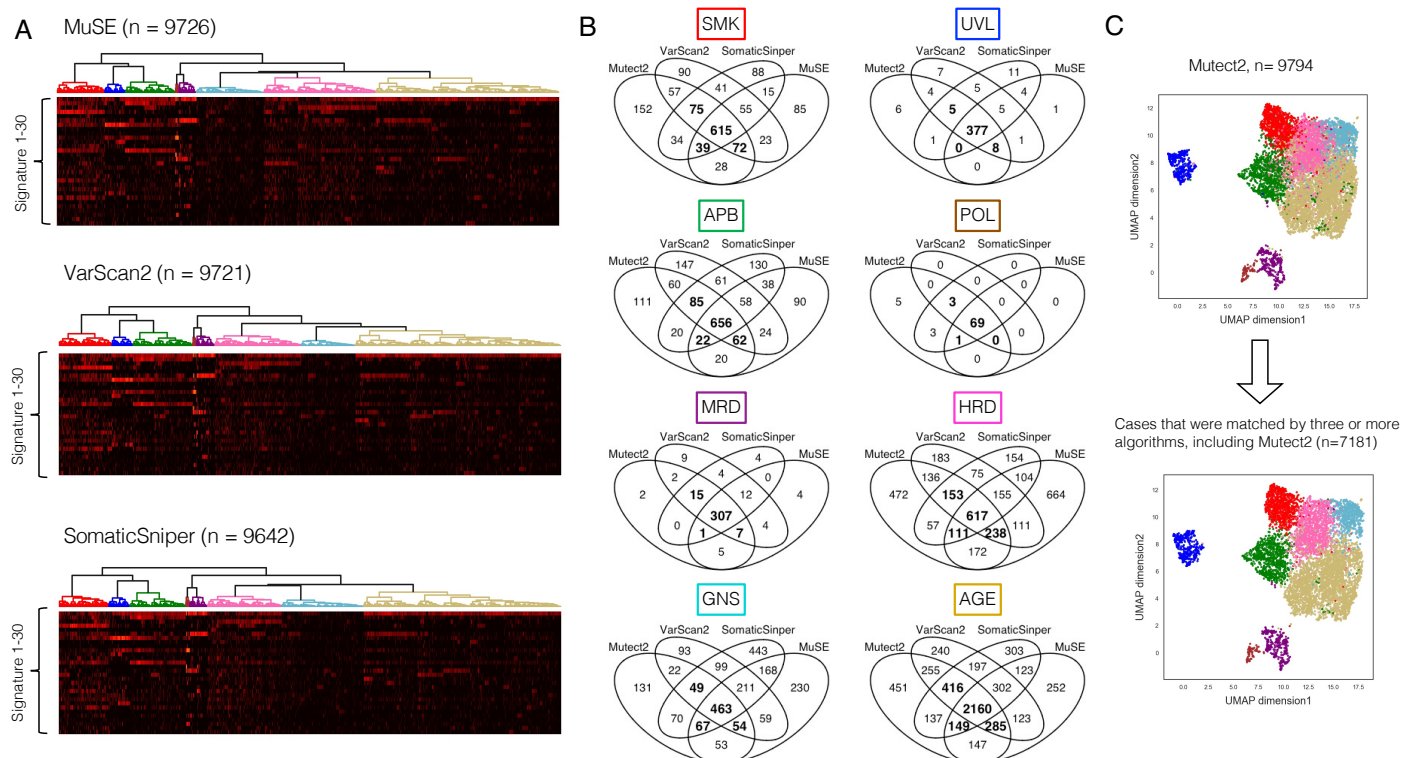


Figure S10. Consistency of clustering results from different variant callers

A) The results of hierarchical clustering based on the somatic mutation profiles derived from MuSE (upper), VarScan2 (middle), and SomaticSniper (lower).

Using these methods of annotation, eight genomic subtypes were created, as in the case using Mutect2.

B) Venn diagrams showing the overlapping of results from the four variant callers per subtype. GNS, HRD, and AGE subtypes showed relatively low consistency. To extract tumors typical for each subtype, cases that were matched by three or more algorithms, including Mutect2, were used for the next analysis (numbers in bold).

C) The UMAP clustering shows that the sample selection described in B reduces the number of cases in the border region of the subtype heatmap.

