class18

Yinuo Song

```
##BiocManager::install("maftools")
```

```
##BiocManager::install("BSgenome.Hsapiens.UCSC.hg19")
```

```
##BiocManager::install("MutationalPatterns")
```

1. Exploring a cancer sequencing data portal

Group 1 **liver hepatocellular carcinoma**

Q1 How many cancer samples are included in the dataset?

372

Q2. Which is the most mutated gene?

TNN

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

Sorafenib

2. Downloading cancer sequencing data

3. Generating mutational matrices and visualizing mutational profiles

```
# Read maf file
library(maftools)
lihc = read.maf('lihc_tcga_pan_can_atlas_2018/data_mutations.txt')
```

```
-Reading
-Validating
--Removed 3332 duplicated variants
-Silent variants: 21154
-Summarizing
--Mutiple centers found
.;--Possible FLAGS among top ten genes:
  TTN
  MUC16
  OBSCN
  FLG
-Processing clinical data
--Missing clinical data
-Finished in 6.046s elapsed (5.113s cpu)
```

```
# Generate mutational matrix (SBS96 context)
mm_lihc = trinucleotideMatrix(maf = lihc, prefix = 'chr', add = TRUE, ref_genome = "BSgeno
```

```
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


Attaching package: 'S4Vectors'

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

    expand.grid, I, unname


Attaching package: 'Biostrings'

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

    strsplit

-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  1.117 % of samples (APOBEC enrichment score > 2
-Creating mutation matrix
--matrix of dimension 358x96
```

```r
mm_lihc = t(mm_lihc$nmf_matrix)
```

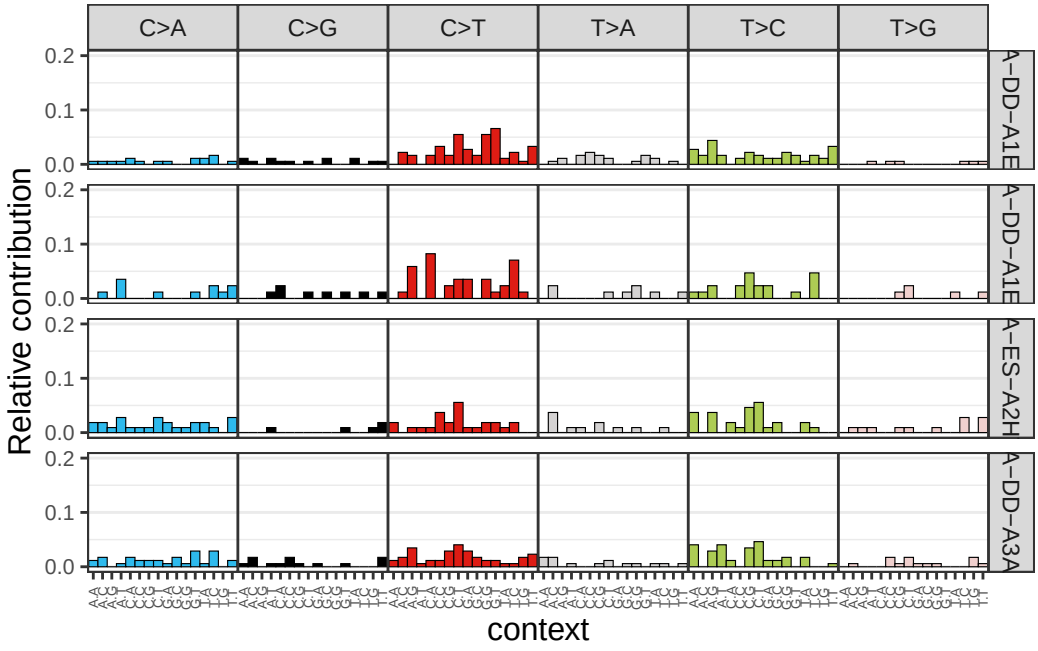
For the **visualization of SBS96 mutational profiles**, we will make use of the `MutationalPatterns` R package. This library is commonly used for all kinds of mutational signature analysis, and we will also use it for the subsequent assignment analysis.

```r
# Generate mutational profiles (4 random samples)
library(MutationalPatterns)
```

```
Loading required package: NMF

Loading required package: registry
```

