

Class 18: Mutational Signatures in Human Cancer

Patrick Tran

Exploring a cancer sequencing data portal

Q1. How many cancer samples are included in the dataset?

448 cancer samples.

Q2. Which is the most mutated gene?

TTN

Q3. Which is the most common treatment undergone by patients?

Radiation 1

Generating mutational matrices and visualizing mutational profiles

Install the `maftools` R package.

```
# Install required packages
if (!require("BiocManager")){
  install.packages("BiocManager")
}
```

Loading required package: `BiocManager`

```
if (!require("maftools")){
  BiocManager::install("maftools")
}
```

Loading required package: maftools

```
if (!require("BSgenome.Hsapiens.UCSC.hg19")){           # reference genome needed to
BiocManager::install("BSgenome.Hsapiens.UCSC.hg19")     # generate mutational matrices
}
```

Loading required package: BSgenome.Hsapiens.UCSC.hg19

Loading required package: BSgenome

Loading required package: BiocGenerics

Attaching package: 'BiocGenerics'

The following objects are masked from 'package:stats':

IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':

anyDuplicated, aperm, append, as.data.frame, basename, cbind,
colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
table, tapply, union, unique, unsplit, which.max, which.min

Loading required package: S4Vectors

Loading required package: stats4

Attaching package: 'S4Vectors'

The following objects are masked from 'package:base':

expand.grid, I, unname

Loading required package: IRanges

Attaching package: 'IRanges'

The following object is masked from 'package:grDevices':

windows

Loading required package: GenomeInfoDb

Loading required package: GenomicRanges

Loading required package: Biostrings

Loading required package: XVector

Attaching package: 'Biostrings'

The following object is masked from 'package:base':

strsplit

Loading required package: rtracklayer

Read the input data and generate mutational matrices

```
# Read maf file
library(maftools)
coad = read.maf('data_mutations.txt')
```

-Reading

-Validating

--Removed 27563 duplicated variants

-Silent variants: 209854

-Summarizing

--Possible FLAGS among top ten genes:

TTN

MUC16
-Processing clinical data
--Missing clinical data
-Finished in 37.6s elapsed (15.6s cpu)

```
# Generate mutational matrix (SBS96 context)
mm_coad = trinucleotideMatrix(maf = coad, prefix = 'chr', add = TRUE,
                              ref_genome = "BSgenome.Hsapiens.UCSC.hg19")
```

-Extracting 5' and 3' adjacent bases
-Extracting +/- 20bp around mutated bases for background C>T estimation
-Estimating APOBEC enrichment scores
--Performing one-way Fisher's test for APOBEC enrichment
---APOBEC related mutations are enriched in 1.818 % of samples (APOBEC enrichment score > 2)
-Creating mutation matrix
--matrix of dimension 440x96

```
mm_coad = t(mm_coad$nmf_matrix)
```

For the visualization of SBS96 mutational profiles, we will make use of the `MutationalPatterns` R package

```
# Install MutationalPatterns package
if (!require("MutationalPatterns")){
  BiocManager::install("MutationalPatterns")
}
```

Loading required package: `MutationalPatterns`

Loading required package: `NMF`

Loading required package: `registry`

Loading required package: `rngtools`

Loading required package: `cluster`

NMF - BioConductor layer [OK] | Shared memory capabilities [NO: windows] | Cores 11/12