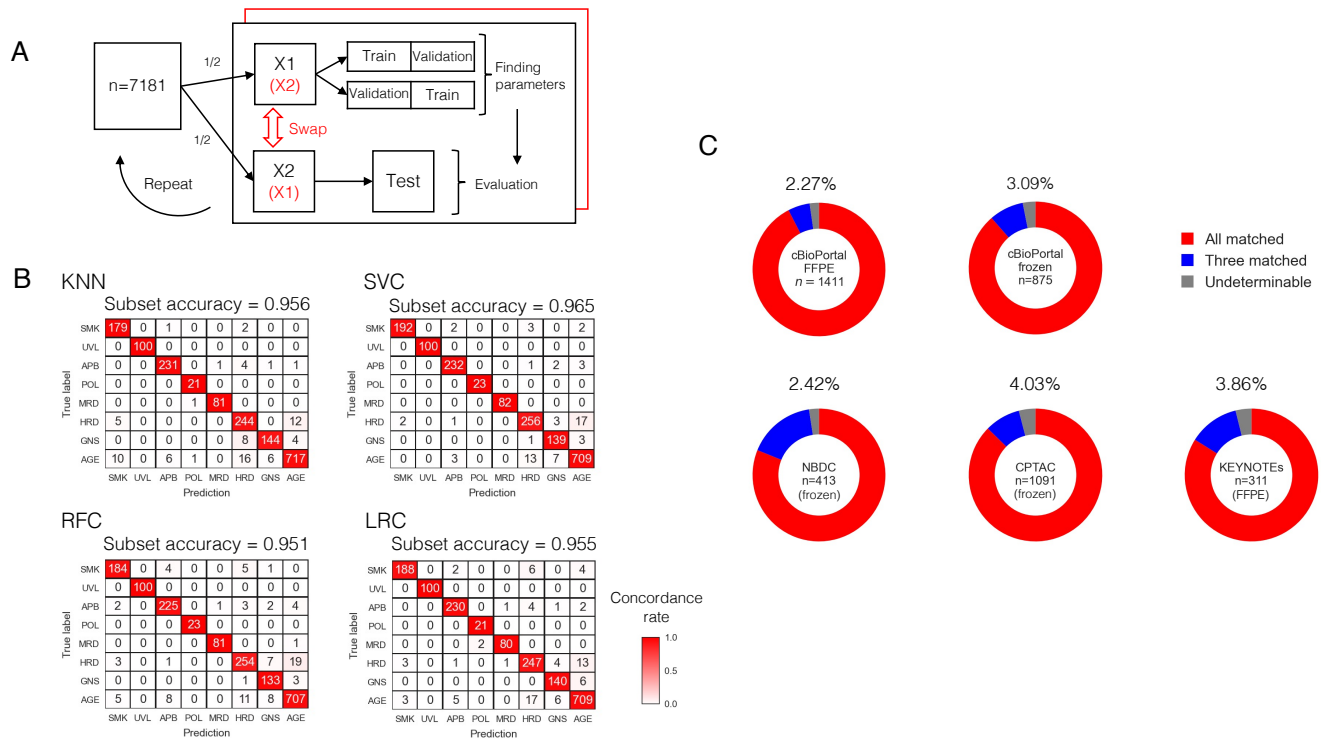


Figure S11. Developing classifiers through machine learning algorithms

- A) Double cross-validation and hyperparameter tuning. First, the selected 7181 TCGA samples were divided into two parts, X1 and X2. Second, parameters were calculated using X1 by two-fold cross-validation, and those parameters were evaluated using X2 as test data. Third, X1 and X2 were swapped, and the same calculations were performed. These processes were repeated 100-1000 times to determine the optimal parameters.
- B) Confusion matrices showing the classification performance on test data for the four classifiers: K-nearest neighbor (KNN), support vector machine (SVC), random forest (RFC), and logistic regression (LRC). Results using 75% of all cases for training and 25% for testing are shown. All showed more than 95% subset accuracy (exact match ratio).
- C) Consistency of subtyping results between the four classifiers per data group. When three or more of the four classifier results do not match, the sample is annotated as "undeterminable". Undeterminable samples were found in approximately 2-4% in the studied data groups, and there was no significant difference in the proportion between FFPE samples and frozen tissue origins or between data groups

