bayesNMF: Signature Inference and Visualization

Jenna Landy

```
library(bayesNMF)

data <- readRDS("examples/3_64_1_cosmic.rds")
refit = FALSE
```

```
if (refit) {
    res <- bayesNMF(
        data$M, 1:20,
        file = "examples/plot_example",
        overwrite = TRUE
    )
} else {
    res <- readRDS("examples/plot_example.rds")
}
```

Inference

The `signature_assignment_inference` function performs the Hungarian algorithm on each P matrix among the posterior samples. Each sample gets to vote for it's assignment with cosine similarity as voting weights.

Input

- `res`: bayesNMF results object.
- `ref_matrix`: reference signatures to align to in the form of a reference P matrix or the string `"cosmic"` (default).

Output

- `assignment`: data.frame specifying which estimated signatures ($n = 1, 2, ...$) are assigned to which reference signatures (`sig` column) and with what score. Score is the sum of voting weights assigned to this reference signature over the total voting weights across reference signatures. The maximum score of 1 indicates that the estimated signature was assigned to that particular reference signature in all posterior samples.
- `MAP$cos_sim`: average cosine similarities between estimated signatures and assigned reference signatures across posterior samples
- `credible_intervals$cos_sim`: 95% credible intervals of cosine similarities between estimated signatures and assigned reference signatures across posterior samples

```
assign <- signature_assignment_inference(res)
names(assign)
```

```
## [1] "assignment"         "MAP"                  "credible_intervals"
```

```
assign$assignment
```

```
##     sig score n
## 1 SBS88     1 1
## 2 SBS53     1 2
## 3 SBS49     1 3
```

```
assign$MAP$cos_sim
```

```
##     SBS88     SBS53     SBS49
## 0.9990684 0.9919638 0.9993091
```

```
assign$credible_intervals$cos_sim
```

```
## [[1]]
##     SBS88     SBS53     SBS49
## 0.9987580 0.9899523 0.9990338
##
## [[2]]
##     SBS88     SBS53     SBS49
## 0.9993219 0.9939621 0.9995534
```

Plots

All functions return `ggplot2` objects, meaning it is simple to add additional layers to change the theme, axes labels, etc. It also means plots can be combined with packages like `patchwork`, `cowplot`, `gridExtra`, and `ggpubr`. The next few sections go over each plotting function separately.

Plotting Individual Signatures

The `plot_sig` function can be used to plot a reference mutational signature, an estimated mutational signature with posterior uncertainty as points with error bars, or both. The default reference is COSMIC, and unless otherwise specified, the reference signature chosen to plot is that with highest cosine similarity to the estimated signature.

Input

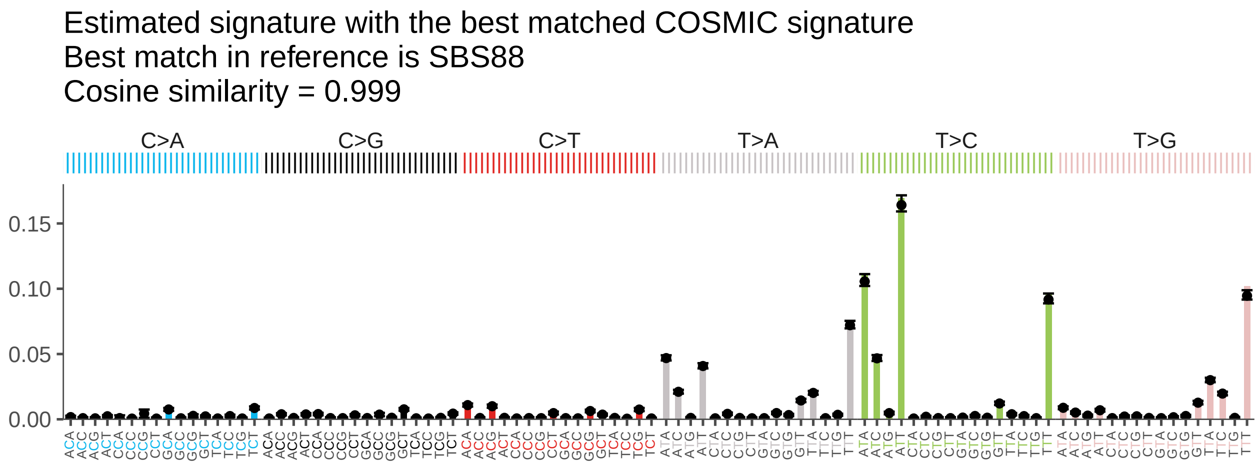
- `res`: bayesNMF object
- `sig`: selection of estimated signature. Integer, index of the estimated signature to plot. Can be NULL (default) to plot a reference signature on its own
- `ref`: selection of reference signature. Integer, column name string, or "best" (default). If "best", the reference signature with highest cosine similarity to the estimated signature will be used, and it's identification will be reported in the title. Can be NULL to plot an estimated signature on its own
- `ref_matrix`: reference signatures to align to in the form of a reference P matrix or the string `"cosmic"` (default)
- `title`: optional first line of the title
- `cosine`: whether to report cosine similarity between estimate and reference in the title (ignored if `sig` or `ref` is NULL)

Output:

- `ggplot2` object

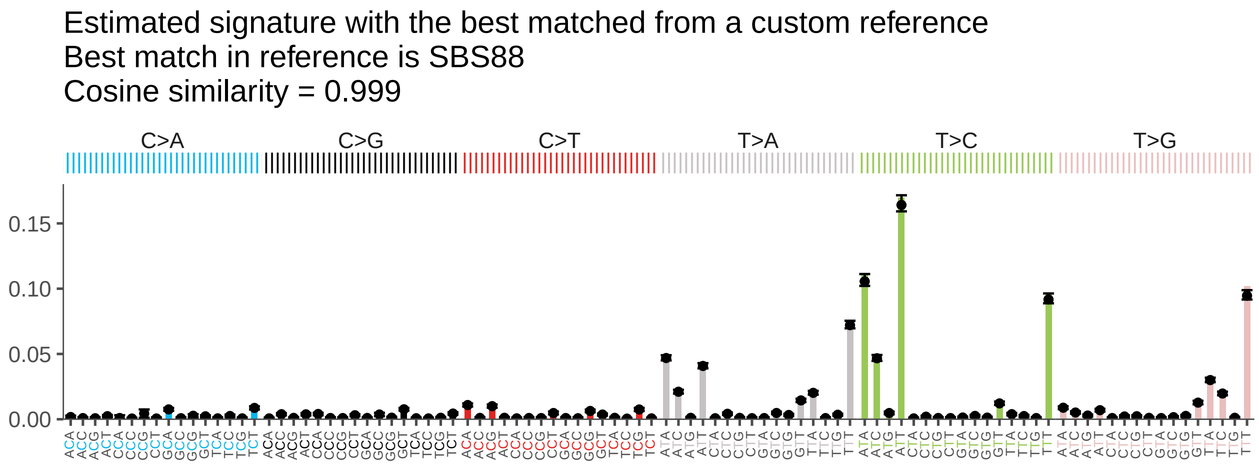
Plotting an estimated signature with the best matched COSMIC signature

```
plot_sig(
    res = res, sig = 1,
    title = "Estimated signature with the best matched COSMIC signature"
)
```



