**CPTAC GENCODE V34 harmonized meta table README**

Yongchao Dou, Zhiao Shi, Sara Savage, Wen Jiang, and Bing Zhang

**Clinical data**

The clinical data used in the portal were collected from CPTAC with the May 2022 update. Age was truncated to 90 years. Tumors with a size <=0 were replaced with NA. For overall survival analysis, cases were removed if a death occurred within 30 days of initial diagnosis and for progression free survival analysis, cases were removed with follow up or a new tumor event that occurred within 10 days of the initial diagnosis.

**CIN score**

The chromosome instability (CIN) score reflects the overall copy number aberration across the whole genome. From the segmentation result, we used a weighted-sum approach to summarize the chromosome instability for each sample[(Vasaikar et al., 2019)](https://paperpile.com/c/4rdwyJ/23ze). The absolute segment level log2 ratios of all segments (indicating the copy number aberration of these segments) within a chromosome were weighted by the segment length and summed up to derive the instability score for the chromosome. The genome-wide chromosome instability index was calculated by summing up the instability score of all 22 autosomes. The R package genomicWidgets was used to implement the method (https://github.com/bzhanglab/genomicWidgets).

**Mutation burden**

Tumor mutation burden (TMB) was extracted from the mutation annotation file (maf) by the maftools R package (V2.10.0)[(Mayakonda et al., 2018)](https://paperpile.com/c/4rdwyJ/cYVUm). It was calculated as the number of non-synonymous variants per million bp.

**Immune deconvolution**

The R package immunedeconv (V2.0.4)[(Sturm et al., 2019)](https://paperpile.com/c/4rdwyJ/Vpgq2) was used to perform immune cell deconvolution using RNA expression data (TPM). Among the seven deconvolution methods in immunedeconv, CIBERSORT[(Newman et al., 2015)](https://paperpile.com/c/4rdwyJ/GD8KR) and xCell[(Aran et al., 2017)](https://paperpile.com/c/4rdwyJ/Qi59W) were selected in our analysis. CIBERSORT was performed in the ‘abs’ mode.

**ESTIMATE**

The ESTIMATE scores reflecting the overall immune and stromal infiltration were calculated by the R package ESTIMATE[(Yoshihara et al., 2013)](https://paperpile.com/c/4rdwyJ/AJqRx) using the normalized RNA expression data (RSEM). We removed genes with 0 expression in >=50% samples of a cohort.

**PROGENy score**

The PROGENy scores were inferred using the R package progeny (V1.10.0)[(Schubert et al., 2018)](https://paperpile.com/c/4rdwyJ/HiCml) with default parameters using the RNA expression data (FPKM). Genes with mean expression = 0 in a cohort were removed from the analysis.

**MSigDB hallmark pathway single sample gene set enrichment analysis (ssGSEA)**

ssGSEA was performed for each cancer type using gene-wise Z-scores of the RNA expression data (RSEM) for the MSigDB Hallmark gene sets v7.0[(Liberzon et al., 2015)](https://paperpile.com/c/4rdwyJ/YPYiQ) via the ssGSEA2.0 R package[(Krug et al., 2019)](https://paperpile.com/c/4rdwyJ/IUSDA). RNA data were filtered to coding genes with < 50% 0 expression. (Parameters: sample.norm.type="rank", weight=0.75, statistic="area.under.RES", nperm=1000, min.overlap=10). Pathway activity scores are normalized enrichment scores from ssGSEA.

**Phosphosite signature scores**

Phosphosite signature scores were calculated using the PTMsigDB v1.9.0 database and the ssGSEA2.0 R package[(Krug et al., 2019)](https://paperpile.com/c/4rdwyJ/IUSDA). The parameters were the same as those used for Hallmark pathway activity (sample.norm.type="rank", weight=0.75, statistic="area.under.RES", nperm=1000, min.overlap=10). Phosphoproteomics data were filtered to the fifteenmer phosphosites with complete data across all samples within a cohort. If there were multiple rows with complete data for identical fifteenmers, one row was selected at random. Each site was z-score transformed. Activity scores are normalized enrichment scores from ssGSEA.

**Mutation signature calling**

The R package SigProfilerMatrixGeneratorR[(Bergstrom et al., 2019)](https://paperpile.com/c/4rdwyJ/9XGD0) (version 1.0) was used to call mutation signatures from WES-derived somatic mutation data. All synonymous and non-synonymous mutations were included. The maximum number of signatures was set to 10 and nmf replicates parameter was set to 100. The activity scores of the decomposed solution suggested by SigProfilerMatrixGenerator were used as signature scores.

**Tumor purity**

The DoAbsolute R package (V2.2)[(Wang et al., 2019)](https://paperpile.com/c/4rdwyJ/fokGG) with ABSOLUTE (V1.0.6)[(Carter et al., 2012)](https://paperpile.com/c/4rdwyJ/pMIiV) was used to infer tumor purity and ploidy from somatic mutations and WES-based CNV. The parameters min.mut.af and max.as.seg.count were set to 0.02 and 5000, respectively. All other parameters were set as default. These results are referred to as Tumor Purity (ABSOLUTE) in the portal. Additionally, CNVEX[(Clark et al., 2019)](https://paperpile.com/c/4rdwyJ/r0gb) was used to infer tumor purity using both whole genome sequencing (WGS) and WES data (<https://github.com/mctp/cnvex>) by the University of Michigan team, and these results are referred to as Tumor Purity (WGS) in the portal.

Related file(s):

XXXX\_meta.txt

Contact:

If you have questions about this pipeline, please contact Yongchao Dou at BCM: [ydou@bcm.edu](mailto:ydou@bcm.edu)