Descriptive statistics

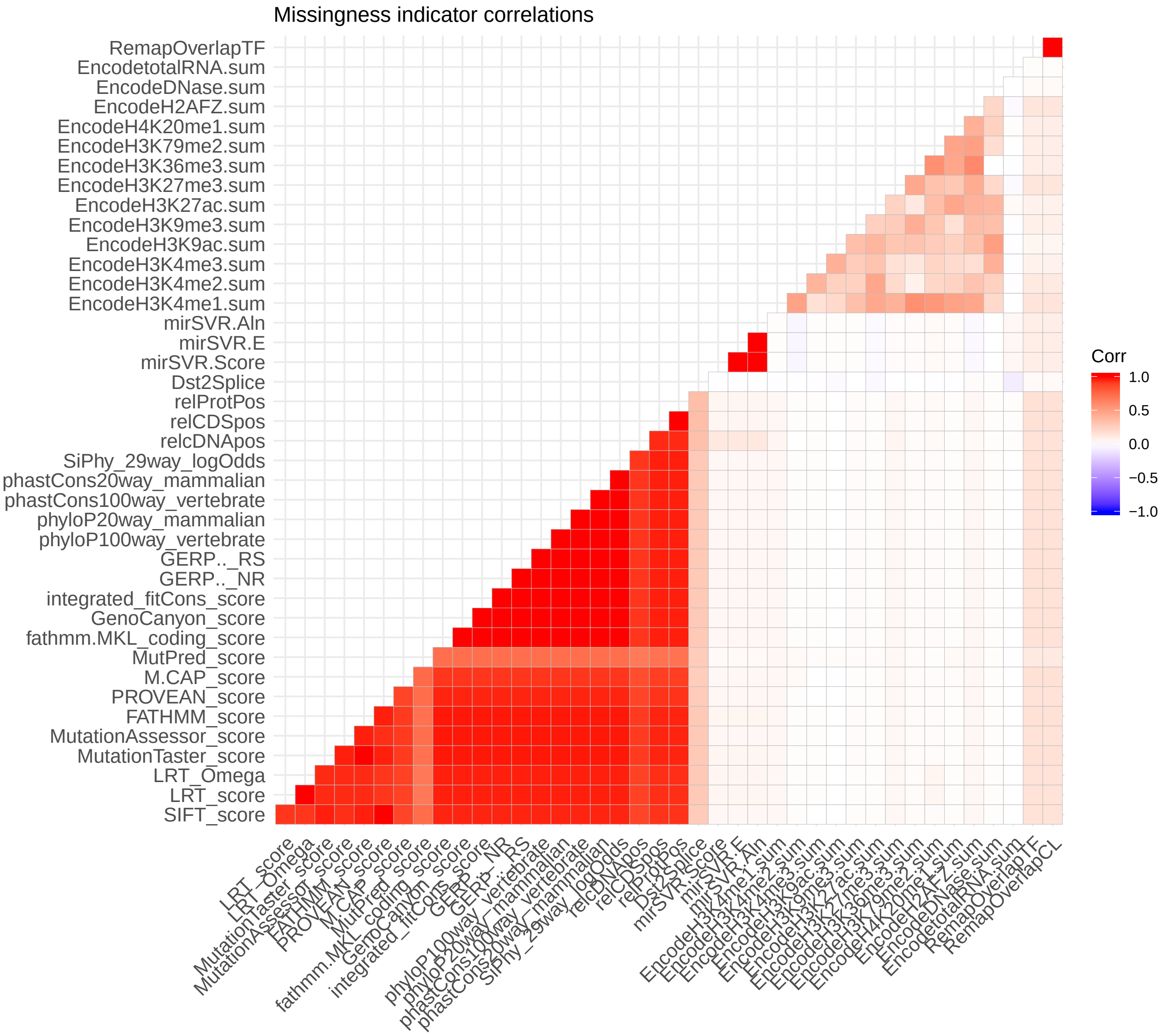
```
library(magrittr)
library(ggplot2)
library(gridExtra)
library(ggcorrplot)

source("../R/visualizations.R")
source("../R/feature_definitions.R")

training_set <- read.csv("../preprocessed_training_data.csv", row.names = 1, as.is = TRUE)
outcome <- read.csv("../training_outcomes.csv", row.names = 1)[,1]

stopifnot(row.names(training_set) == row.names(outcome))

features <- colnames(training_set)
```

Correlations

Plot correlation matrices of missingness indicators against missingness indicators, observed values against observed values, and missingness indicators against observed values.