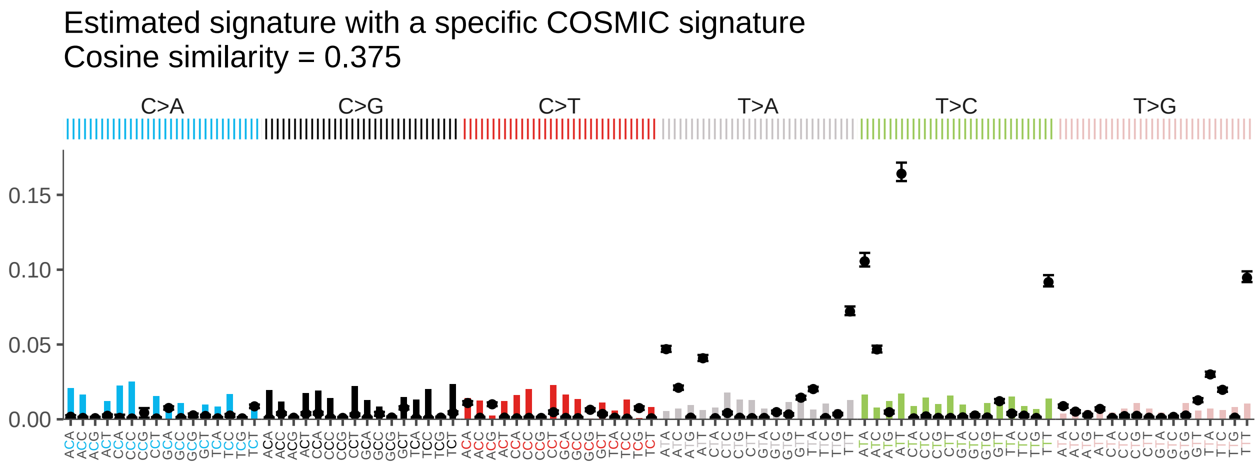
Plotting an estimated signature with the best matched signature from a custom reference

```r
plot_sig(
    res = res, sig = 1, ref_matrix = data$P,
    title = "Estimated signature with the best matched from a custom reference"
)
```



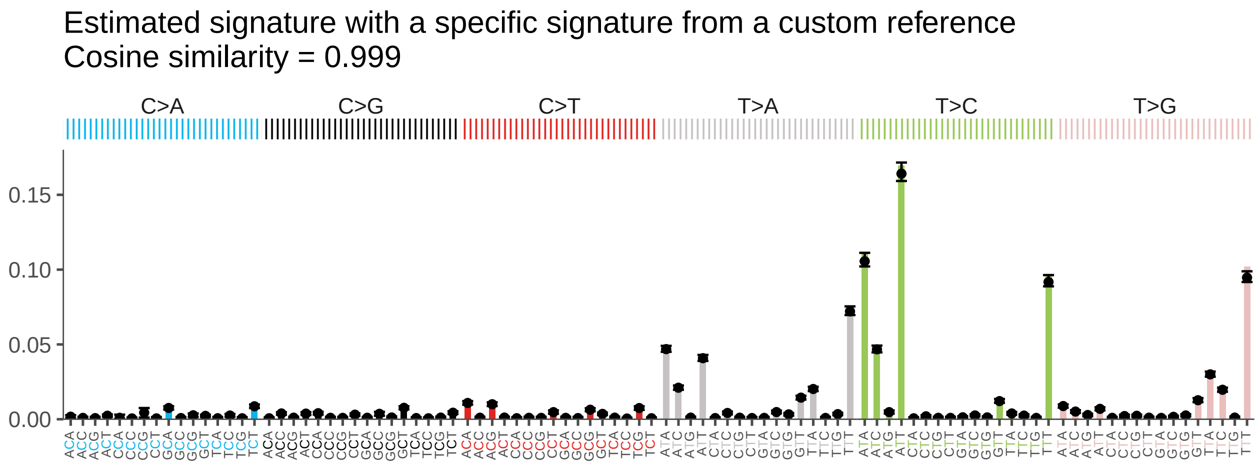
Plot an estimated signature with a specific COSMIC signature

```r
plot_sig(
    res = res, sig = 1, ref = "SBS3",
    title = "Estimated signature with a specific COSMIC signature"
)
```



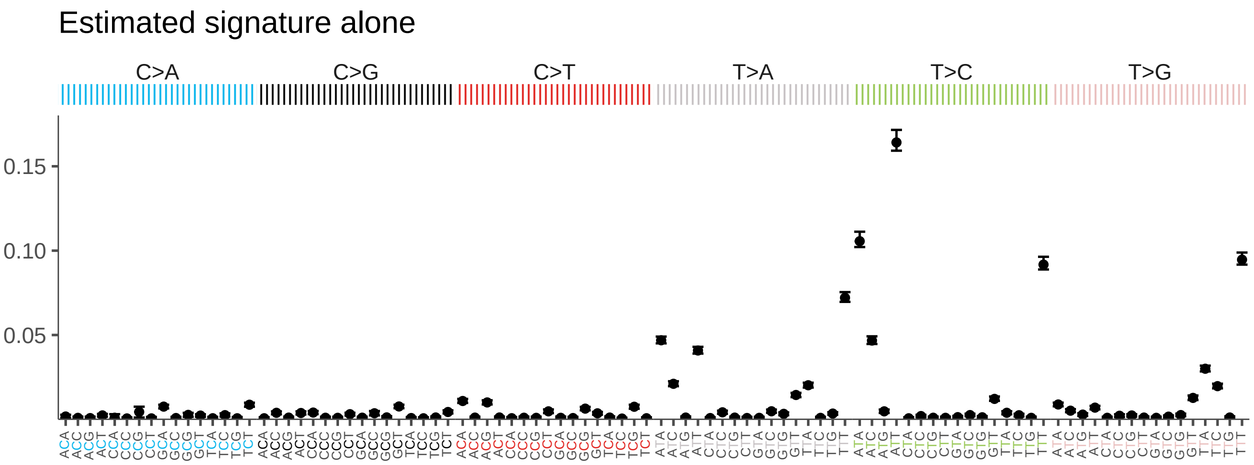
Plot an estimated signature with a specific signature from a custom reference

```r
plot_sig(
    res = res, sig = 1, ref = "SBS88", ref_matrix = data$P,
    title = "Estimated signature with a specific signature from a custom reference"
)
```