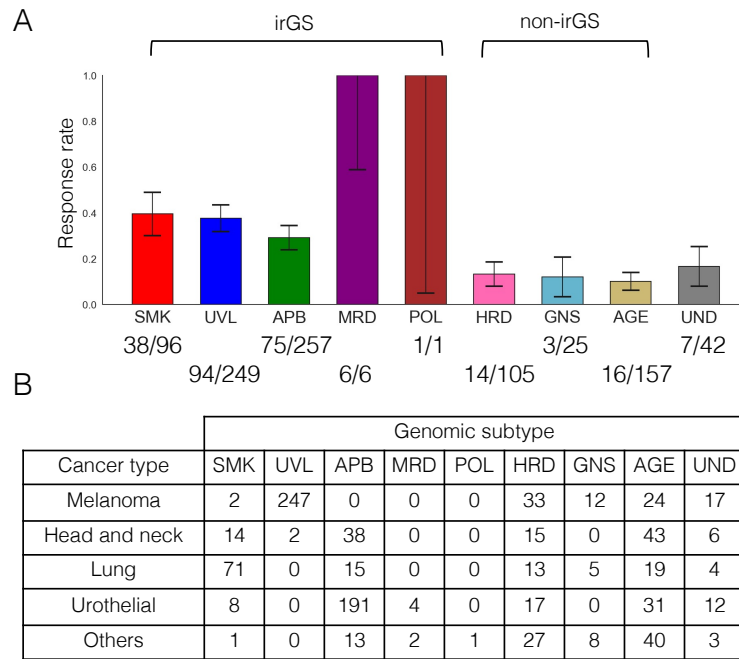


Figure S12. Relationship between tumor genomic subtype, response rate and cancer type in the whole cohort

- A) ICI response rate for each tumor genomic subtype. Subtypes included in irGS (SMK, UVL, APB, MRD, POL) showed a higher response rate than subtypes included in non-irGS (HRD, GNS, AGE).
- B) Distribution of the samples by tumor genomic subtype and by cancer type.

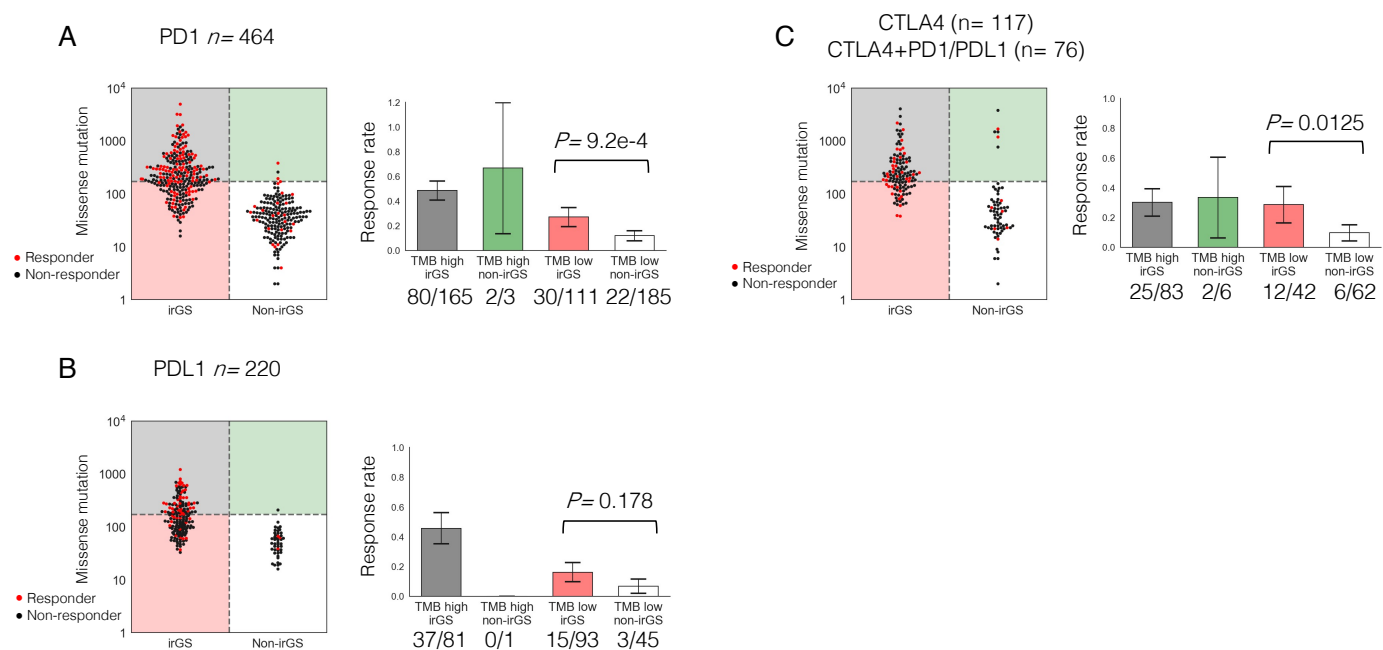


Figure S13. Studies examined by type of drug

A) Study in the cases treated with anti-PD1 antibody.

B) Study in the cases treated with anti-PDL1 antibody.

C) Study in the cases treated with anti-CTLA antibody or anti-CTLA antibody plus other ICIs.

In all studies, ICI response rate tended to be higher in irGS compared to non-irGS within TMB low cases.

