

Supplementary Meterial for “Substitution mutational signatures across pan-squamous cell carcinomas”

Shun Wang, Guowei Zhao, Hongrui Li, Yi Zhang, Mengfei Yao, Jianming Ying and Wenbin Li

Supplementary Tables

Table 1 Squamous cell carcinoma cohorts in this study

SCC	cohort	WGS/WES	sample	sample (full Sig cos ≥ 0.8)	sample (rep Sig cos ≥ 0.8)	Source
CECSC	tcga_cesc	WES	291	278	207	[1, 2]
ESCC	tcga_esca*	WES	97	77	26	[3]
	icgc_escc_cn	WGS	332	213	171	[4]
	icgc_escc	WGS	88	62	43	[5, 6]
	pumc_escc	WGS	508	404	354	[7]
	ucla_escc	WES	137	41	26	[8, 9]
HNSC	broad_hnsc	WES	74	55	51	[10, 11]
	jhu_hnsc	WES	32	13	7	[12, 13]
	mdanderson_hnsc	WES	40	27	7	[14, 15]
	cptac_hnsc**	WES	108	92	46	[16]
	tcga_hnsc	WES	509	462	200	[17, 18]
	icgc_orca	WES	178	136	108	[19]
LSCC	cptac_lsc	WES	108	98	69	[20, 21]
	icgc_lsc	WGS	170	136	62	[22]
	pumc_lsc	WES	109	87	77	†
	tcga_lsc	WES	485	461	366	[23, 24]

* Samples whose primary pathology histological types marked as Esophagus Squamous Cell Carcinoma were enrolled in our study

** Only HPV-negative samples

† Unpublished data in our hospital, details can be accessed upon request

Table 2 NCC-CLINDx Samples for Validation

	sample	sample (rep Sig cos ≥ 0.2)
CECSC	38	30
ESCC	159	144
HNSC	267	212
<i>Hypopharynx</i>	94	82
<i>Larynx</i>	27	23
<i>Oral Cavity</i>	58	37
<i>Oropharynx</i>	62	49
<i>Others</i>	26	21
LUSC	220	186

Table 3 SBS fractions of each cohort, original data of Figure 1

Cohort	APOBEC	DDR	SBS1	SBS10b	SBS5	SBS7ab
tcga_cesc	48.91%	5.20%	0.00%	4.82%	25.78%	2.98%
tcga_esca	19.85%	2.62%	0.00%	1.86%	31.13%	1.00%
icgc_escc_cn	20.81%	5.15%	0.00%	3.89%	32.18%	2.35%
icgc_escc	20.67%	4.81%	0.00%	3.58%	30.63%	3.65%
pumc_escc	25.42%	4.67%	0.00%	4.15%	34.62%	2.19%
ucla_escc	28.96%	2.87%	0.00%	3.24%	22.40%	2.11%
broad_hnsc	20.35%	3.92%	0.00%	3.16%	28.25%	1.31%
cptac_hnsc	18.40%	3.87%	0.00%	1.52%	36.91%	7.03%
jhu_hnsc	12.14%	14.45%	11.84%	4.20%	0.00%	5.97%
mdanderson_hnsc	12.26%	5.13%	0.00%	2.54%	33.36%	3.15%
tcga_hnsc	24.64%	3.60%	0.00%	2.43%	32.70%	3.41%
cptac_lsc	10.73%	0.56%	0.00%	0.55%	29.00%	1.23%
icgc_lsc	7.85%	2.38%	0.00%	0.22%	41.43%	1.37%
pumc_lsc	8.81%	2.43%	0.00%	0.67%	30.44%	0.77%
tcga_lsc	12.05%	1.10%	0.00%	0.26%	22.28%	1.70%
icgc_orca	12.54%	4.65%	0.00%	2.68%	19.89%	3.62%
Cohort	SBS4	SBS29	SBS24	SBS18	SBS16	Others
tcga_cesc	0.05%	0.26%	0.09%	0.38%	0.00%	11.53%
tcga_esca	1.53%	3.04%	0.20%	4.24%	14.61%	19.93%
icgc_escc_cn	0.17%	0.90%	0.94%	3.14%	2.00%	28.47%
icgc_escc	0.75%	0.65%	1.70%	5.06%	3.76%	24.73%
pumc_escc	0.08%	0.29%	0.14%	2.13%	2.48%	23.84%
ucla_escc	0.00%	0.41%	0.00%	0.46%	4.24%	35.31%
broad_hnsc	1.48%	1.45%	0.00%	2.43%	7.11%	30.53%
cptac_hnsc	1.45%	1.12%	0.73%	1.00%	6.56%	21.41%
jhu_hnsc	0.42%	1.25%	2.08%	0.66%	3.54%	43.46%
mdanderson_hnsc	1.02%	0.51%	0.00%	0.00%	5.79%	36.24%
tcga_hnsc	3.19%	0.92%	0.53%	1.65%	5.27%	21.66%
cptac_lsc	14.77%	5.74%	1.48%	1.37%	1.00%	33.56%
icgc_lsc	6.51%	2.47%	0.59%	1.04%	2.35%	33.79%
pumc_lsc	17.16%	3.96%	1.91%	1.73%	1.03%	31.09%
tcga_lsc	14.45%	3.85%	1.50%	2.05%	0.94%	39.82%
icgc_orca	0.80%	26.43%	6.26%	6.32%	0.43%	16.38%