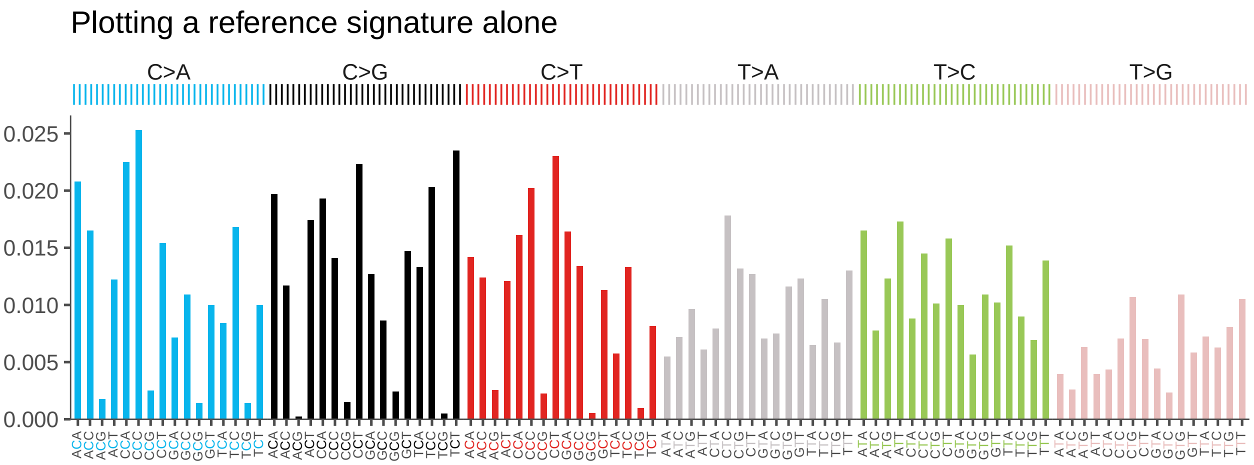
Plotting an estimated signature alone

```r
plot_sig(
    res = res, sig = 1, ref = NULL,
    title = "Estimated signature alone"
)
```



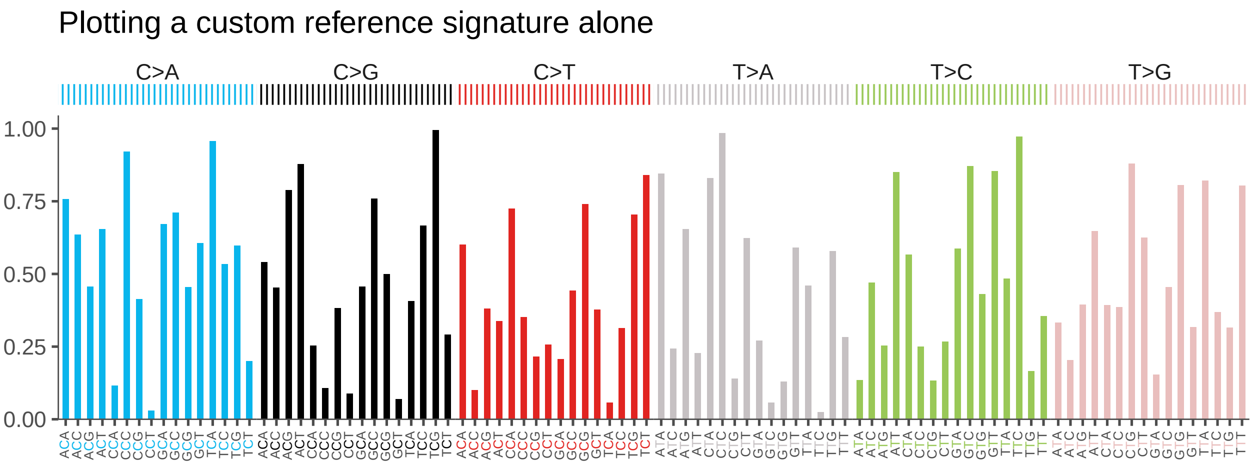
Plotting a reference signature alone

```r
plot_sig(ref = "SBS3", title = "Plotting a reference signature alone")
```



Plotting a custom reference signature alone

```r
plot_sig(ref = runif(96), title = "Plotting a custom reference signature alone")
```



Plotting all signatures

The `plot_results` function displays which reference signatures are recovered in one or more bayesNMF results objects. It reports MAP (posterior average) cosine similarity to the assigned reference as size, and median (across samples with contribution > 0) MAP contributions as color. The function returns the plot as well as a data frame with all of the visualized information if `return_df = TRUE`.

Input:

- `res_list`: named list containing one or more bayesNMF objects. Names will become identifiers along the top of the plot.
- `ref_matrix`: reference signatures to align to in the form of a reference P matrix or the string `"cosmic"` (default).
- `title`: optional plot title
- `return_df`: whether to return data frame with all visualized information (default `TRUE`)