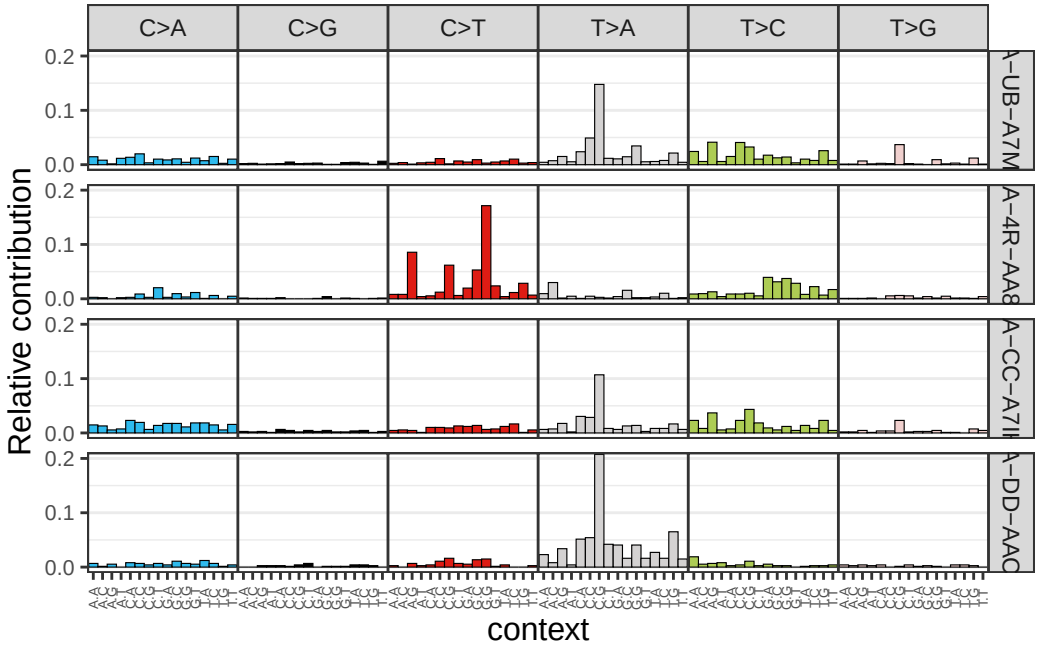
```
Loading required package: rngtools

Loading required package: cluster

NMF - BioConductor layer [OK] | Shared memory capabilities [NO: bigmemory] | Cores 2/2

  To enable shared memory capabilities, try: install.extras('
NMF
')


Attaching package: 'NMF'

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

    nrun
```

```r
set.seed(11111) # fixing the seed for random number generation