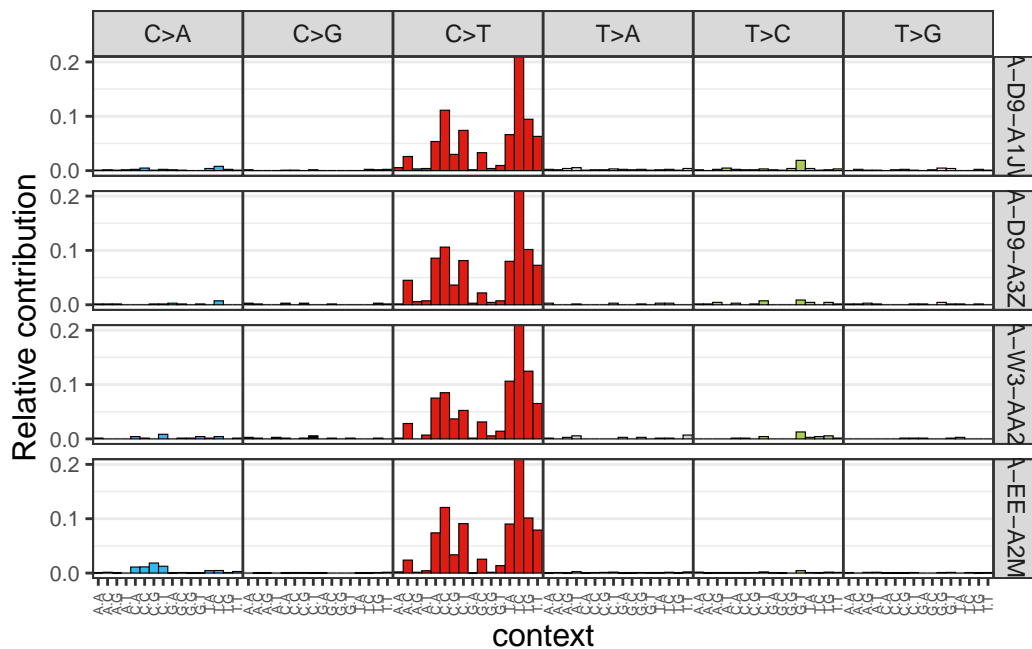
Attaching package: 'NMF'

The following object is masked from 'package:S4Vectors':

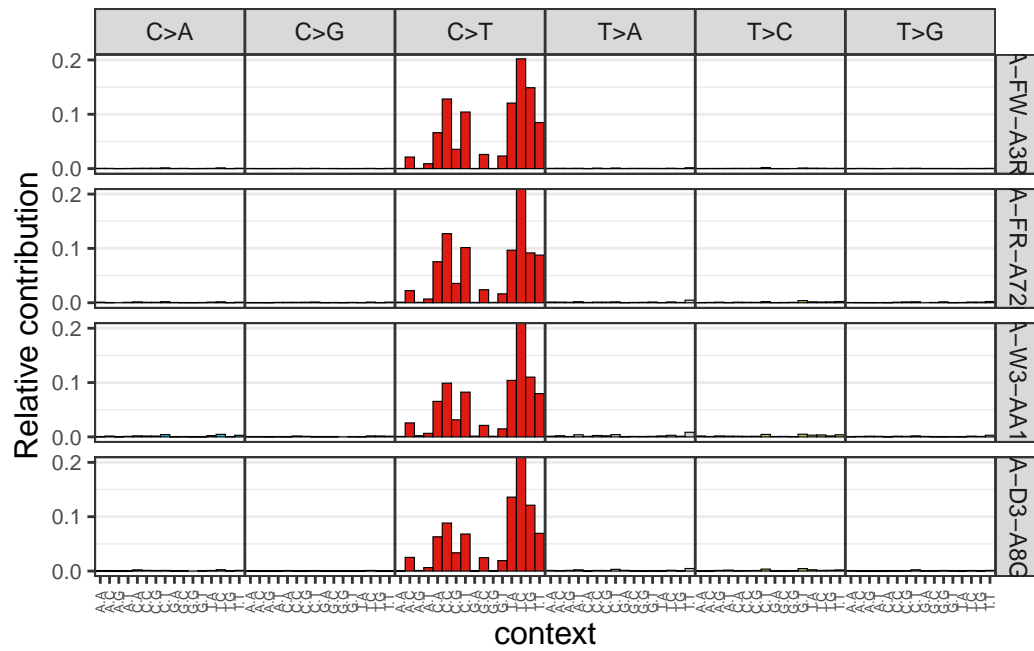
nrun

```
# Generate mutational profiles (4 random samples)
library(MutationalPatterns)
set.seed(11111) # fixing the seed for random number generation

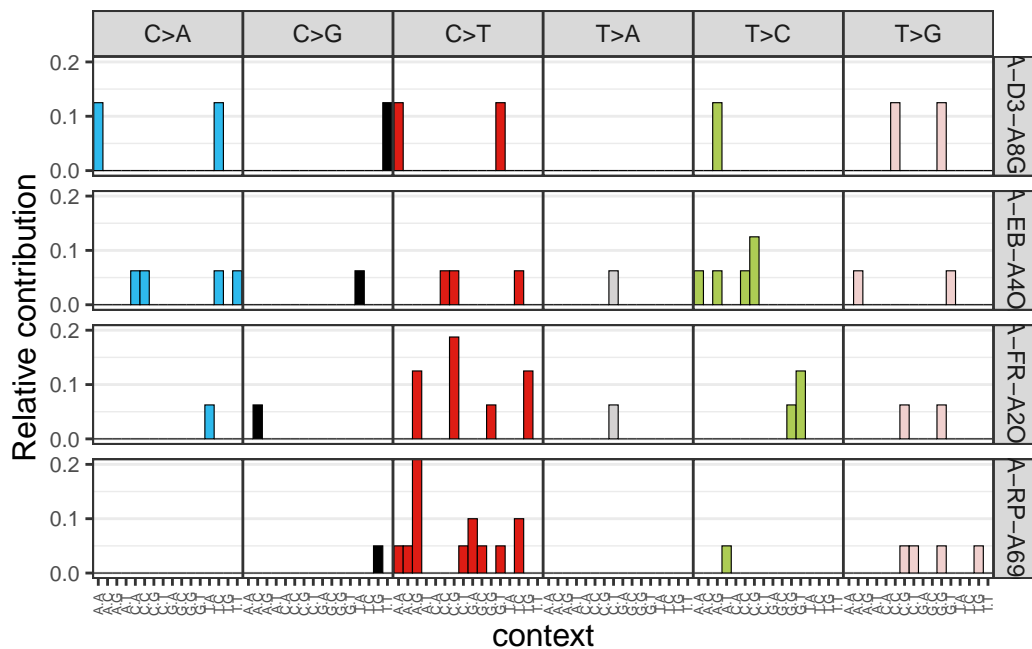
samples_to_plot = sample(1:ncol(mm_coad),4) # selecting 4 random samples
plot_96_profile(mm_coad[,samples_to_plot], condensed = T)
```