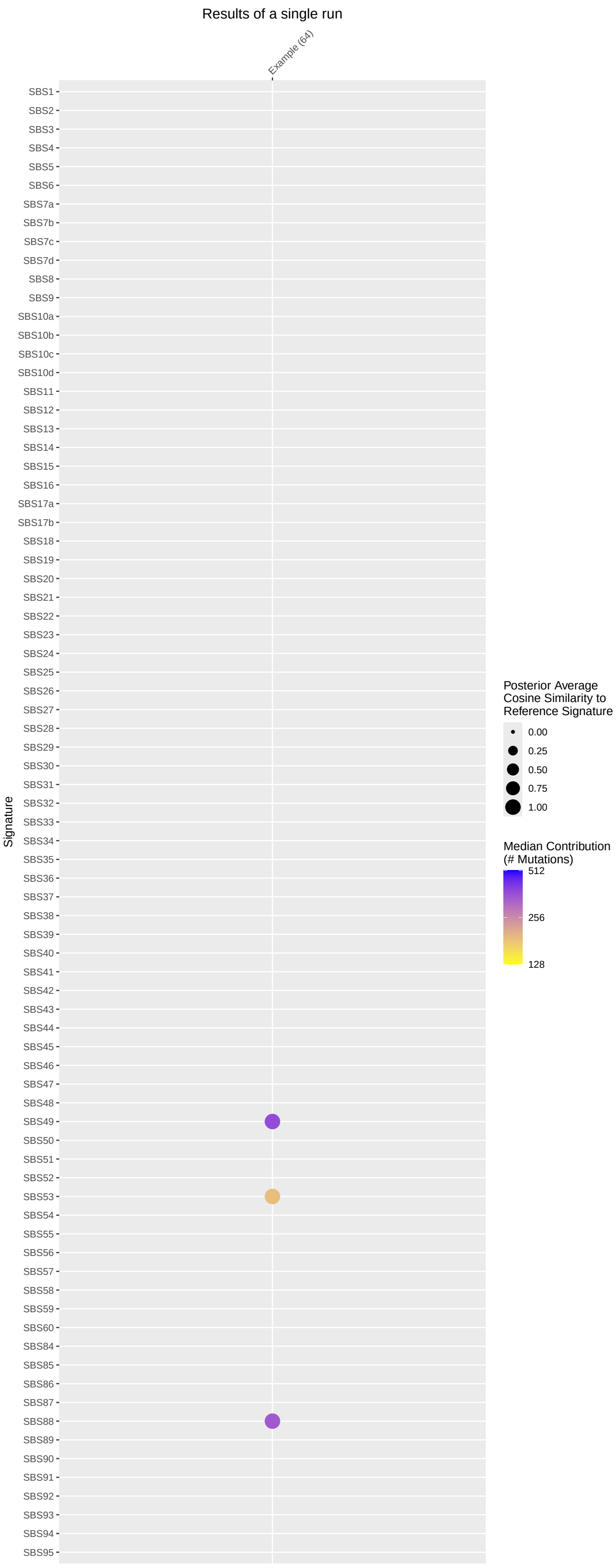
Output - `ggplot2` object - `data.frame` if `return_df = TRUE`

Results of a single run with respect to COSMIC reference

```
pl <- plot_results(
    list("Example" = res),
    title = "Results of a single run"
)
pl$df
```

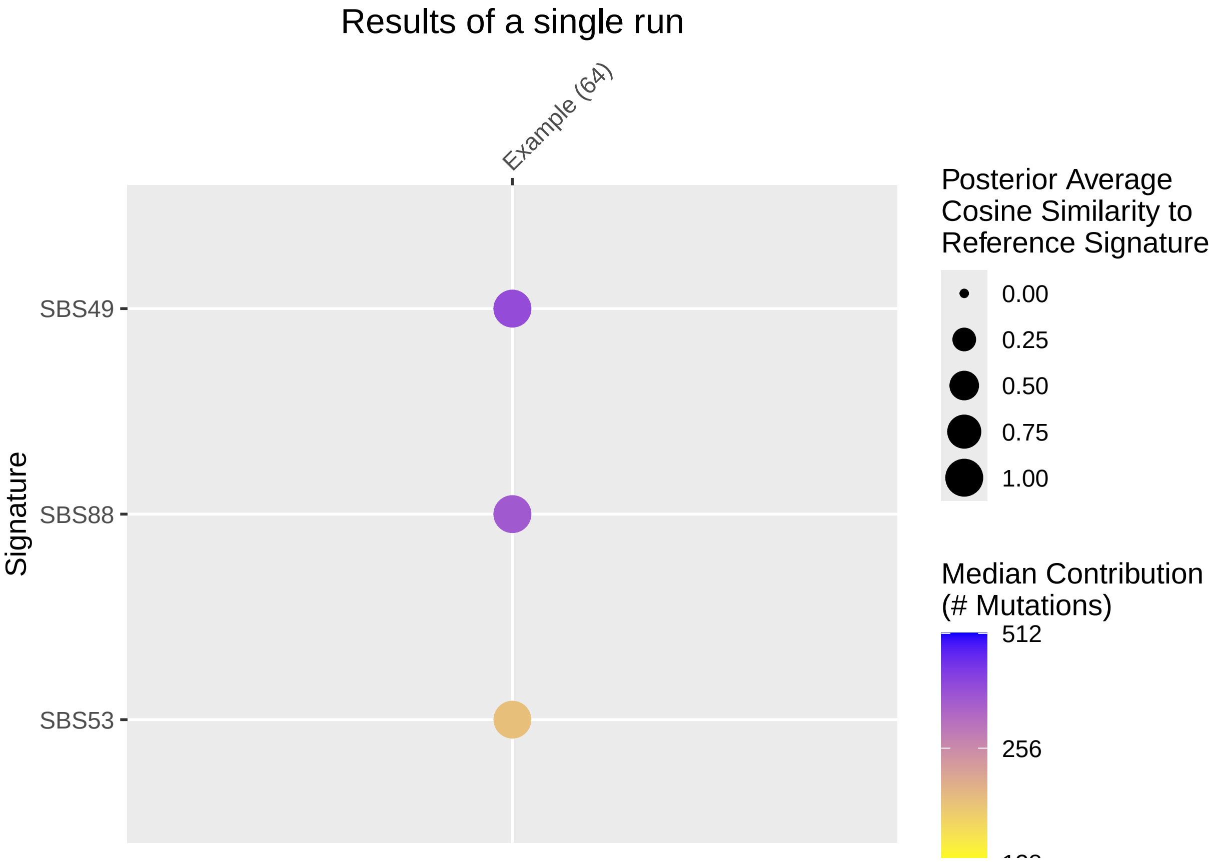
```
##           Name Signature  G Med_Contribution Cosine_Similarity
## 1 Example (64)     SBS49 64         370.0759         0.9993091
## 2 Example (64)     SBS53 64         186.3619         0.9919638
## 3 Example (64)     SBS88 64         343.9838         0.9990684
```

```
pl$plot
```


Results of a single run
Example (64)
Signature
SBS1
SBS2
SBS3
SBS4
SBS5
SBS6
SBS7a
SBS7b
SBS7c
SBS7d
SBS8
SBS9
SBS10a
SBS10b
SBS10c
SBS10d
SBS11
SBS12
SBS13
SBS14
SBS15
SBS16
SBS17a
SBS17b
SBS18
SBS19
SBS20
SBS21
SBS22
SBS23
SBS24
SBS25
SBS26
SBS27
SBS28
SBS29
SBS30
SBS31
SBS32
SBS33
SBS34
SBS35
SBS36
SBS37
SBS38
SBS39
SBS40
SBS41
SBS42
SBS43
SBS44
SBS45
SBS46
SBS47
SBS48
SBS49
SBS50
SBS51
SBS52
SBS53
SBS54
SBS55
SBS56
SBS57
SBS58
SBS59
SBS60
SBS84
SBS85
SBS86
SBS87
SBS88
SBS89
SBS90
SBS91
SBS92
SBS93
SBS94
SBS95
Posterior Average Cosine Similarity to Reference Signature
0.00
0.25
0.50
0.75
1.00
Median Contribution (# Mutations)
512
256
128