References

- [1] Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, Barretos Cancer Hospital, Baylor College of Medicine, Beckman Research Institute of City of Hope, et al. Integrated genomic and molecular characterization of cervical cancer. *Nature*. 2017 mar;(7645):378–384.
- [2] Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA, Firehose Legacy);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=cesc_tcga.
- [3] Esophageal Carcinoma (TCGA, Firehose Legacy);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=esca_tcga.
- [4] Esophageal Cancer - CN;. Accessed October 4, 2023. <https://dcc.icgc.org/projects/ESCA-CN>.
- [5] Esophageal Squamous Cell Carcinoma (ICGC, Nature 2014);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=escc_icgc.
- [6] Song Y, Li L, Ou Y, Gao Z, Li E, Li X, et al. Identification of genomic alterations in oesophageal squamous cell cancer. *Nature*. 2014 may;509(7498):91–5.
- [7] Cui Y, Chen H, Xi R, Cui H, Zhao Y, Xu E, et al. Whole-genome sequencing of 508 patients identifies key molecular features associated with poor prognosis in esophageal squamous cell carcinoma. *Cell Research*. 2020;30(10):902–913.
- [8] Esophageal Squamous Cell Carcinoma (UCLA, Nat Genet 2014);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=escc_ucla_2014.
- [9] Lin DC, Hao JJ, Nagata Y, Xu L, Shang L, Meng X, et al. Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nature genetics*. 2014 may;46(5):467–73.
- [10] Head and Neck Squamous Cell Carcinoma (Broad, Science 2011);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=hnsk_broad.
- [11] Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. *Science*. 2011 aug;333(6046):1157–1160.
- [12] Head and Neck Squamous Cell Carcinoma (Johns Hopkins, Science 2011);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=hnsk_jhu.
- [13] Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science*. 2011 aug;333(6046):1154–1157.
- [14] Oral Squamous Cell Carcinoma (MD Anderson, Cancer Discov 2013);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=hnsk_mdanderson_2013.
- [15] Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, et al. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discovery*. 2013 jul;3(7):770–781.
- [16] Huang C, Chen L, Savage SR, Eguez RV, Dou Y, Li Y, et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell*. 2021;39(3):361–379.e16.
- [17] Head and Neck Squamous Cell Carcinoma (TCGA, Firehose Legacy);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=hnsk_tcga.

- [18] Lawrence MS, Sougnez C, Lichtenstein L, Cibulskis K, Lander E, Gabriel SB, et al. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015 jan;517(7536):576–582.
- [19] ORAL CANCER - IN;. Accessed October 4, 2023. <https://dcc.icgc.org/projects/ORCA-IN>.
- [20] Lung Squamous Cell Carcinoma (CPTAC, Cell 2021);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=lusc_cptac_2021.
- [21] Satpathy S, Krug K, Jean Beltran PM, Savage SR, Petralia F, Kumar-Sinha C, et al. A proteogenomic portrait of lung squamous cell carcinoma. *Cell*. 2021 aug;184(16):4348–4371.e40.
- [22] LUNG CANCER - KR;. Accessed October 4, 2023. <https://dcc.icgc.org/projects/LUSC-KR>.
- [23] Lung Squamous Cell Carcinoma (TCGA, Firehose Legacy);. Accessed October 4, 2023. https://www.cbioportal.org/study?id=lusc_tcga.
- [24] Hammerman PS, Voet D, Lawrence MS, Voet D, Jing R, Cibulskis K, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012 sep;489(7417):519–525.