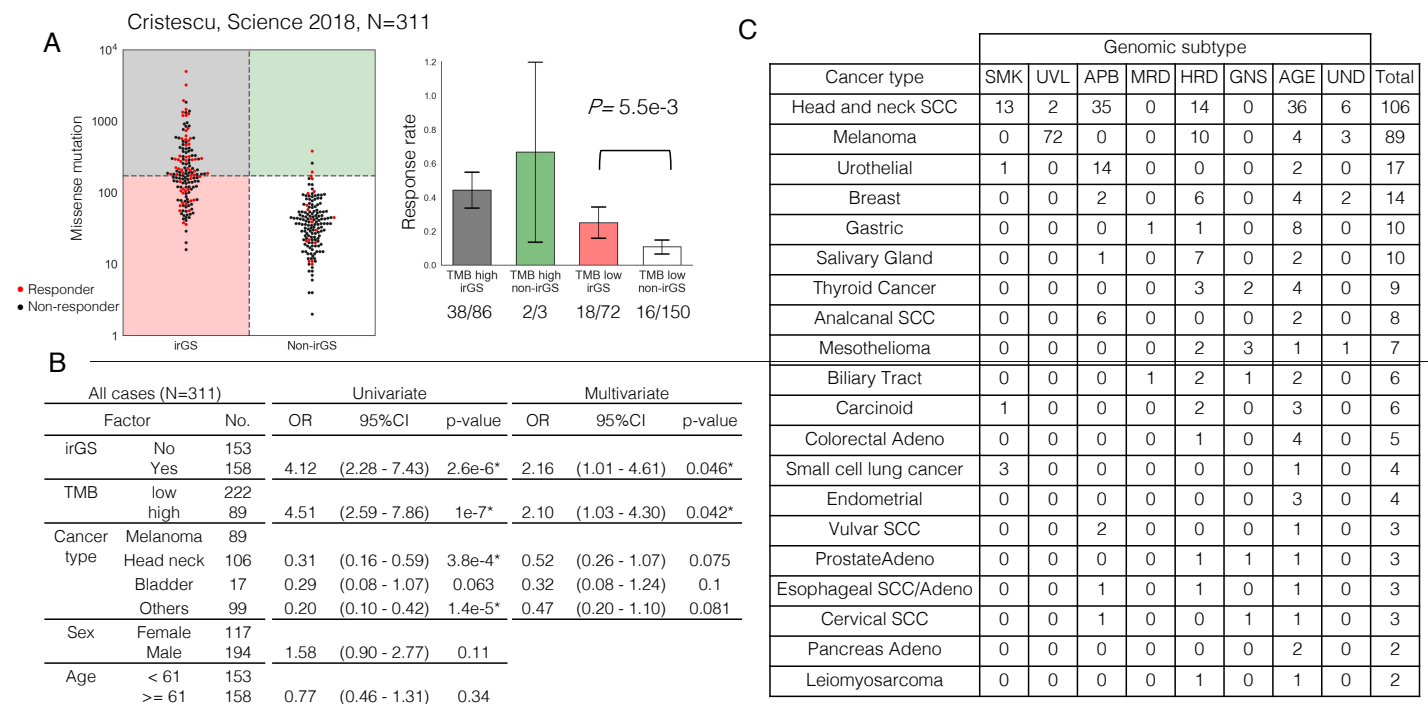


Figure S14. Study in the dataset from the KEYNOTE trials, which were prospective cohort studies of patients treated with solely pembrolizumab (n=311)

- A) Association between TMB and ICI response per sample divided by irGS status (left) and comparison of ICI response rate in the four groups stratified by irGS and TMB status (right). irGS showed a significantly higher response rate than non-irGS within the sample classified as TMB low.
- B) Univariate and multivariate logistic regression analysis for ICI response. irGS status was significantly associated with the ICI response after adjusting by TMB status and cancer type.
- C) Distribution of the samples by tumor genomic subtype and by cancer type. Although the KEYNOTE trials excluded patients with clinically diagnosed MSI high tumors at enrollment, two tumors from the cohort (one each with gastric cancer and biliary tract cancer) were classified as MRD subtype, and both of them responded to ICI.

irGS, immune-reactive genomic subtype; TMB, Tumor mutational burden; OR, Odds ratio; CI, confidence interval; SCC, squamous cell carcinoma