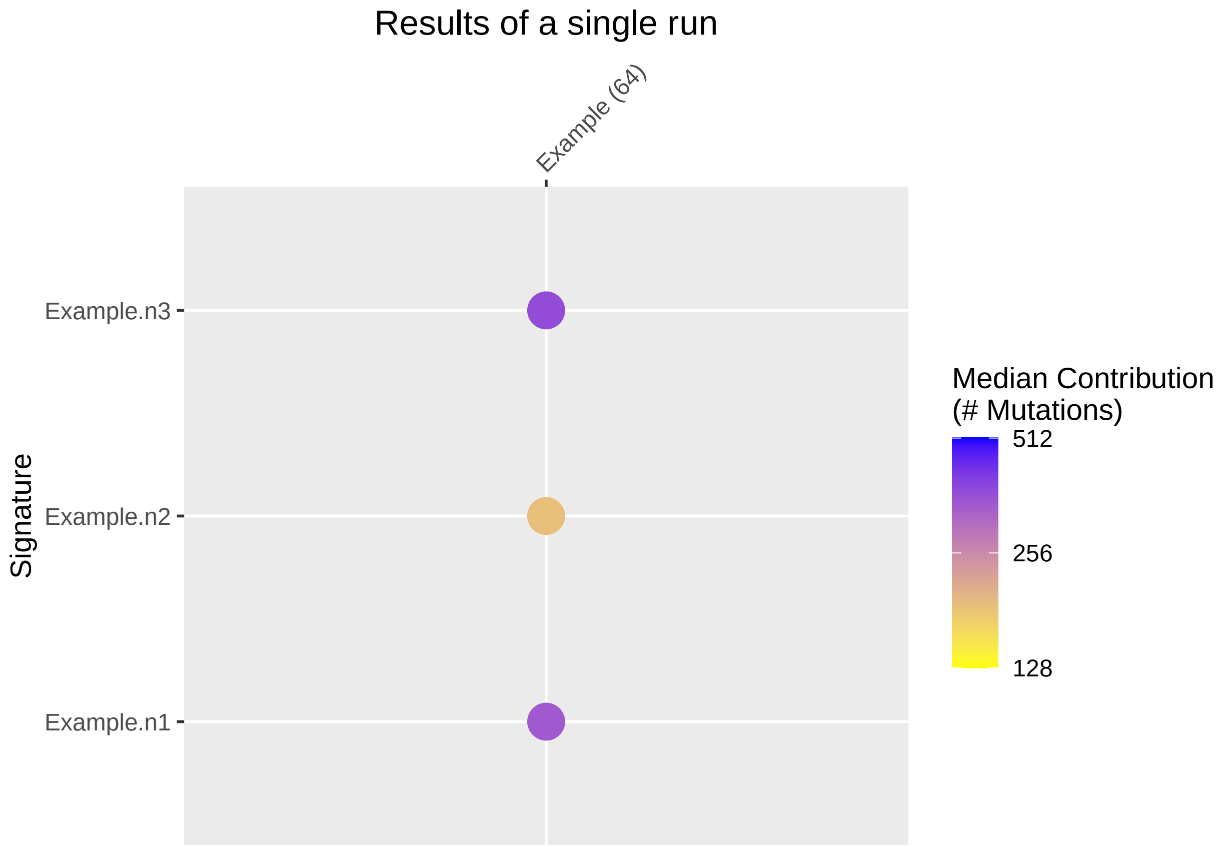
Results of a single run with respect to a custom reference

```r
pl <- plot_results(
    list("Example" = res), ref_matrix = data$P,
    title = "Results of a single run"
)
pl$plot
```



Results of a single run with no reference

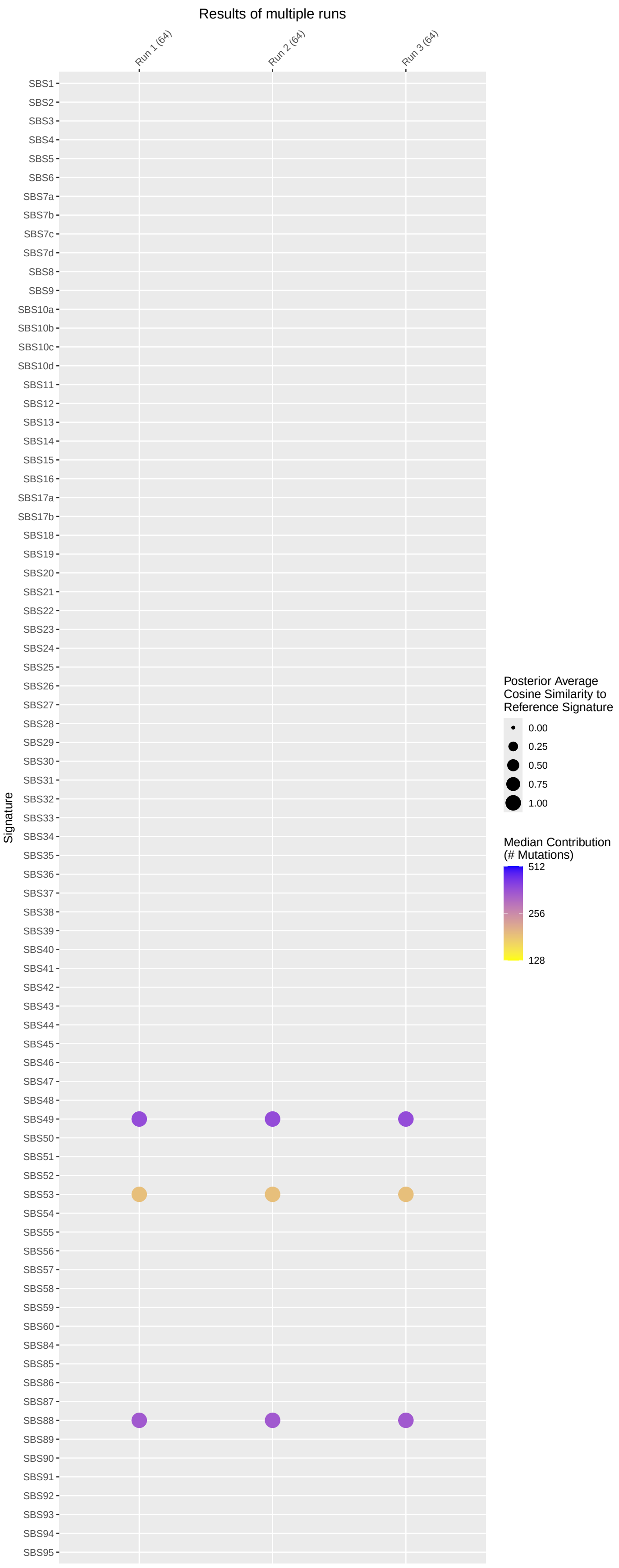
```r
pl <- plot_results(
    list("Example" = res), ref_matrix = NULL,
    title = "Results of a single run"
)
pl$plot
```



Results of multiple runs against COSMIC reference

For the sake of example, we are just using the same run three times, but in practice you could have results of bayesNMF on different cancer types as we show in the paper, across technical batches, etc.