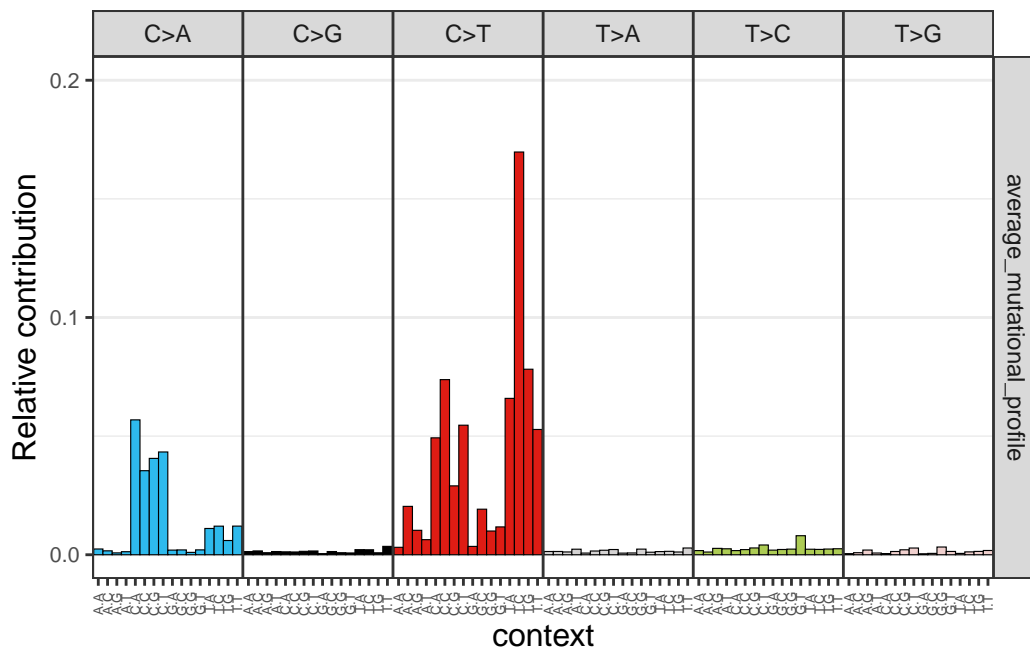
```
# Generate mutational profiles (top 4 mutated samples and top 4 less mutated)
mutations_in_samples = colSums(mm_coad)
mutations_in_samples = sort(mutations_in_samples, decreasing = T)
samples_to_plot = names(mutations_in_samples)[1:4]
plot_96_profile(mm_coad[,samples_to_plot], condensed = T)
```



```
mutations_in_samples = sort(mutations_in_samples, decreasing = F)
samples_to_plot = names(mutations_in_samples)[1:4]
plot_96_profile(mm_coad[,samples_to_plot], condensed = T)
```



```
# Generate average mutational profiles
relative_mutational_profile = apply(mm_coad, 2, prop.table) # obtained relative
                                                             # mutational matrix
average_mutational_profile = rowMeans(relative_mutational_profile)
average_mutational_profile = data.frame(average_mutational_profile)
plot_96_profile(average_mutational_profile, condensed = T)
```