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

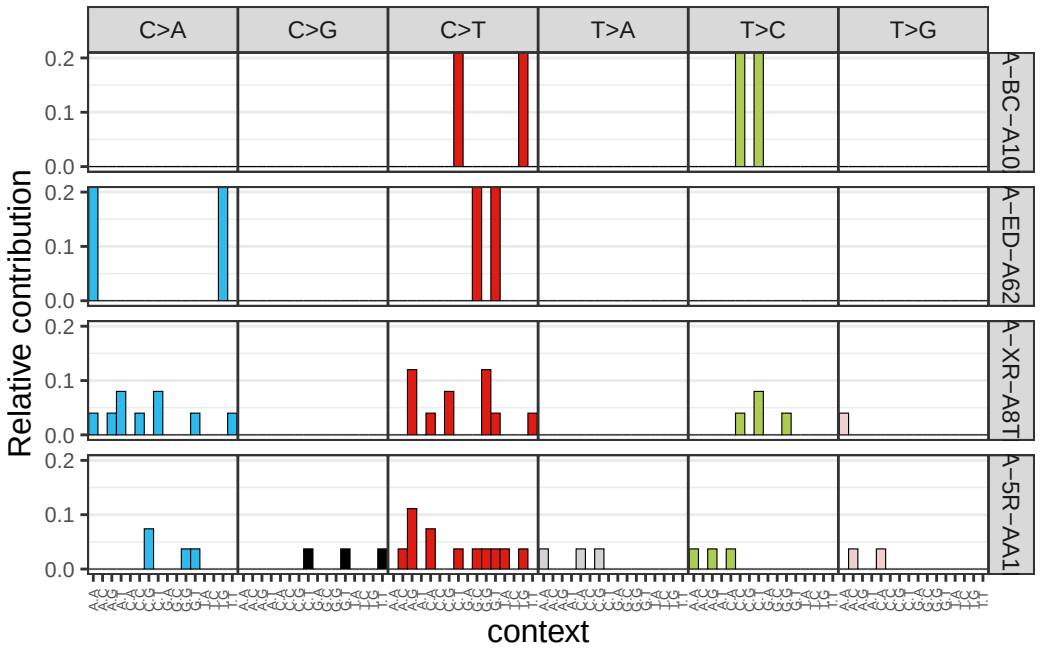


Plot the samples with more mutations

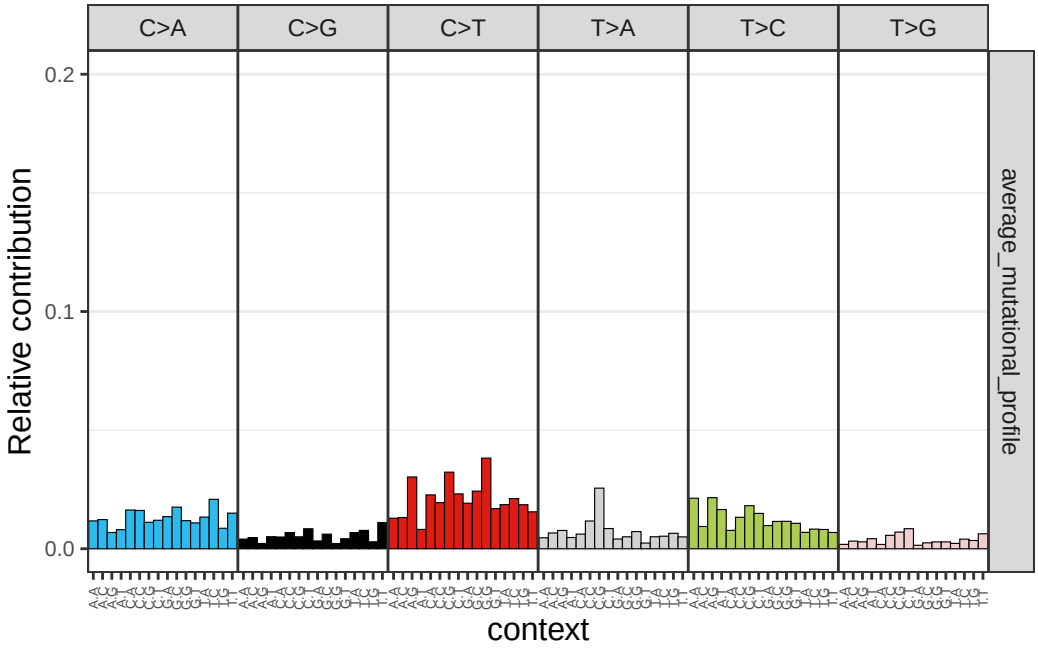
```
# Generate mutational profiles (top 4 mutated samples and top 4 less mutated)
mutations_in_samples = colSums(mm_lihc)
top_4_mutated_cases= order(mutations_in_samples, decreasing = T)[1:4]
plot_96_profile(mm_lihc[,top_4_mutated_cases], condensed = T)
```