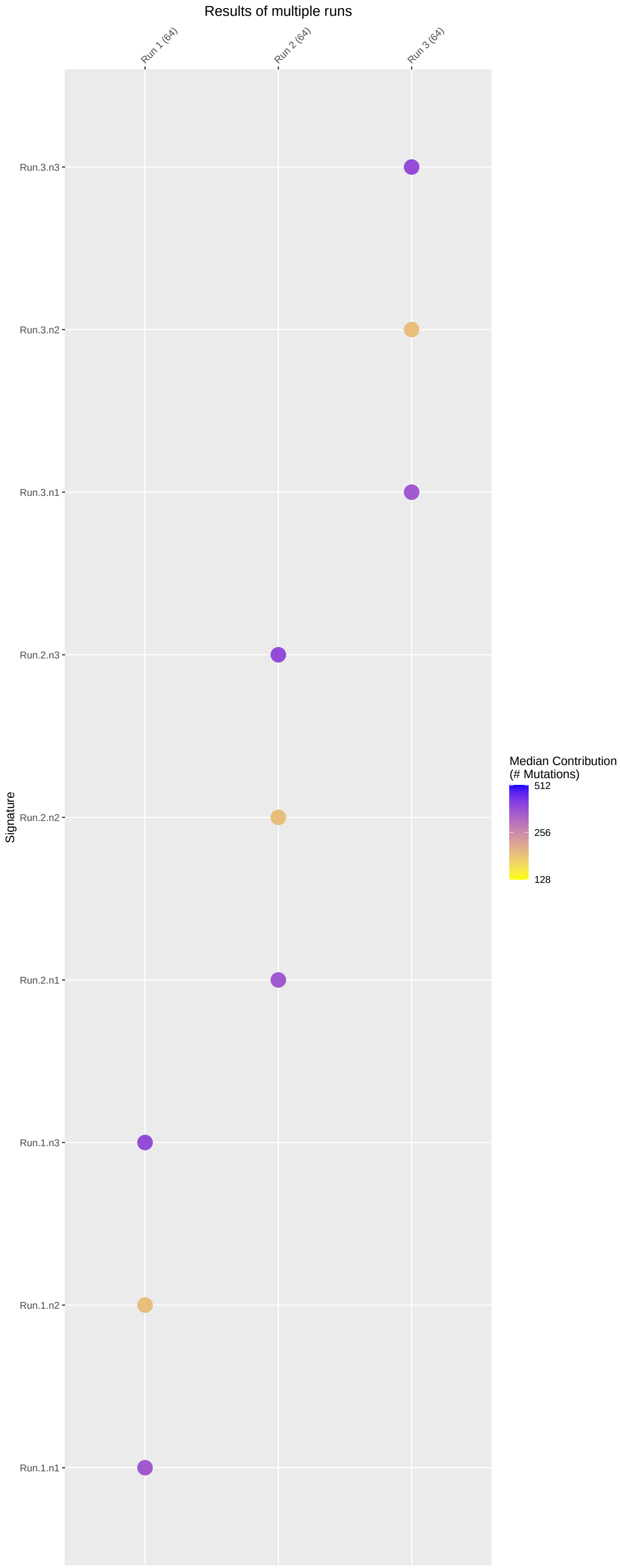
```
pl <- plot_results(
    list("Run 1" = res, "Run 2" = res, "Run 3" = res),
    title = "Results of multiple runs"
)
pl$plot
```


Results of multiple runs
Run 1 (64)
Run 2 (64)
Run 3 (64)
Signature
SBS1
SBS2
SBS3
SBS4
SBS5
SBS6
SBS7a
SBS7b
SBS7c
SBS7d
SBS8
SBS9
SBS10a
SBS10b
SBS10c
SBS10d
SBS11
SBS12
SBS13
SBS14
SBS15
SBS16
SBS17a
SBS17b
SBS18
SBS19
SBS20
SBS21
SBS22
SBS23
SBS24
SBS25
SBS26
SBS27
SBS28
SBS29
SBS30
SBS31
SBS32
SBS33
SBS34
SBS35
SBS36
SBS37
SBS38
SBS39
SBS40
SBS41
SBS42
SBS43
SBS44
SBS45
SBS46
SBS47
SBS48
SBS49
SBS50
SBS51
SBS52
SBS53
SBS54
SBS55
SBS56
SBS57
SBS58
SBS59
SBS60
SBS84
SBS85
SBS86
SBS87
SBS88
SBS89
SBS90
SBS91
SBS92
SBS93
SBS94
SBS95
Posterior Average
Cosine Similarity to
Reference Signature
0.00
0.25
0.50
0.75
1.00
Median Contribution
(# Mutations)
512
256
128

Results of multiple runs against no reference. When signatures are not aligned to a common reference, they are treated as separate. Even though we know the three estimated signatures are repeated (since all three results objects are the same), they are not aligned to one another since a reference is excluded.

```
pl <- plot_results(
    list("Run 1" = res, "Run 2" = res, "Run 3" = res),
    ref_matrix = NULL,
    title = "Results of multiple runs"
)

pl$plot
```


Results of multiple runs
Run 1 (64)
Run 2 (64)
Run 3 (64)
Run.3.n3
Run.3.n2
Run.3.n1
Run.2.n3
Run.2.n2
Run.2.n1
Run.1.n3
Run.1.n2
Run.1.n1
Signature
Median Contribution
(# Mutations)
512
256
128

Plotting signature distribution on the mutation-type level

The `plot_signature_dist` function plots the distribution of mutational counts across signatures for each mutation type across all, a subset, or one subject. This visualizes, for each trinucleotide mutation, how many are present and what signatures they are attributed to.