Supplementary Figures

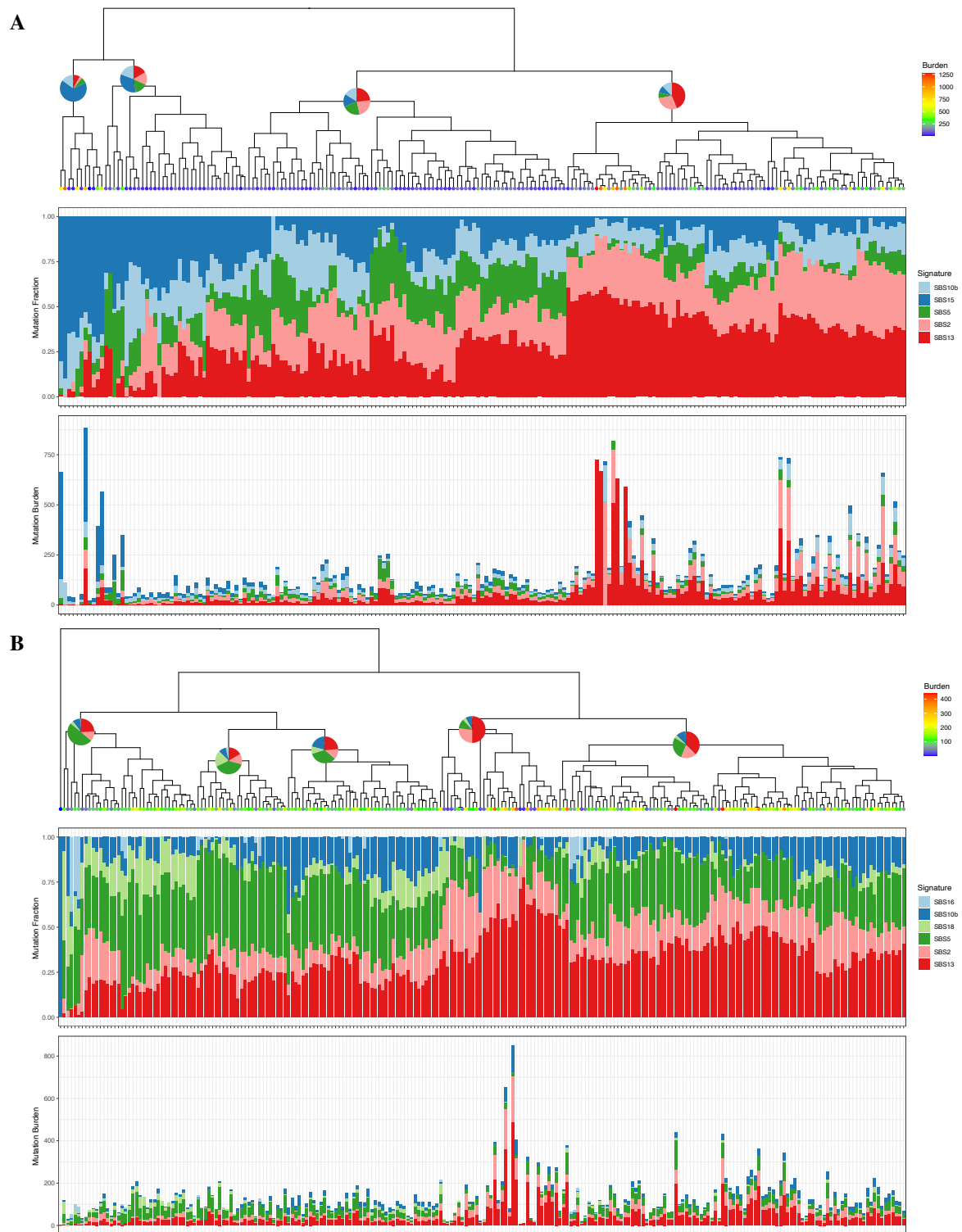


Figure 1 Continuing in next page

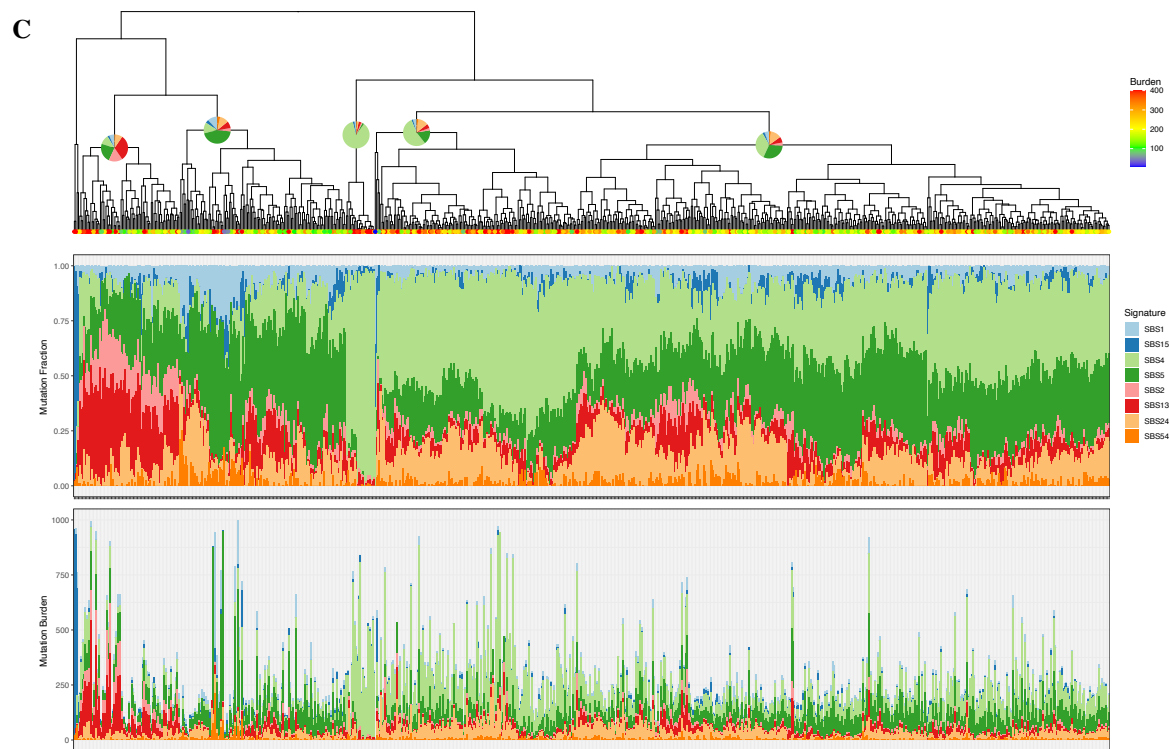
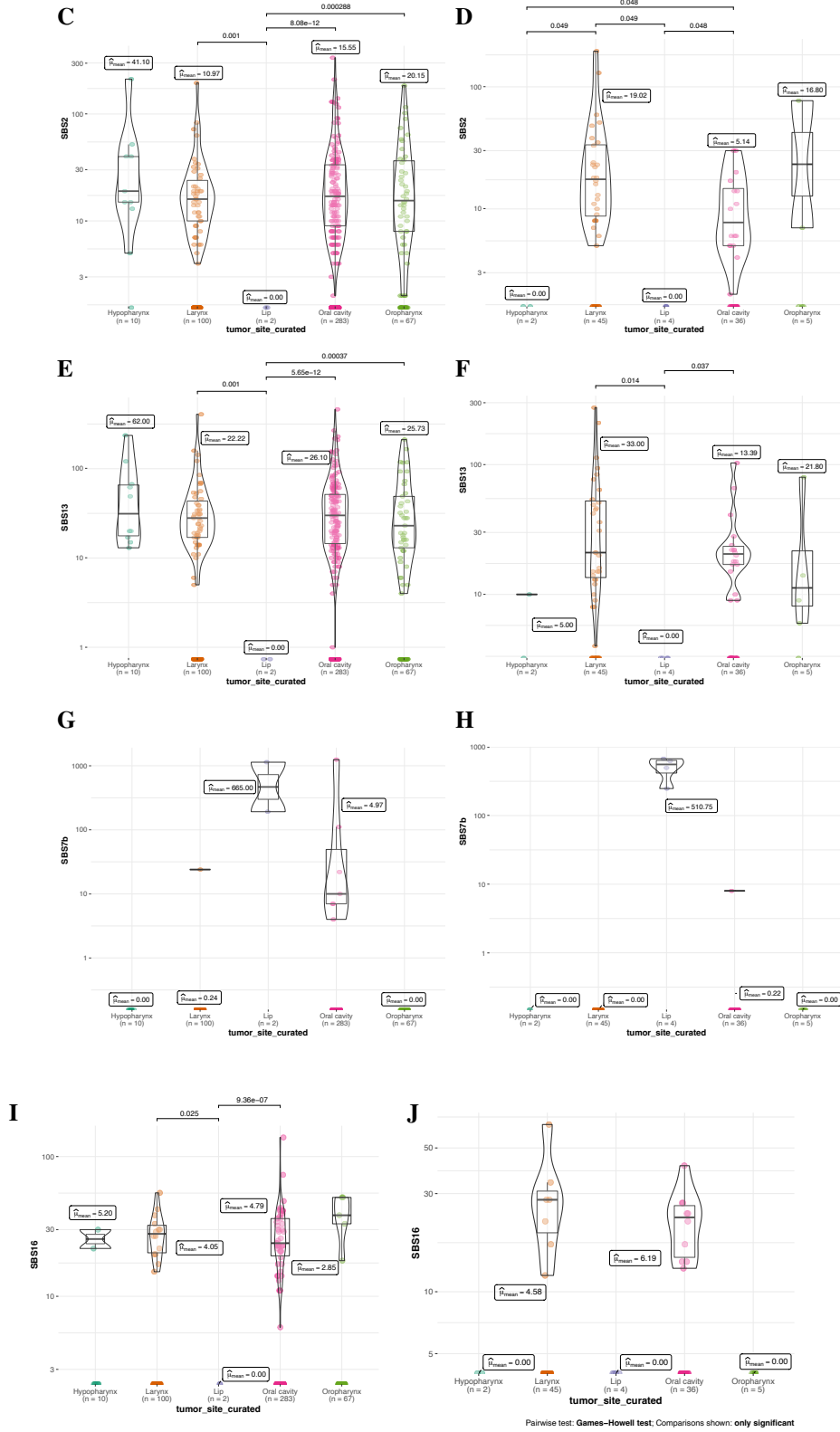


Figure 1 Landscape of mutational processes in SCCs.
A-C: CESC, ESCC and LSCC of WES data



(a) Association between other signature and tumor sites in two HNSC cohort.