Assigning reference mutational signatures

```
# Mutational signature assignment
cosmic_signatures = get_known_signatures(source = 'COSMIC_v3.2')
fit_res = fit_to_signatures(mm_coad, cosmic_signatures)

# Top contributing signatures
contributions = fit_res$contribution

top_contributing_signatures_abs = rowMeans(contributions)
top_contributing_signatures_abs = sort(top_contributing_signatures_abs,
                                       decreasing = T)[1:4]

## Top 4 contributing signatures (absolute values)
top_contributing_signatures_abs
```

SBS7a	SBS7b	SBS38	SBS4
366.97614	340.91011	204.44450	99.49106


```

relative_contributions = apply(contributions,2,prop.table)
top_contributing_signatures_rel = rowMeans(relative_contributions)
top_contributing_signatures_rel = sort(top_contributing_signatures_rel,
                                       decreasing = T)[1:4]

## Top 4 contributing signatures (relative values)
top_contributing_signatures_rel

```

```

      SBS7b      SBS7a      SBS38      SBS4
0.26336351 0.26019455 0.10885595 0.07240978

```

```

# Mutational signature assignment strict
fit_res_strict = fit_to_signatures_strict(mm_coad, cosmic_signatures)
fit_res_strict = fit_res_strict$fit_res
contributions_strict = fit_res_strict$contribution