Input:

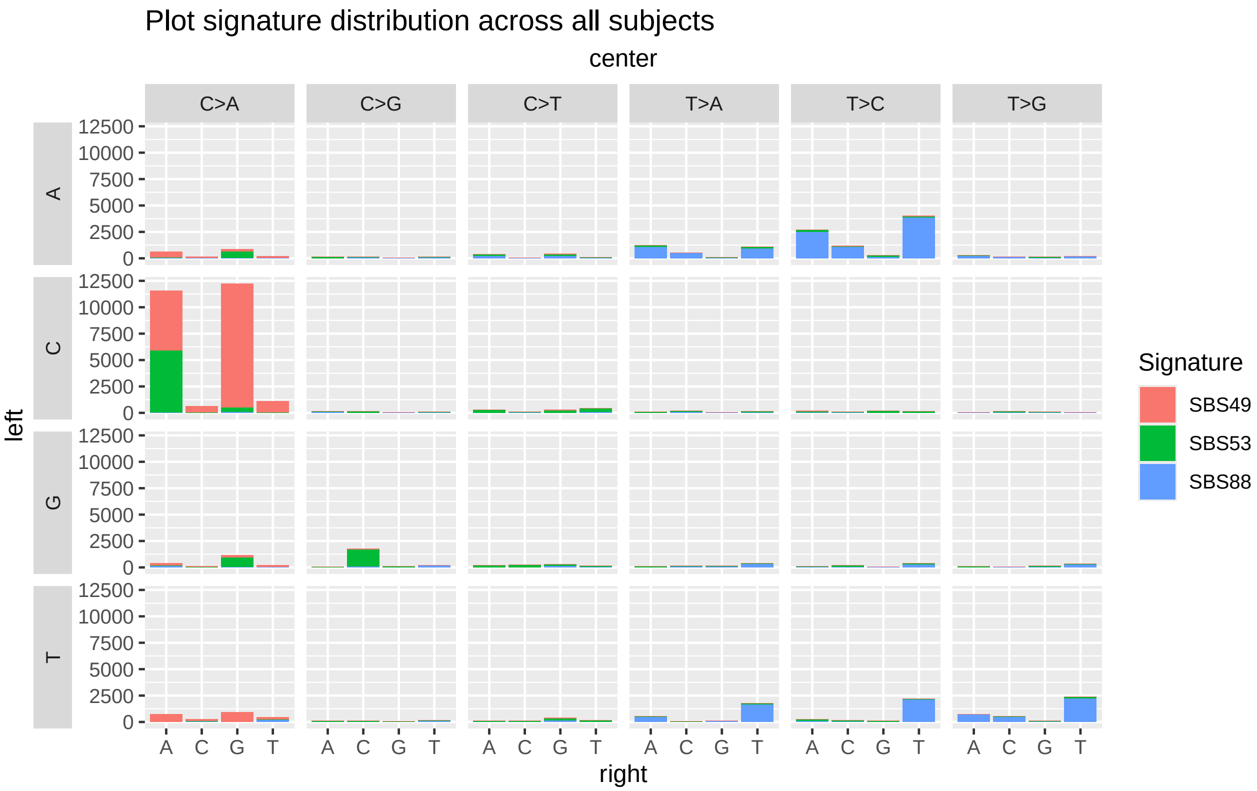
- `res`: bayesNMF results object
- `subjects`: vector of subject indices to include in plot, defaults to all subjects
- `ref_matrix`: reference signatures to align to in the form of a reference P matrix or the string `"cosmic"` (default).
- `title`: plot title, default "Distribution of Signature Allocation"

Output:

- `ggplot2` object

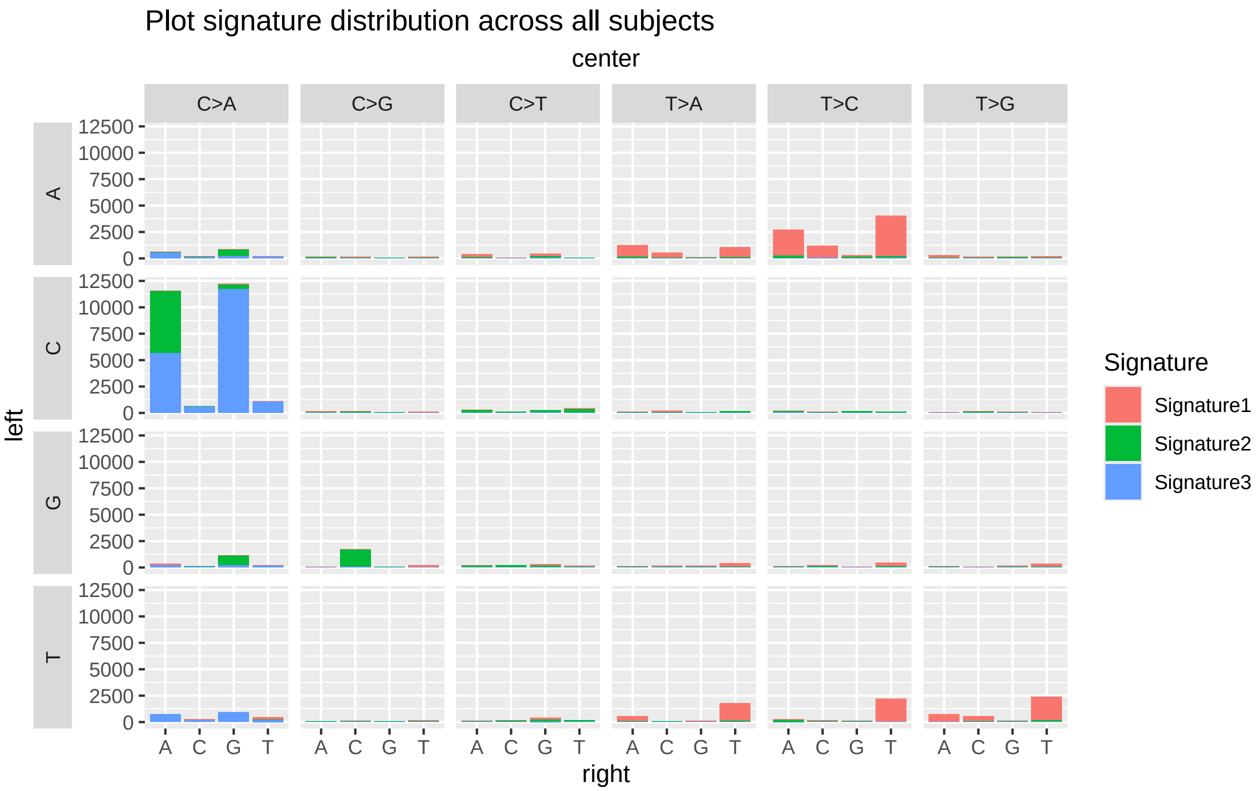
Plot signature distribution across all subjects.

```
plot_signature_dist(
    res,
    title = "Plot signature distribution across all subjects"
)
```