C: SBS2 in TCGA HNSC; D: SBS2 in CPTAC HNSC;
 E: SBS13 in TCGA HNSC; F: SBS13 in CPTAC HNSC;
 G: SBS7b in TCGA HNSC; H: SBS7b in CPTAC HNSC;
 I: SBS16 in TCGA HNSC; J: SBS16 in CPTAC HNSC;

Figure 2 Association between signature and tumor sites in TCGA and CPTAC HNSC cohort

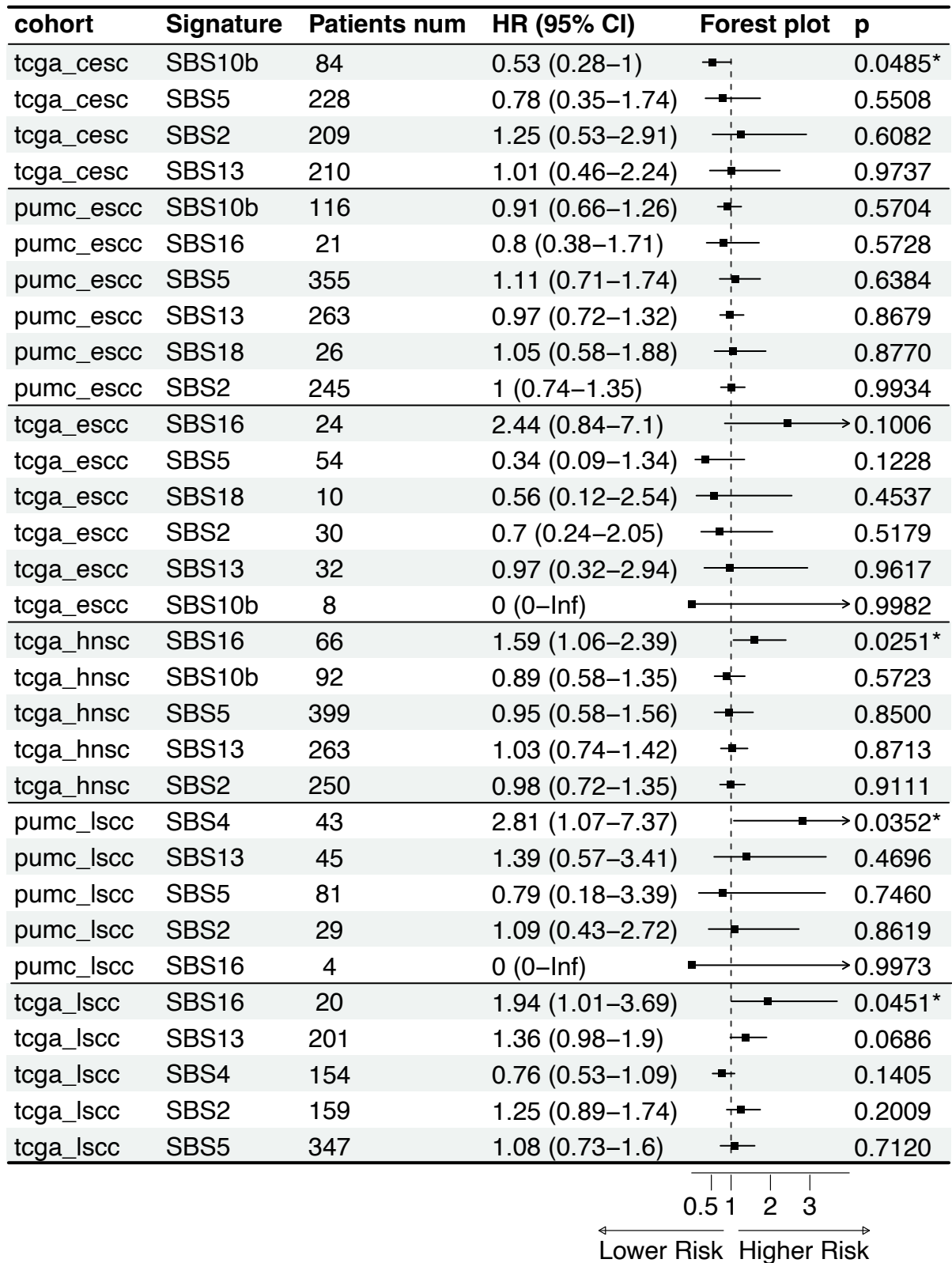
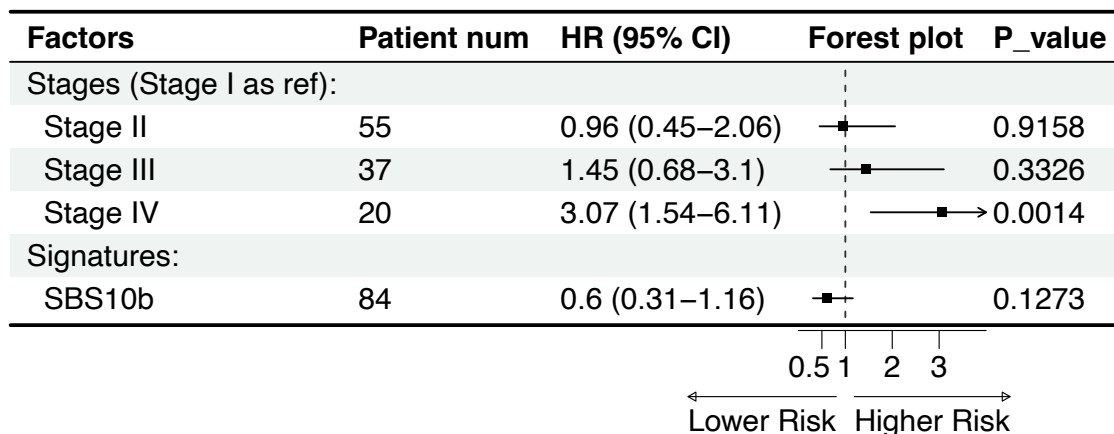
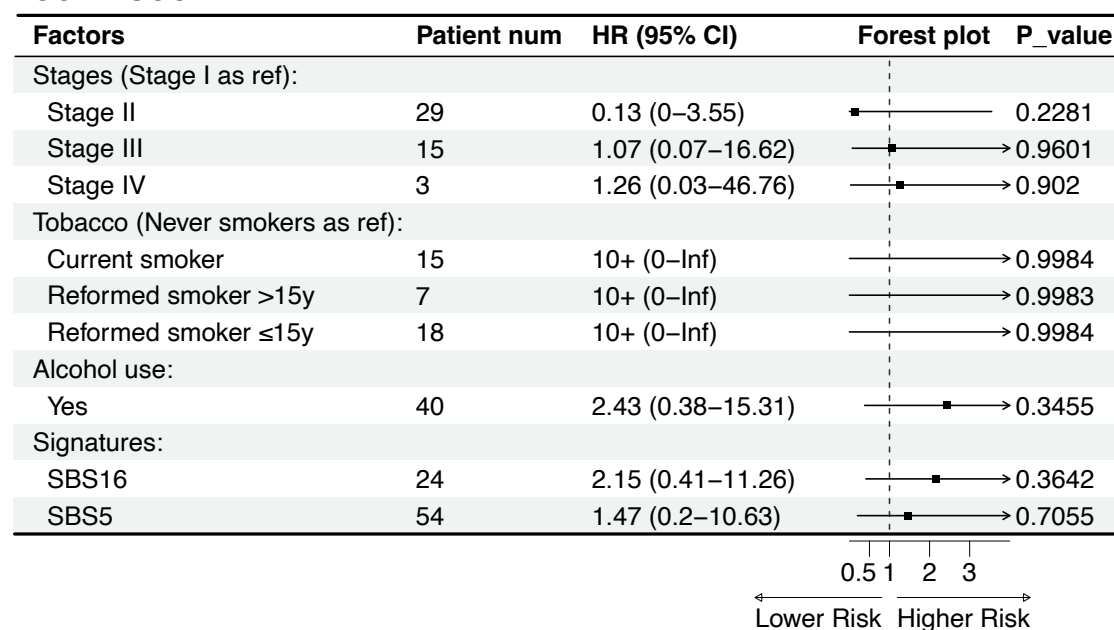