```

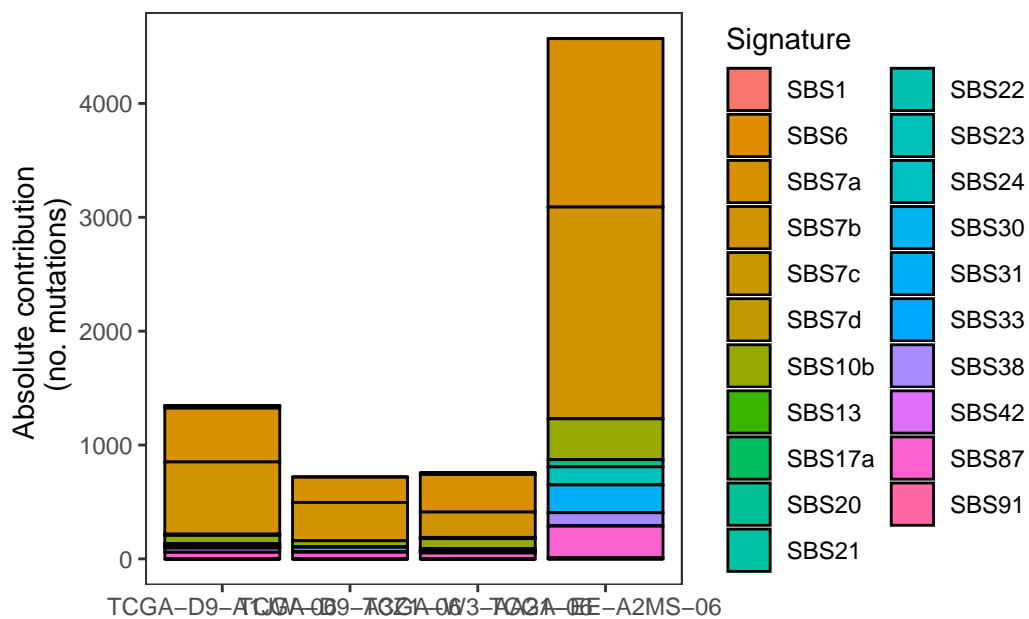
Visualizing mutational signature assignment results

```

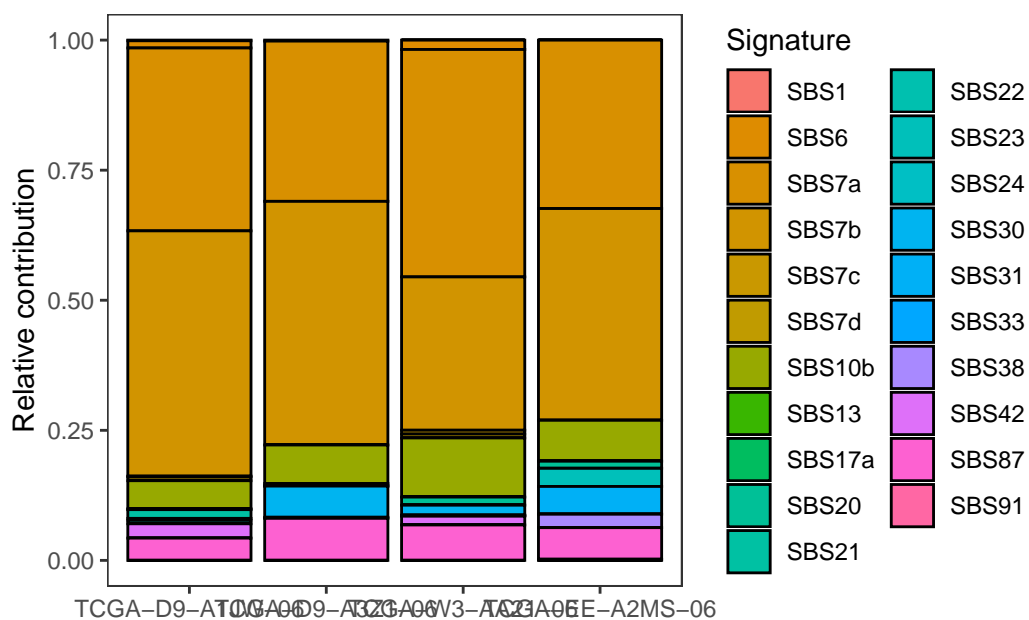
# Visualization of signature assignment results (fit_to_signatures)
set.seed(11111)
samples_to_plot = sample(1:ncol(mm_coad),4)

plot_contribution(contributions[,samples_to_plot], mode = "absolute")

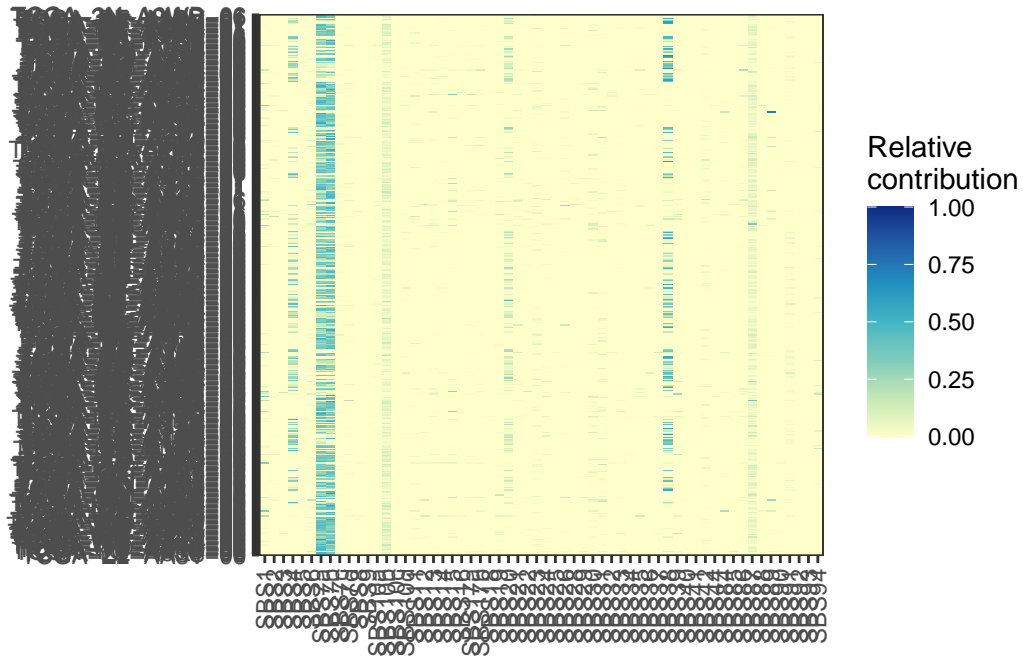
```