```
top_4_less_mutated_cases <- order(mutations_in_samples, decreasing = F)[1:4]
plot_96_profile(mm_lihc[,top_4_less_mutated_cases], condensed = T)
```




```
# Generate average mutational profiles
relative_mutational_profile = apply(mm_lihc, 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)
```



4. COSMIC reference mutational signatures

5. Assigning reference mutational signatures

Signature Assignment

```
# Mutational signature assignment
cosmic_signatures = get_known_signatures(source = 'COSMIC_v3.2')
fit_res = fit_to_signatures(mm_lihc, cosmic_signatures)
# Top contributing signatures
contributions = fit_res$contribution
```

Top 4 contributing signatures in liver hepatocellular carcinoma

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

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

```
    SBS22     SBS24     SBS26      SBS4
15.174219 13.283108  9.783660  8.506521
```

To get relative values for mutations in signatures

```
relative_contributions = apply(contributions,2,prop.table)
top_relative = sort(rowMeans(relative_contributions),decreasing = T)


## Top 4 contributing signatures (relative values)
top_contributing_signatures_rel = top_relative[1:4]
top_contributing_signatures_rel
```

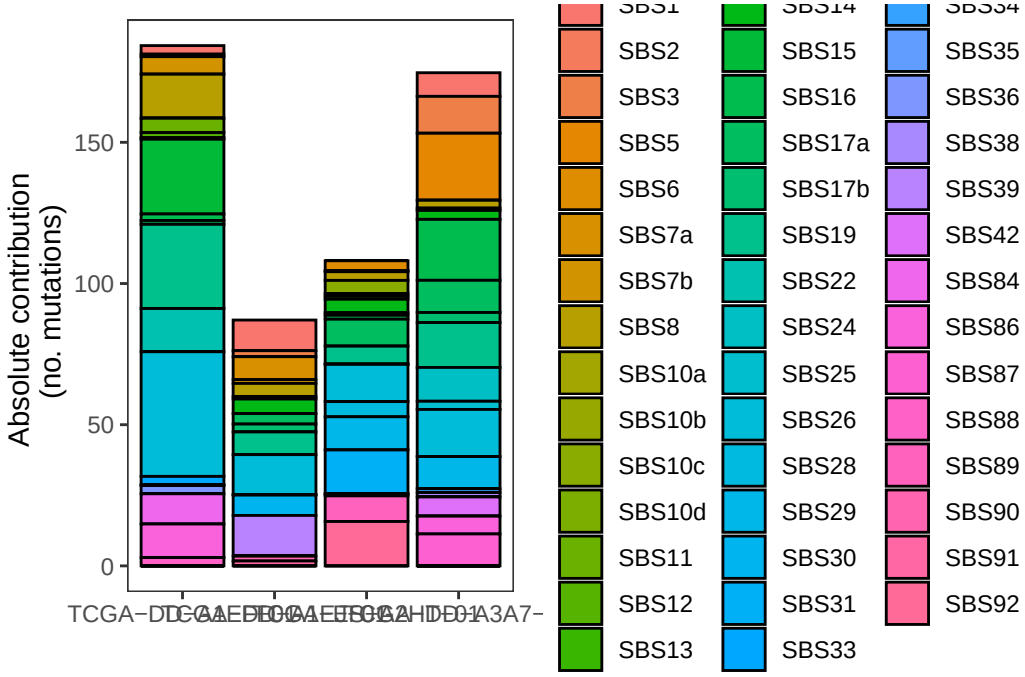
```
    SBS24      SBS22      SBS26      SBS87
0.08643508 0.05763556 0.05750834 0.04976188
```

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

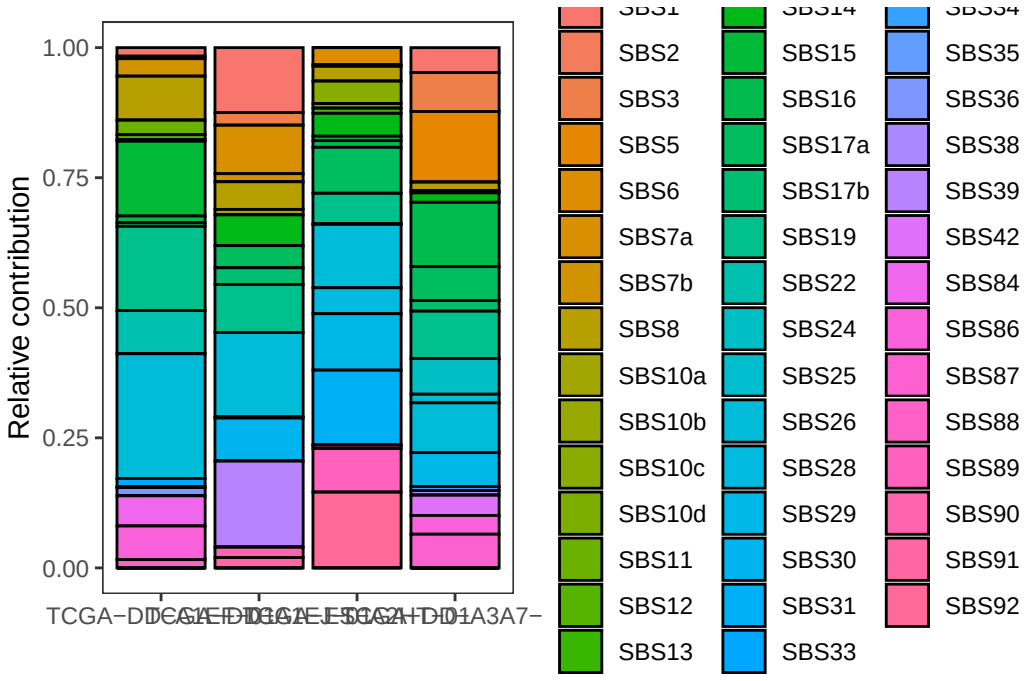
6. Visualizing mutational signature assignment results

To visualize the mutational signature assignment results, we will use the default visualizations available in the `MutationalPatterns` package. However, other visualizations are also present as part of `maftools` (please check the appropriate section in their vignette) or can be created using `ggplot2` and the contributions output matrix from the mutational signature assignment analysis (`contributions` or `contributions_strict`).

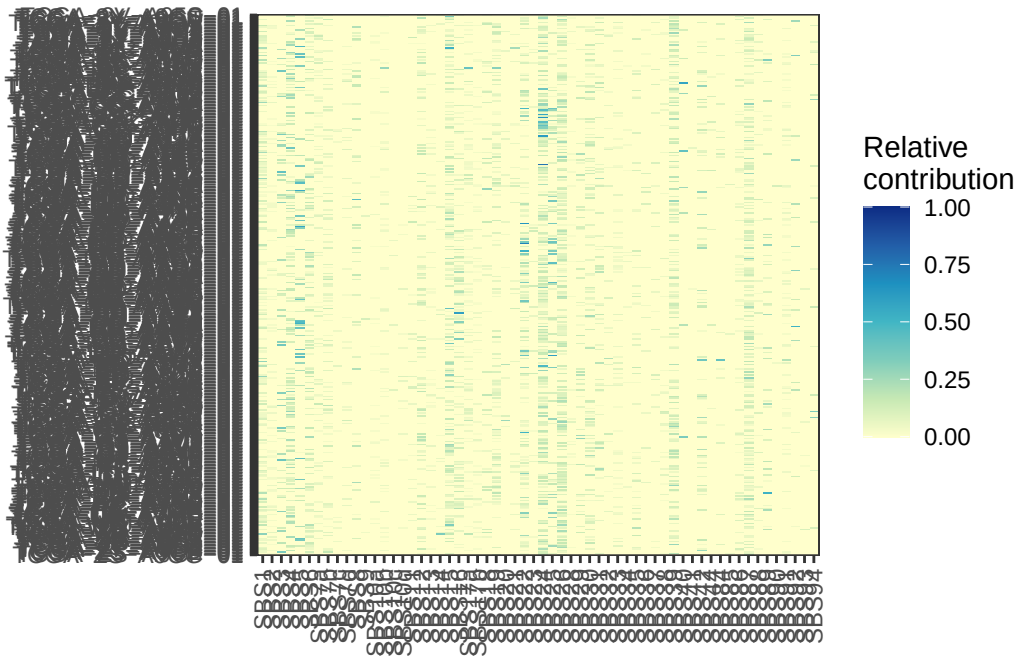
```
# Visualization of signature assignment results (fit_to_signatures)
set.seed(11111)
samples_to_plot = sample(1:ncol(mm_lihc),4)