Figure 3 Forest plot of univariate Cox proportional hazards regression for representative signatures. Additionally, SBS16 was enrolled in LSCC cohorts analysis.

A

TCGA CESC

B

TCGA ESCC

C

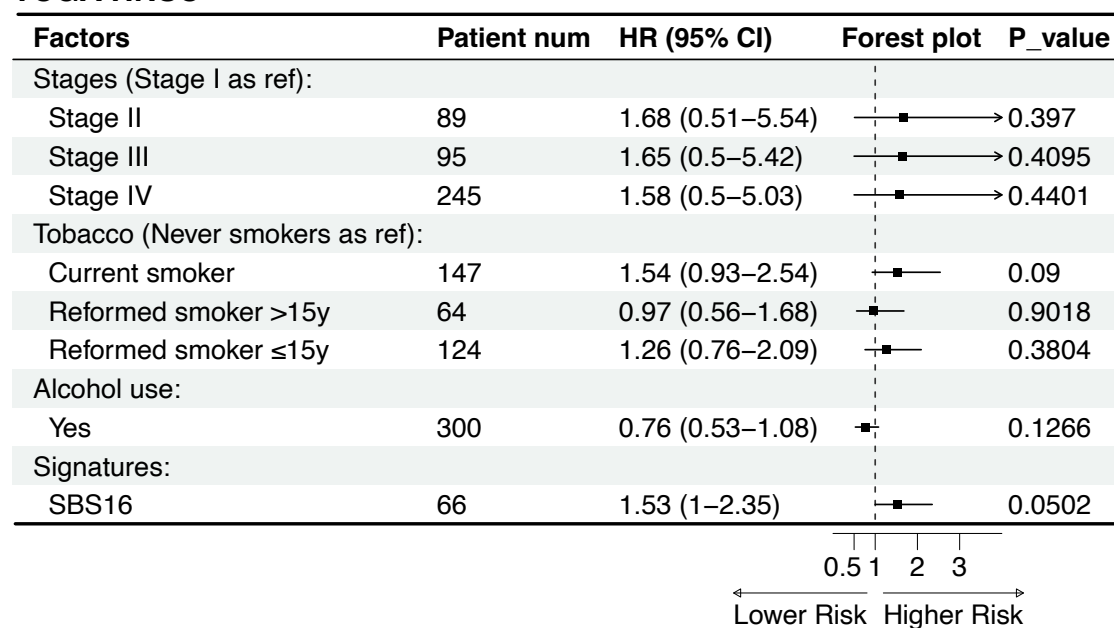
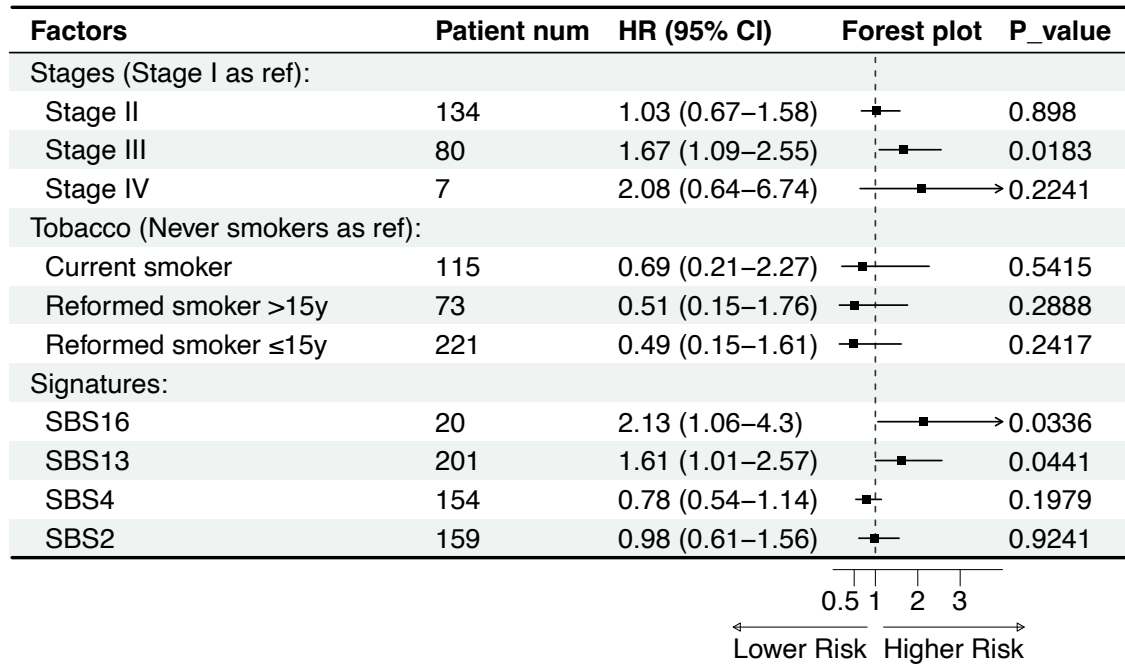
TCGA HNSC