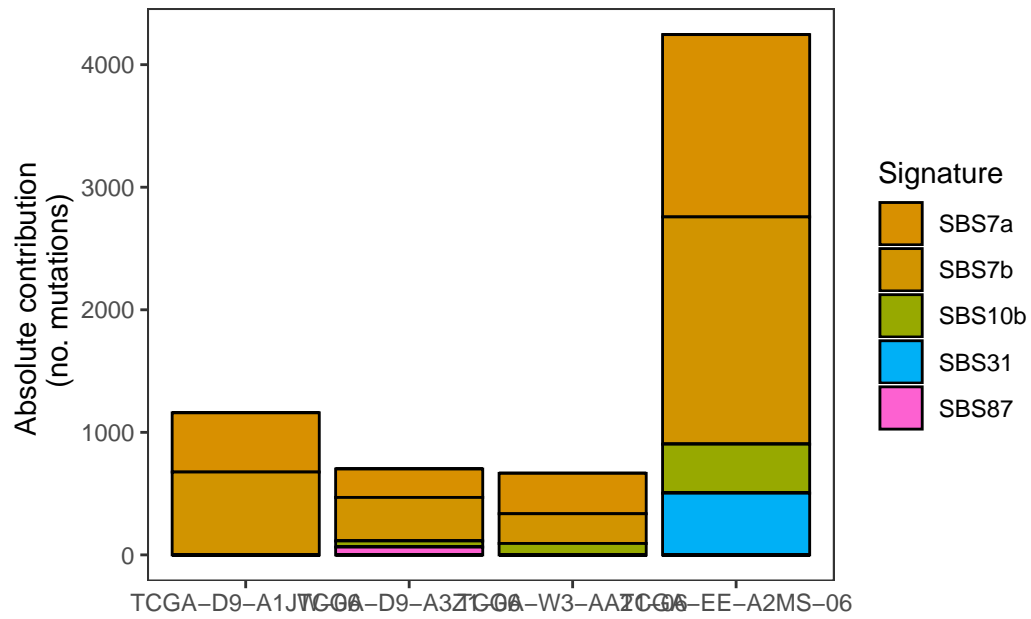
```
plot_contribution(contributions[,samples_to_plot], mode = "relative")
```



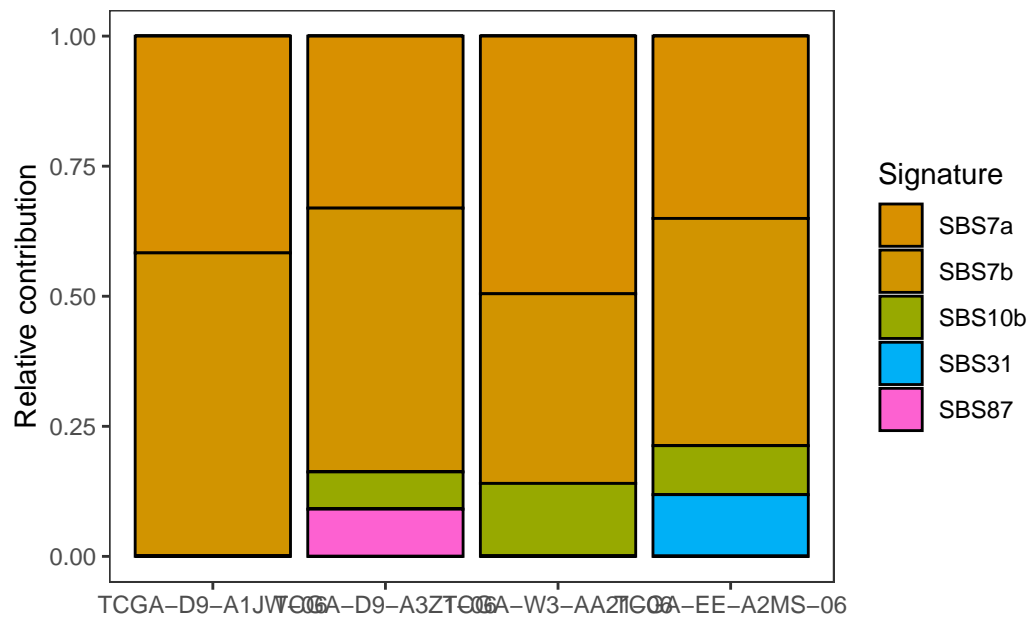
```
plot_contribution_heatmap(contributions, cluster_samples = F)
```