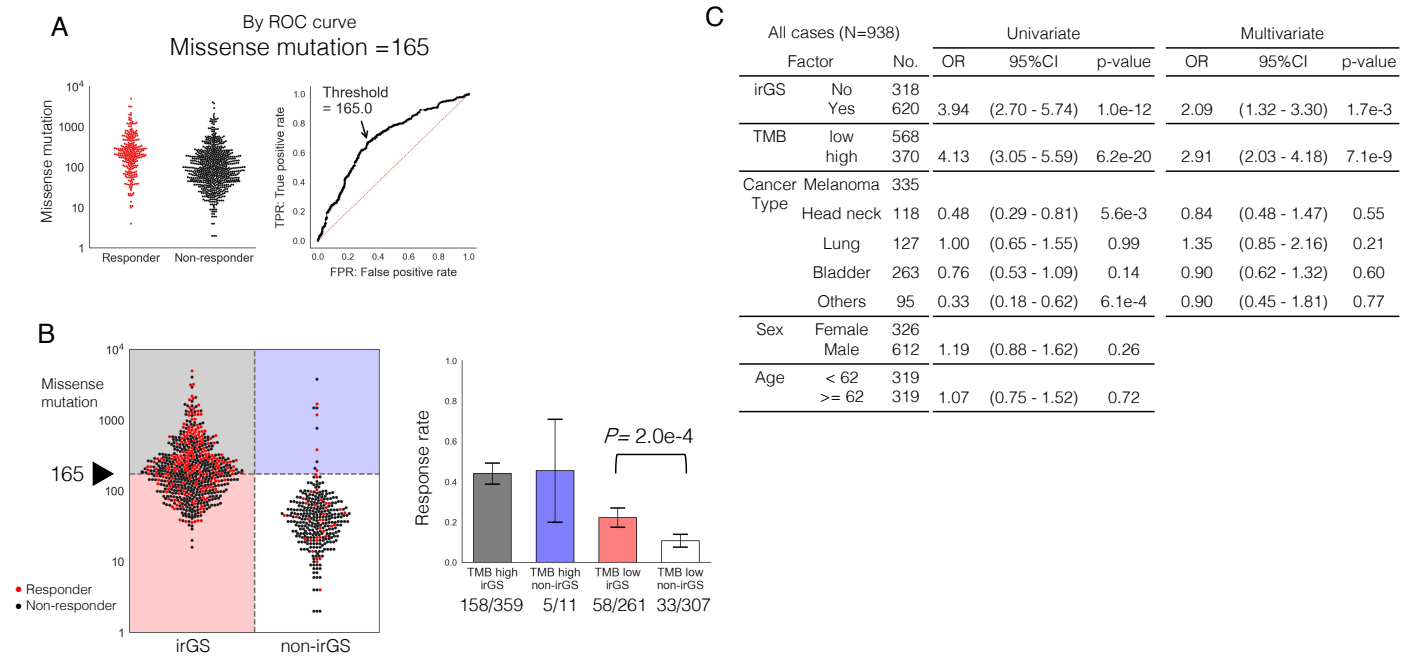


Figure S15. Study using the cohorts' optimal cutoff of TMB high

- C) The cohort's optimal TMB cutoff determined by the ROC curve and the Youden index for objective responses in the whole cohort (N=938) was 163 missense mutations, which was close to 173, the value calculated Figure3B.
- D) Association between TMB and ICI response per sample divided by irGS status (left) and comparison of ICI response rate in the four groups stratified by irGS and TMB status (right). irGS showed a significantly higher response rate than non-irGS within the sample classified as TMB low.
- E) Univariate and multivariate logistic regression analysis for ICI response in the validation cohort (N=938). irGS status was significantly associated with the ICI response after adjusting by TMB status and cancer type.

irGS, immune-reactive genomic subtype; TMB, Tumor mutational burden; OR, Odds ratio; CI, confidence interval