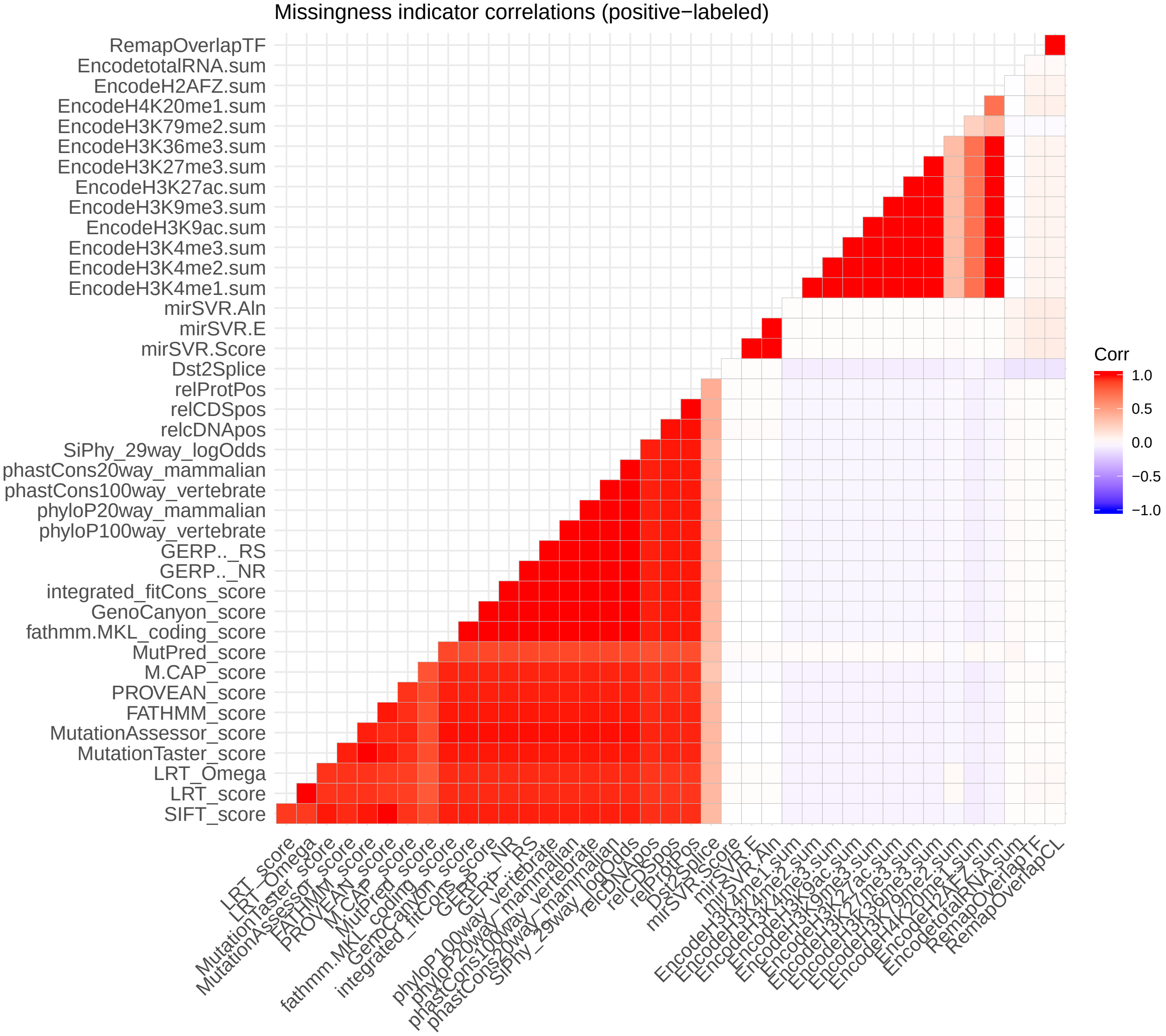
```
positive_data <- training_set[outcome == "positive", ]
negative_data <- training_set[outcome == "negative", ]

# Missingness indicator correlations
plot_missingness_correlations(training_set, numeric_features, "Missingness indicator correlations")
```

```
## Warning in cor(miss_data[, features]): the standard deviation is zero
```