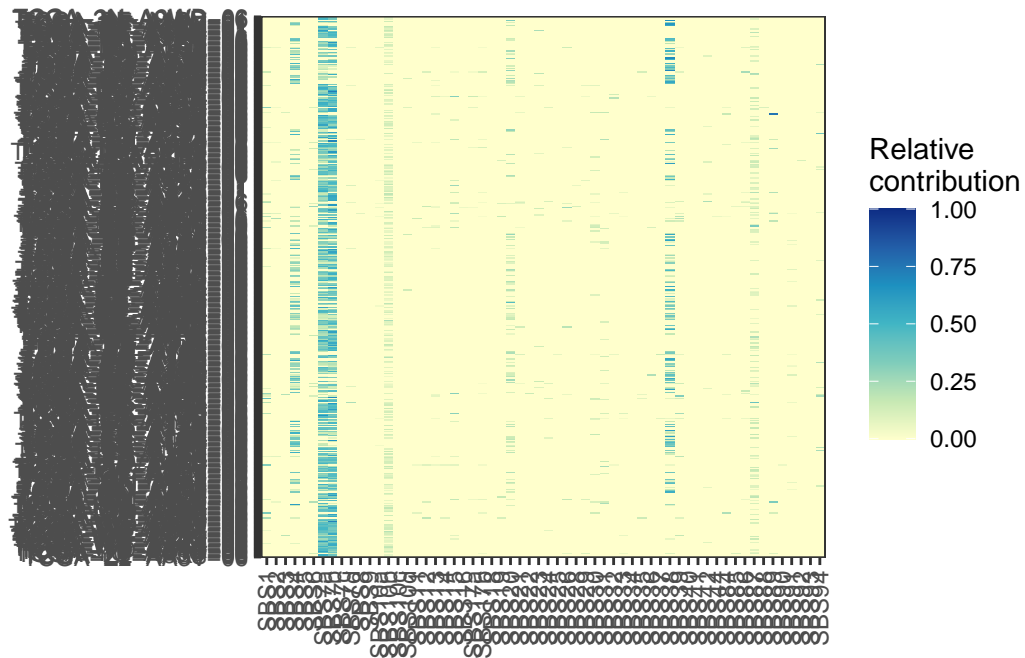
```
# Visualization of signature assignment results (strict)
plot_contribution(contributions_strict[,samples_to_plot], mode = "absolute")
```



```
plot_contribution(contributions_strict[,samples_to_plot], mode = "relative")
```

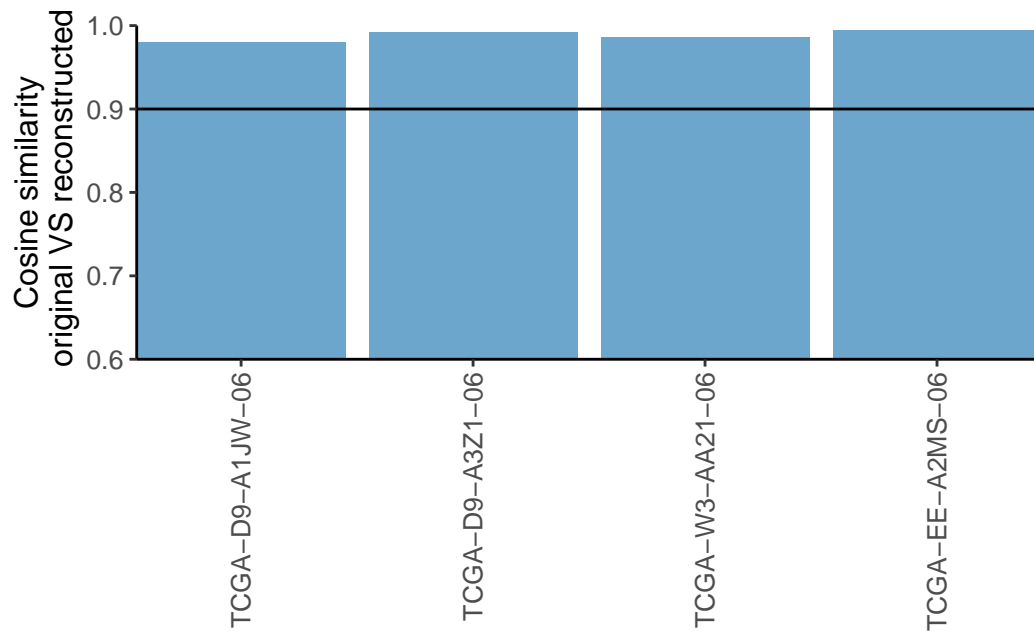


```
plot_contribution_heatmap(contributions_strict, cluster_samples = F)
```