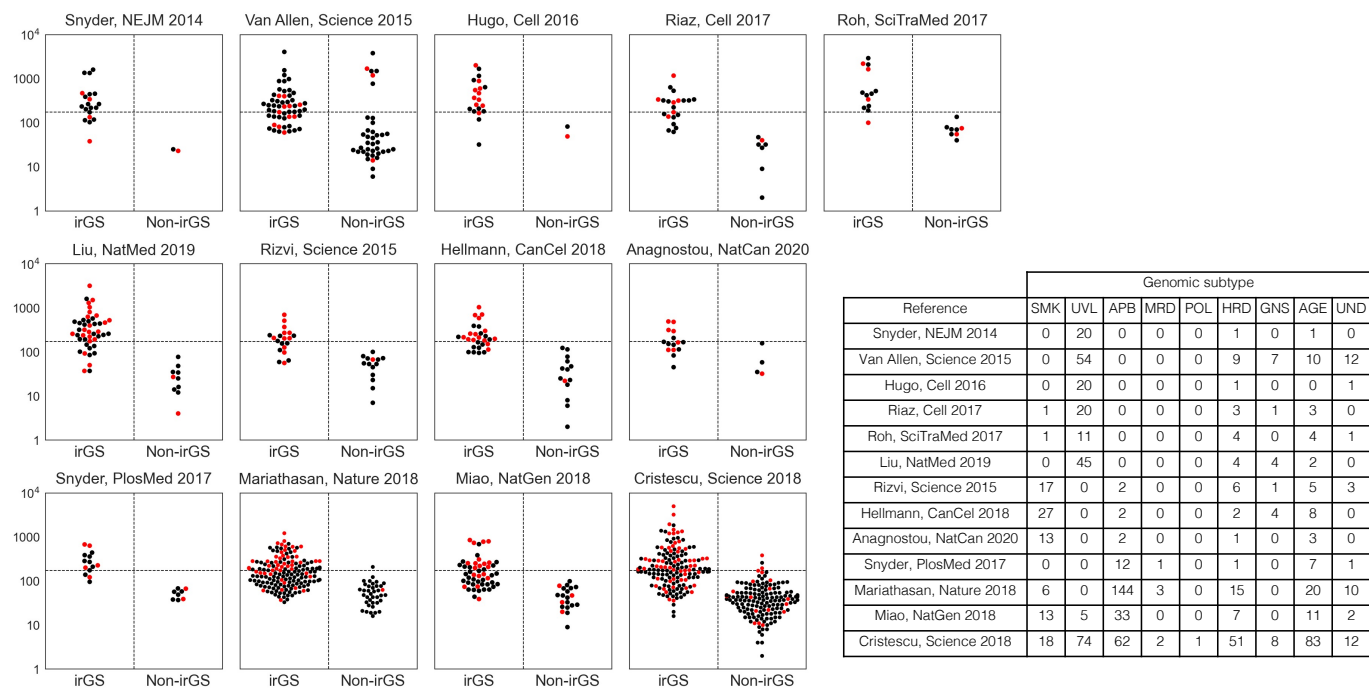


Figure S16. Association between TMB and ICI response divided by irGS status per dataset

The right table indicates the distribution of samples for each subtype per dataset. See also Table S3.

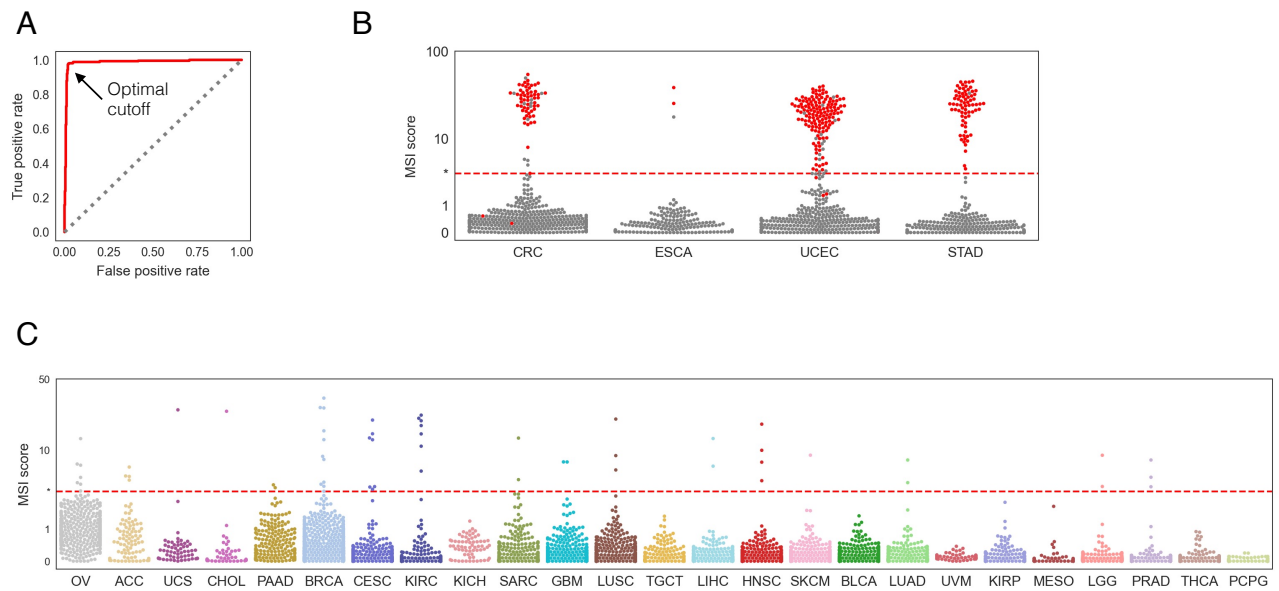


Figure S17. Determination of MSI-high cases using MSIsensor

- A,B) Optimal cutoff of MSI score to discriminate cases with MSI-high annotations was calculated using ROC curve and Youden index from the datasets of CRC, ESCA, STAD and UCEC.
- C) The relationship between MSI score and the cutoff value in cancer types other than the four above. Cases exceeding the cutoff were annotated as MSI high.