Figure 4 Continuing in next page

D TCGA LSCC



E PUMC LSCC

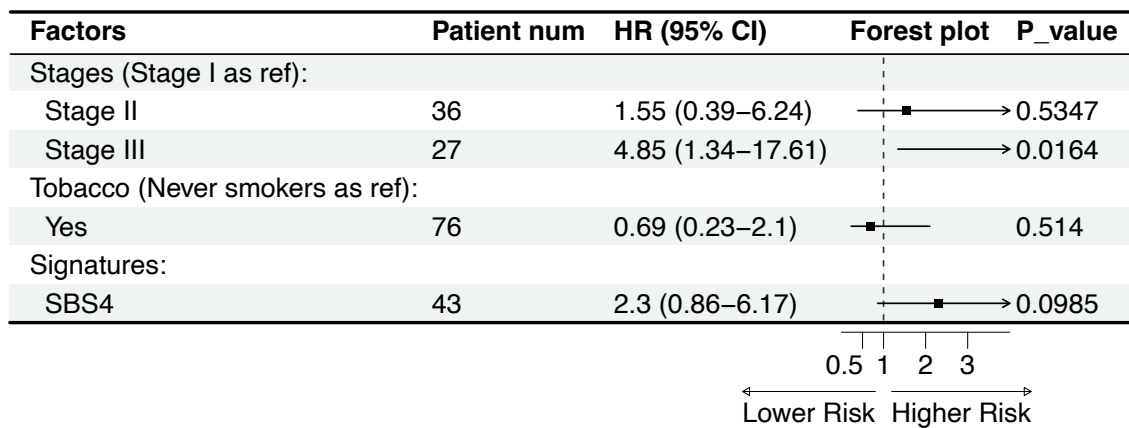


Figure 4 Forest plot of multi variate Cox proportional hazards regression for representative signatures.

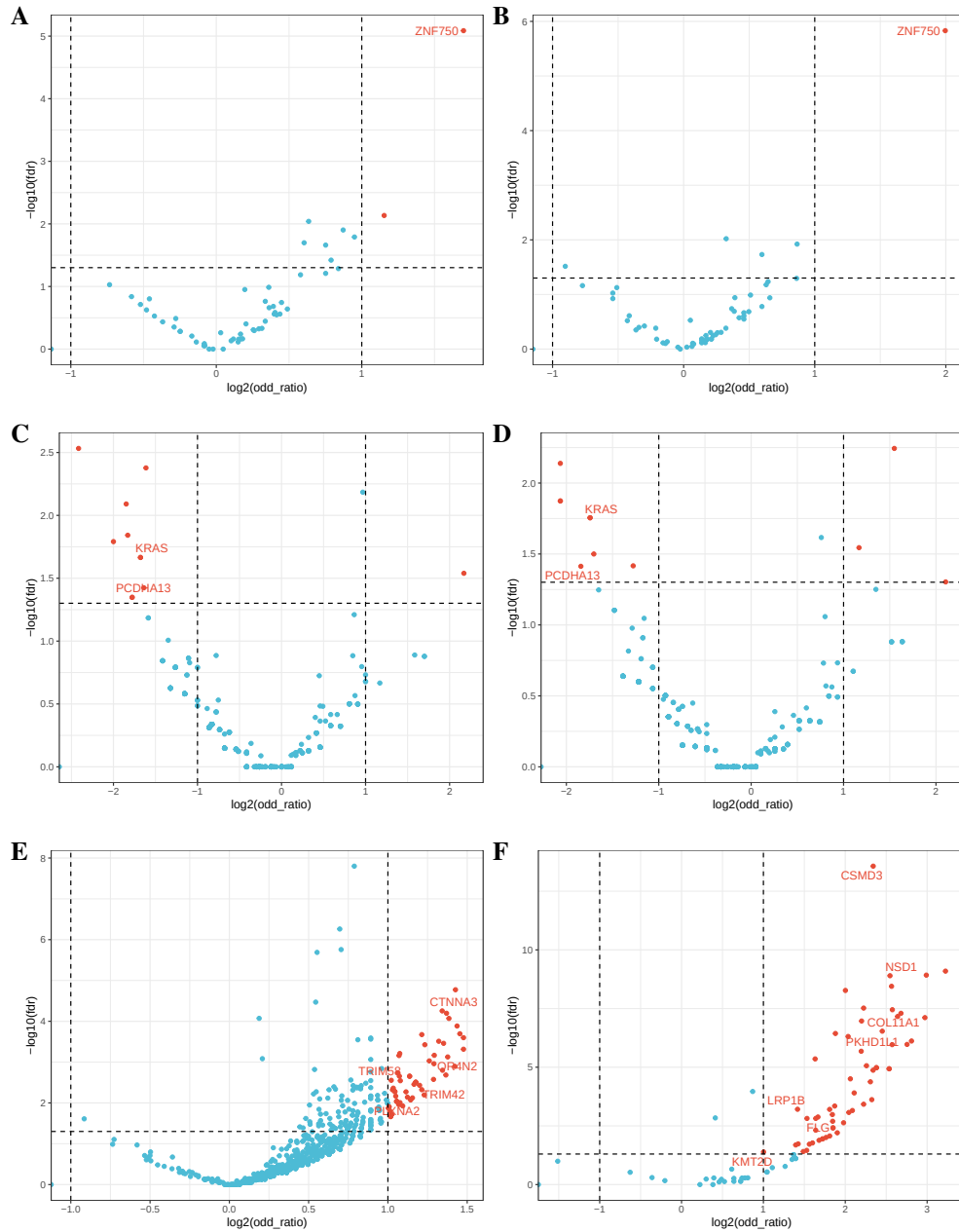


Figure 5 Mutated Genes vs Signatures: Volcano plot indicating odd ratios of mutational rates (x-axis) for each gene (with at least 5% mutation rates in each cohort, represented as dots), and significance (y-axis, negative-log scaled fdr value). Red dots are genes with more significant mutation odd-ratio (≥ 2 fold changes or ≤ 0.5 fold changes) and $\text{fdr} \leq 0.05$.