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



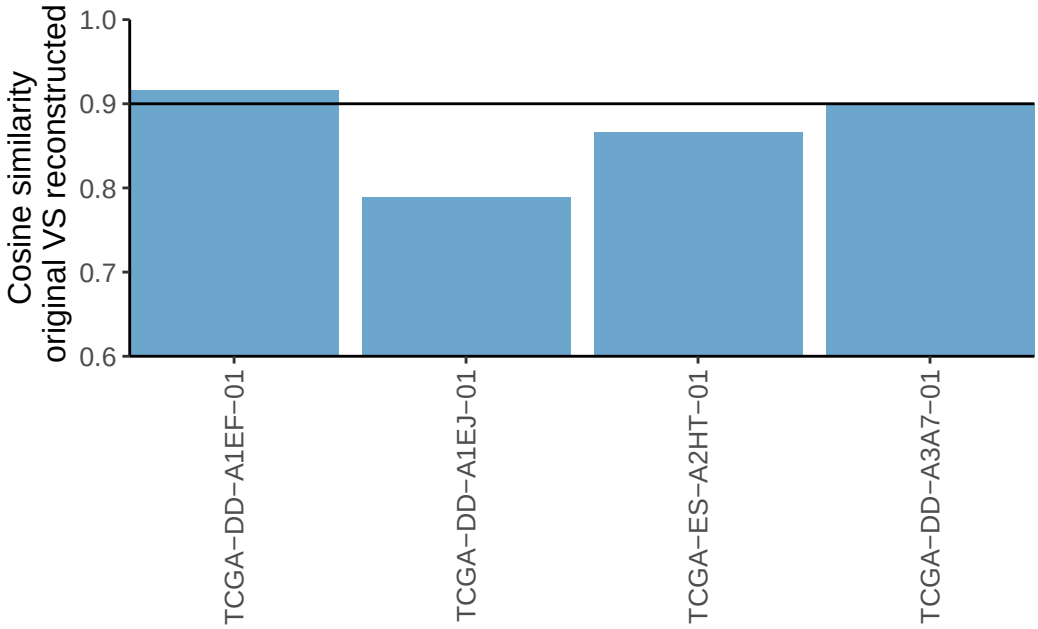

```
plot_contribution(contributions[,samples_to_plot], mode = "relative")
```



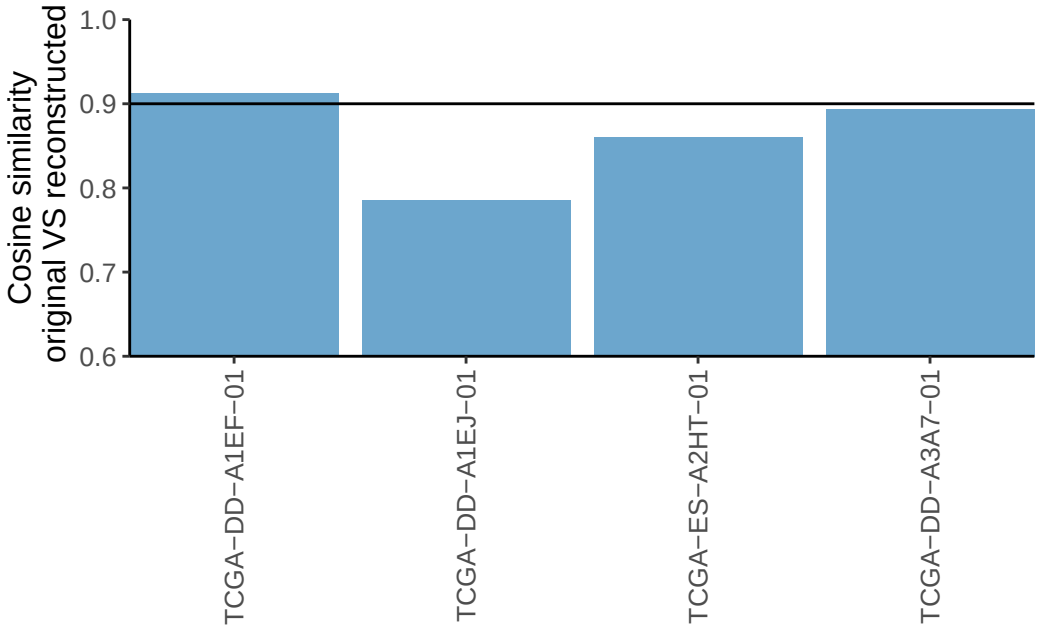
```
plot_contribution_heatmap(contributions_strict, cluster_samples = F)
```




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

plot_original_vs_reconstructed(mm_lihc[,samples_to_plot],fit_res$reconstructed[,samples_to
```




```
# Cosine similarity reconstruction vs. original mutational profile (strict)
plot_original_vs_reconstructed(mm_lihc[,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.