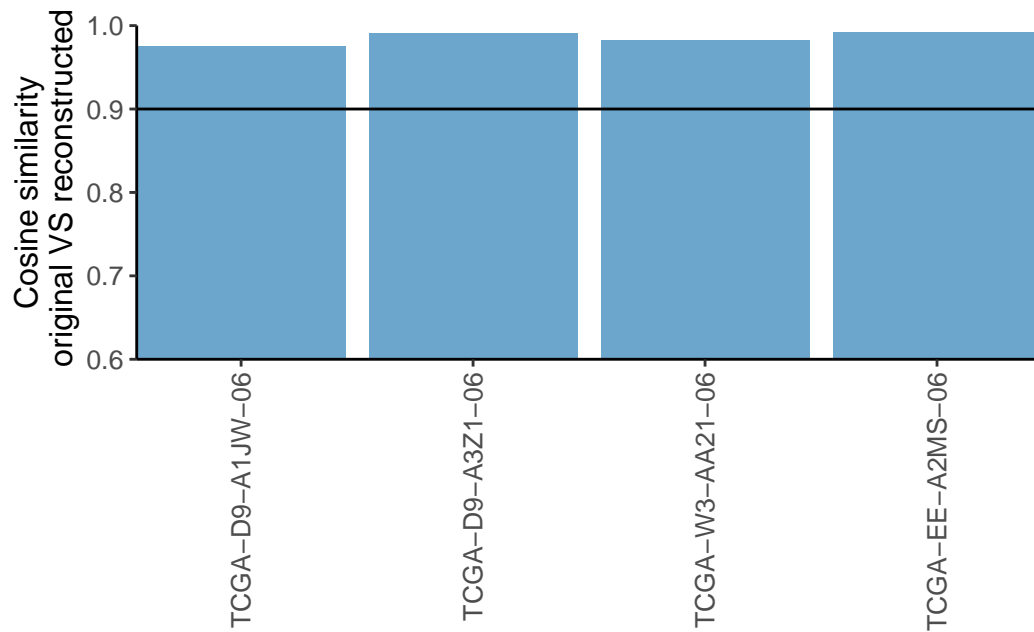


```
# Cosine similarity reconstruction vs. original mutational profile (fit_to_signatures)
set.seed(11111)
samples_to_plot = sample(1:ncol(mm_coad), 4)

plot_original_vs_reconstructed(mm_coad[, samples_to_plot],
                               fit_res$reconstructed[, samples_to_plot],
                               y_intercept = 0.90)
```



```
# Cosine similarity reconstruction vs. original mutational profile (strict)
plot_original_vs_reconstructed(mm_coad[,samples_to_plot],
                               fit_res_strict$reconstructed[,samples_to_plot],
                               y_intercept = 0.90)
```



Q4. Which is the etiology of the top absolute contributing signature for liver cancer?

Aristolochic acid exposure

Q5. Which is the most prominent mutational context for the top contributing signature in skin cancer?

C>T

Q6. The etiology of the top contributing signature for lung cancer corresponds to an endogenous cellular mechanism.

FALSE

Q7. SBS4 is one of the most common signatures found in lung cancer and is associated with tobacco smoking.

TRUE

Q8. SBS7d is one of the most common signatures in skin cancer and is associated with UV light exposure and high numbers of C>T mutations.

FALSE