A,B: Mutated genes vs APOBEC in ESCC (A: SBS2, B: SBS13)

C,D: Mutated genes vs APOBEC in CESC (A: SBS2, B: SBS13)

E,F: Mutated genes associated with SBS4 in LSCC and HNSC (A: LSCC, B: HNSC)

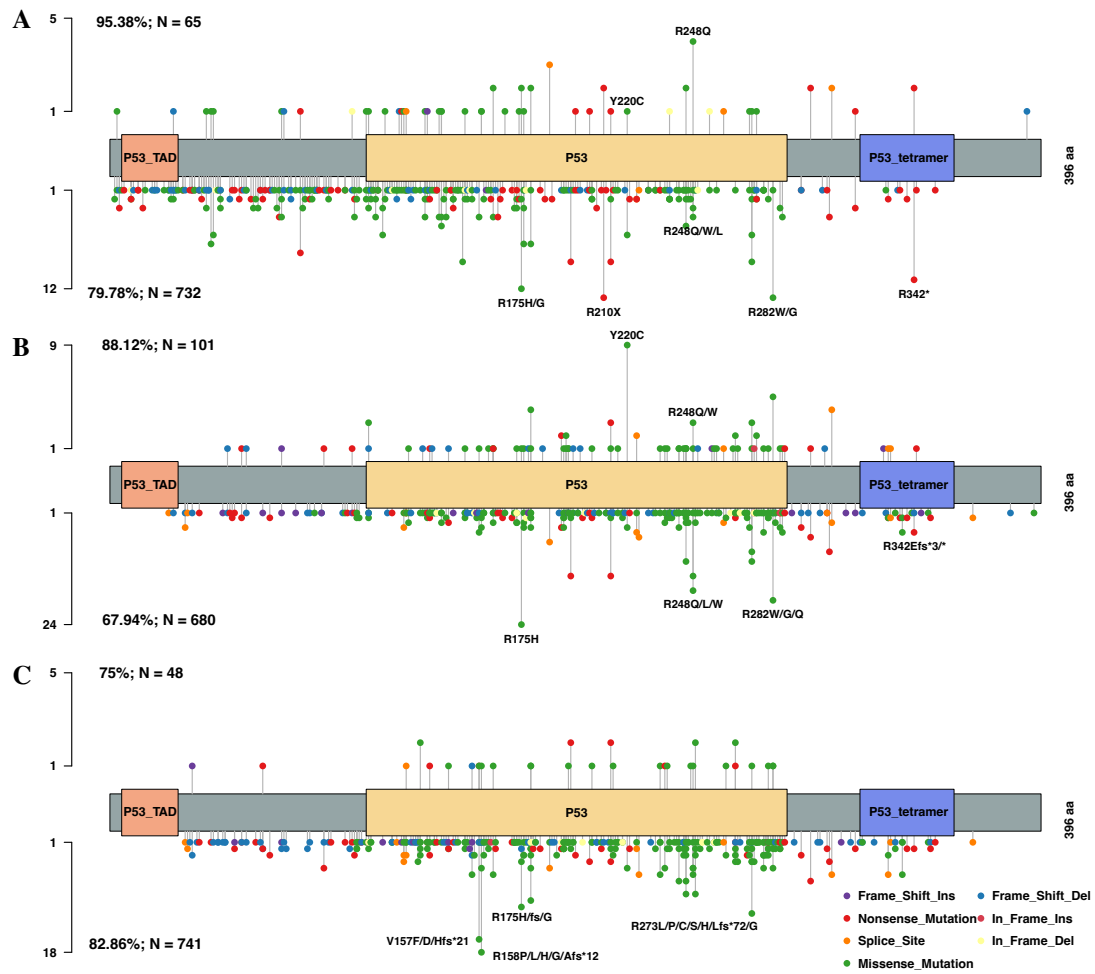


Figure 6 A,B,C: TP53 mutation sites in patients with (top) and without (bottom) SBS16 mutation signatures in ESCC (A), HNSC (B) and LSCC (C) cohorts.

Technological Notes

Methodological Framework for Parsing and Filtering Mutation Signatures: take tcga_hnsc cohort as an example

This methodological framework delineates a systematic pipeline for mutational signature analysis, encompassing data acquisition, signature decomposition, signature curation, and sample stratification. Using tcga_hnsc cohort as a paradigmatic model, we demonstrate the implementation of standardized computational approaches for characterizing somatic mutation patterns.

1. Data Acquisition and SBS96 Matrix Construction

Mutation datasets from the TCGA consortium were retrieved in RData format through the *TCGAmutations* package.

```
RData <- '~/Data/TCGA/MC3.RDs/HNSC.RDs'
laml_object <- readRDS(RData)
```

Genomic coordinates were converted to variant allele counts using hg19 human reference genome annotations. The *sigminer* R package was used to generate the SBS96 count matrix, which serves as the input for mutational signature extraction.

```
# Generate SBS96 count matrix
library(sigminer)
mt_tally <- sig_tally(
  laml_object,
  mode = 'SBS',
  ref_genome = "BSgenome.Hsapiens.UCSC.hg19",
  use_syn = FALSE # Exclude synonymous variants
)

# Export SBS96 spectrum matrix
SBS96_matrix <- t(mt_tally$all_matrices$SBS_96)
SBS96_matrix <- SBS96_matrix[order(rownames(SBS96_matrix)), ]
write.table(SBS96_matrix, "hnsc_tcga.SBS_96.txt", sep="\t", row.names=TRUE, quote=FALSE)
```

2. Mutational Signature Extraction

Multiple tools (like *maftools* or *sigminer* etc.) can calculate mutation signatures. In this study, Signatures were decomposed using *SigProfilerAssignment* pathon package, implementing non-negative matrix factorization with default hyperparameters.

```
python3 ../../run_SigProfilerAssignment.py hnsc_tcga.SBS_96.txt
```

After execution, load the key results from the output folder *SigProfilerAssignment_out*:

```
Sig_fit = list()
Sig_fit$expo = read.table("SigProfilerAssignment_out/Assignment_Solution/Activities/Assignment_Solution_Activities.txt", row.names = 1, header=T)
Sig_fit$cosine = read.table("SigProfilerAssignment_out/Assignment_Solution/Solution_Stats/Assignment_Solution_Samples_Stats.txt", row.names = 1, header=T, sep="\t")
Sig_fit$cosine = Sig_fit$cosine$Cosine.Similarity
```

Here *Sig_fit\$expo* contains the contribution (exposure) of each signature per sample, and *Sig_fit\$cosine* provides reconstruction accuracy (cosine similarity between original and reconstructed mutation profiles).