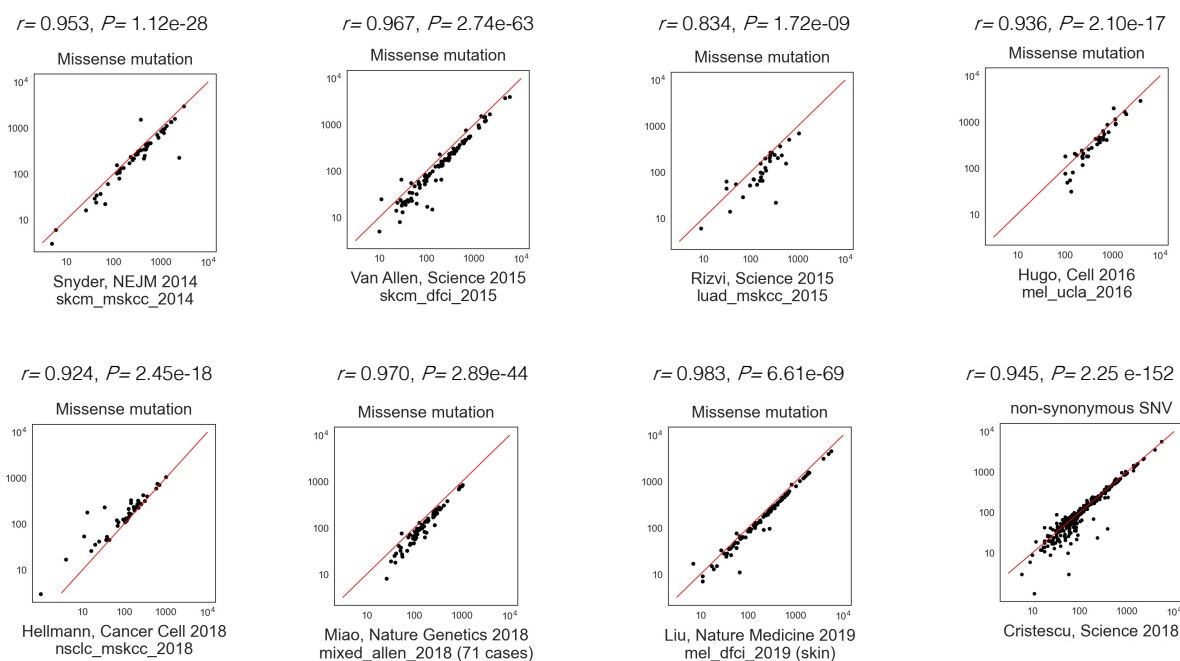


Figure S18. Comparison of the number of missense mutations or non-synonymous SNVs from our WES pipeline and previously published data

The red line represents a straight line with slope 1 reaching the zero point. The r value and p value at the top of each panel were calculated using the Pearson's correlation. Most of the datasets have Pearson correlations greater than 0.9.