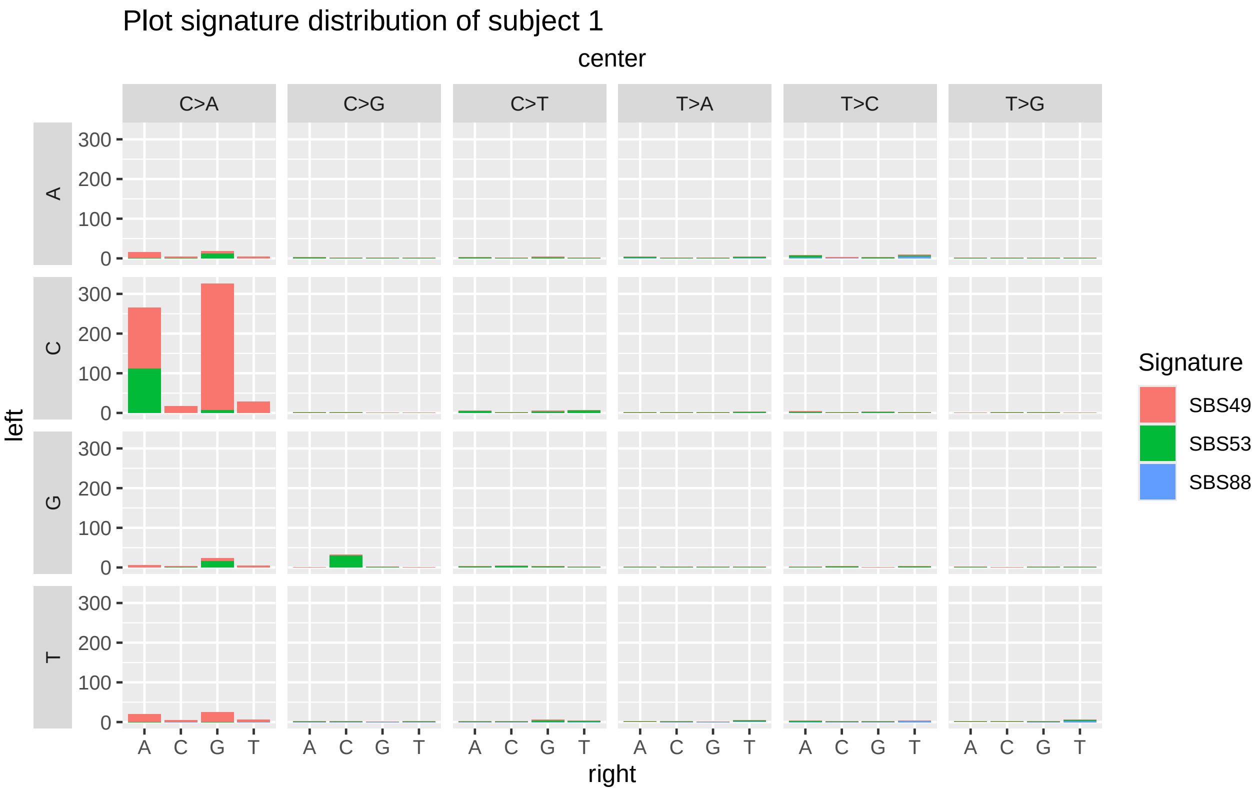
Plot signature distribution across all subjects without a reference.

```
plot_signature_dist(
    res,
    title = "Plot signature distribution across all subjects",
    ref_matrix = NULL
)
```



Plot signature distribution of a particular subject

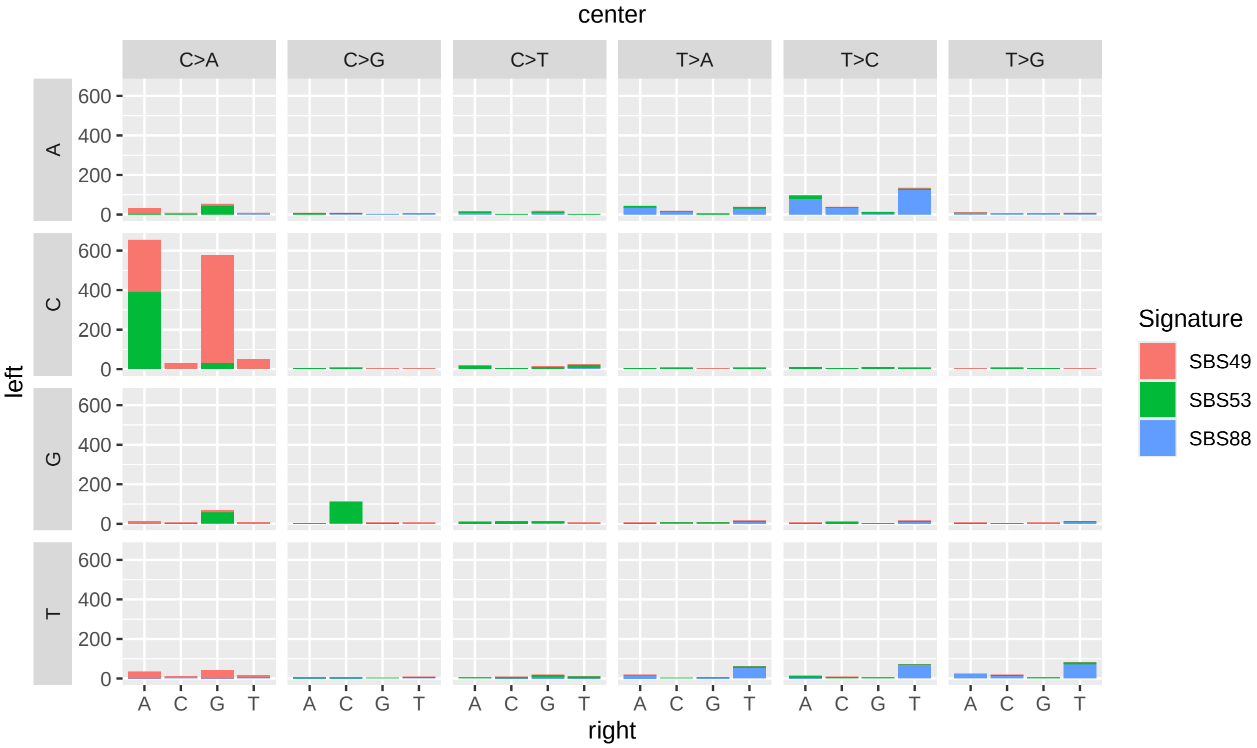
```
plot_signature_dist(
    res,
    subject = 1,
    title = "Plot signature distribution of subject 1"
)
```



Plot signature distribution of a subset of subjects

```
plot_signature_dist(
    res, subject = c(1,4,10),
    title = "Plot signature distribution of subjects 1, 4, and 10"
)
```

Plot signature distribution of subjects 1, 4, and 10


center
C>A
C>G
C>T
T>A
T>C
T>G
A
C
G
T
left
right
A
C
G
T
Signature
🟥 SBS49
🟩 SBS53
🟦 SBS88