3. Cohort-Specific Signatures selection

We seek the smallest set of cohort-specific signatures that fully explain the observed mutation patterns. Therefore, the selected signatures must satisfy two criteria: (a) a sufficiently high average contribution (≥ 0.02) across

samples, and (b) minimal mutual similarity, with pairwise cosine similarity remaining below 0.7. Using following codes:

```
# load SBS signatures with sigminer package:
library(sigminer)
SigDB = get_sig_db(sig_db = "SBS")
SigCosine = sigminer::cosine(SigDB$db, SigDB$db)
rownames(SigCosine) = colnames(SigDB$db)
colnames(SigCosine) = colnames(SigDB$db)
SigCosine = as.data.frame(SigCosine)

# choose signatures with average burden >= 0.02
x = Sig_fit$expo
rownames(x) = x$sample
x = x[, -1]
x = (rowSums(x) ** -1) * x

pct = colSums(x, na.rm = T) / nrow(x)
pct = sort(pct, decreasing = T)
SigSelected = rep(1, length(pct))
names(SigSelected) = names(pct)
SigSelected[pct < 0.02] = 0

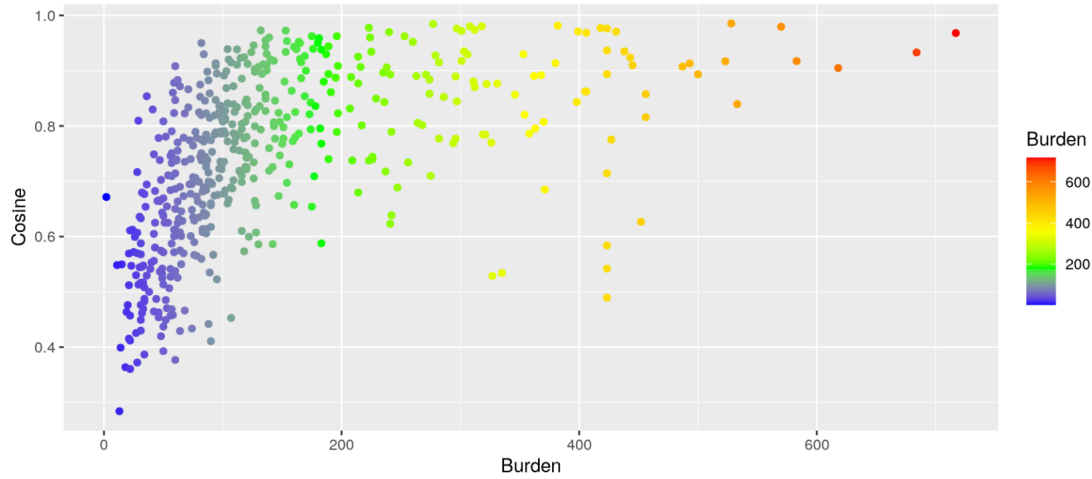
# make sure selected signatures whose cosine similarities between each other below 0.7
for(i in c(1:length(pct))){
  if(SigSelected[i] == 0){next}
  BlackList = SigCosine[names(pct)[i], ]
  BlackList = names(BlackList)[BlackList > 0.7 ]
  SigSelected[BlackList] = 0
  SigSelected[names(pct)[i]] = 1
}
```

Representative signatures of this cohort are:

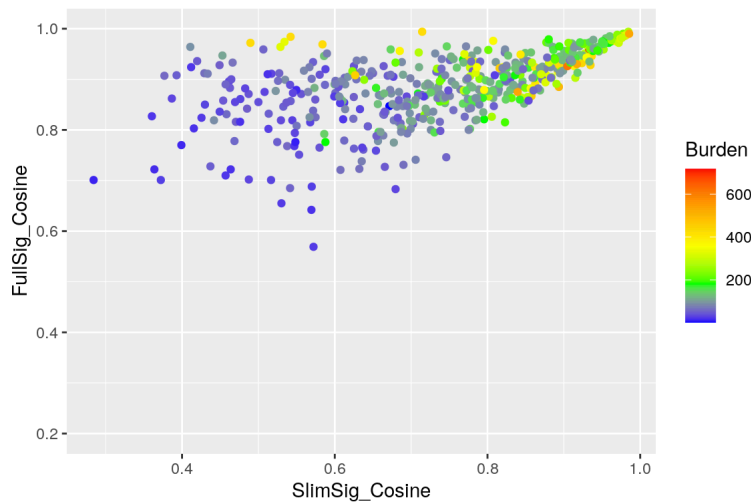
	Aetiology	Average Fraction
SBS5	Unknown (clock-like signature)	0.326977193759765
SBS13	Activity of APOBEC family of cytidine deaminases	0.150876240320061
SBS2	Activity of APOBEC family of cytidine deaminases	0.0955624505155295
SBS16	Unknown	0.0526750070254761
SBS4	Tobacco smoking	0.0318554159032029
SBS10b	Polymerase epsilon exonuclease domain mutations	0.0242918090722473

4. Sample Filtering

Based on full SBS signature extractions, we calculated the cosine similarity between the reconstructed mutation profiles and the original profiles. Higher mutation burdens correlated with better reconstruction accuracy, with 458 samples achieving cosine similarities ≥ 0.8 :



When deconvolved using the six representative signatures, reconstructed mutation profiles exhibited lower cosine similarities to the original spectra compared to those derived from the full SBS set, with only 200 samples reached a similarity threshold of ≥ 0.8 .



5. How are Filtered Samples Used in Downstream Analyses?

For clinical association analysis, samples with ≥ 0.8 cosine similarity (derived from the full SBS catalog) were retained. This ensures that clinical correlations (e.g., survival, treatment response) based on binary presence/absence of signatures are biologically reliable.

For sample clustering and tumor subtyping, we used only samples that maintained ≥ 0.8 cosine similarity when reconstructed via representative signatures. This stringent criterion ensures: (a) mutational patterns are driven by biologically dominant signatures rather than reconstruction noise, and (b) cluster boundaries reflect genuine biological divergence, not technical artifacts.