References

1. Cerami E et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2012 May;2(5):401-4. doi: 10.1158/2159-8290.CD-12-0095.
2. Broad GDAC Firebrowse. Available at: <https://gdac.broadinstitute.org/>
3. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol.* 2013 Mar;31(3):213-9. doi: 10.1038/nbt.2514. Epub 2013 Feb 10.
4. Genomic Data Commons Data Portal. Available at: <https://portal.gdc.cancer.gov/>
5. Vaddepally RK et al. Review of Indications of FDA-Approved Immune Checkpoint Inhibitors per NCCN Guidelines with the Level of Evidence. *Cancers (Basel).* 2020 Mar 20;12(3):738. doi: 10.3390/cancers12030738.
6. Yarchoan M et al. Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *N Engl J Med.* 2017 Dec 21;377(25):2500-2501. doi: 10.1056/NEJMc1713444.
7. Yarchoan M et al. PD-L1 expression and tumor mutational burden are independent biomarkers in most cancers. *JCI Insight.* 2019 Mar 21;4(6):e126908. doi: 10.1172/jci.insight.126908. eCollection 2019 Mar 21.
8. Gómez-Raposo C et al. Immune checkpoint inhibitors in endometrial cancer. *Crit Rev Oncol Hematol.* 2021 May;161:103306. doi: 10.1016/j.critrevonc.2021.103306. Epub 2021 Apr 8.
9. Blokzijl F et al. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* 2018 Apr 25;10(1):33. doi: 10.1186/s13073-018-0539-0.
10. Alexandrov LB et al. Clock-like mutational processes in human somatic cells. *Nat Genet.* 2015 Dec;47(12):1402-7.
11. Koboldt DC et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012 Mar;22(3):568-76. doi: 10.1101/gr.129684.111. Epub 2012 Feb 2.
12. Fan Y et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* 2016 Aug 24;17(1):178. doi: 10.1186/s13059-016-1029-6.
13. Larson DE et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics.* 2012 Feb 1;28(3):311-7. doi: 10.1093/bioinformatics/btr665. Epub 2011 Dec 6.
14. Huang KL et al. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell.* 2018 Apr 5;173(2):355-370.e14. doi: 10.1016/j.cell.2018.03.039.
15. Takamatsu S et al. Utility of Homologous Recombination Deficiency Biomarkers Across Cancer Types. *JCO Precis Oncol.* 2021 Aug 11;5:PO.21.00141. doi: 10.1200/PO.21.00141. eCollection 2021 Aug.
16. Knijnenburg TA et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* 2018 Apr 3;23(1):239-254.e6. doi: 10.1016/j.celrep.2018.03.076.
17. Catalogue Of Somatic Mutations In Cancer (COSMIC) Small insertions and deletions (ID) Signatures (v3.2 March 2021). Available at: <https://cancer.sanger.ac.uk/signatures/id/>
18. Hübschmann D et al. Analysis of mutational signatures with yet another package for signature analysis. *Genes Chromosomes Cancer.* 2021 May;60(5):314-331. doi: 10.1002/gcc.22918. Epub 2020 Dec 31.
19. Niu B et al. MSIensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics.* 2014 Apr 1;30(7):1015-6. doi: 10.1093/bioinformatics/btt755. Epub 2013 Dec 25.
20. Jurtz V et al. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol.* 2017 Nov 1;199(9):3360-3368. doi: 10.4049/jimmunol.1700893. Epub 2017 Oct 4.
21. Thorsson V et al. The Immune Landscape of Cancer. *Immunity.* 2018 Apr 17;48(4):812-830.e14. doi: 10.1016/j.immuni.2018.03.023. Epub 2018 Apr 5.

22. Rooney MS et al. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*. 2015 Jan 15;160(1-2):48-61. doi: 10.1016/j.cell.2014.12.033.
23. Molecular Signatures Database (MSigDB). Available at: <https://www.gsea-msigdb.org/gsea/msigdb/>
24. Barbie DA et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009 Nov 5;462(7269):108-12. doi: 10.1038/nature08460. Epub 2009 Oct 21.
25. Cristescu R et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science*. 2018 Oct 12;362(6411):eaar3593. doi: 10.1126/science.aar3593.
26. National Bioscience Database Center (NBDC) Human Database. Available at: <https://humandbs.biosciencedbc.jp>
27. Snyder A et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med*. 2014 Dec 4;371(23):2189-2199. doi: 10.1056/NEJMoa1406498. Epub 2014 Nov 19.
28. Anagnostou V et al. Multimodal genomic features predict outcome of immune checkpoint blockade in non-small-cell lung cancer. *Nat Cancer*. 2020 Jan;1(1):99-111. doi: 10.1038/s43018-019-0008-8. Epub 2020 Jan 13.
29. McKenna A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19.
30. Babraham Bioinformatics. Trim Galore! Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore
31. Li H et al. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18.
32. Hellmann MD et al. Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. *Cancer Cell*. 2018 May 14;33(5):843-852.e4. doi: 10.1016/j.ccell.2018.03.018. Epub 2018 Apr 12.
33. Litchfield K et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell*. 2021 Feb 4;184(3):596-614.e14. doi: 10.1016/j.cell.2021.01.002. Epub 2021 Jan 27.
34. Henderson S et al. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep*. 2014 Jun 26;7(6):1833-41. doi: 10.1016/j.celrep.2014.05.012. Epub 2014 Jun 5. PMID: 24910434.
35. Haradhvala NJ et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat Commun*. 2018 May 1;9(1):1746. doi: 10.1038/s41467-018-04002-4. PMID: 29717118; PMCID: PMC5931517.
36. Takamatsu S et al. Utility of Homologous Recombination Deficiency Biomarkers Across Cancer Types. *JCO Precis Oncol*. 2021 Aug 11;5:PO.21.00141. doi: 10.1200/PO.21.00141. PMID: 34423229; PMCID: PMC8373547.