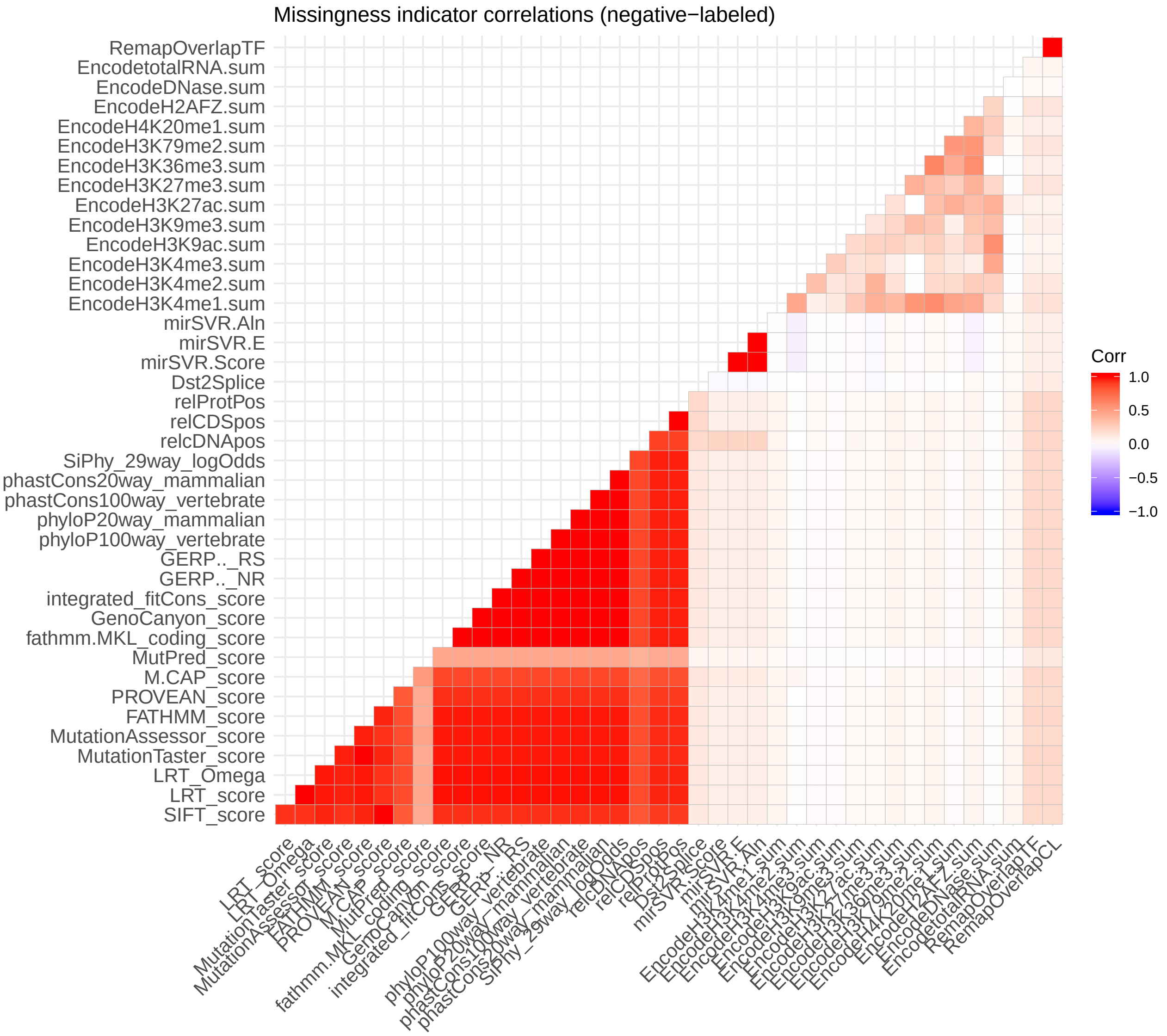
```
plot_missingness_correlations(positive_data, numeric_features, "Missingness indicator correlations (pos
```

```
## Warning in cor(miss_data[, features]): the standard deviation is zero
```



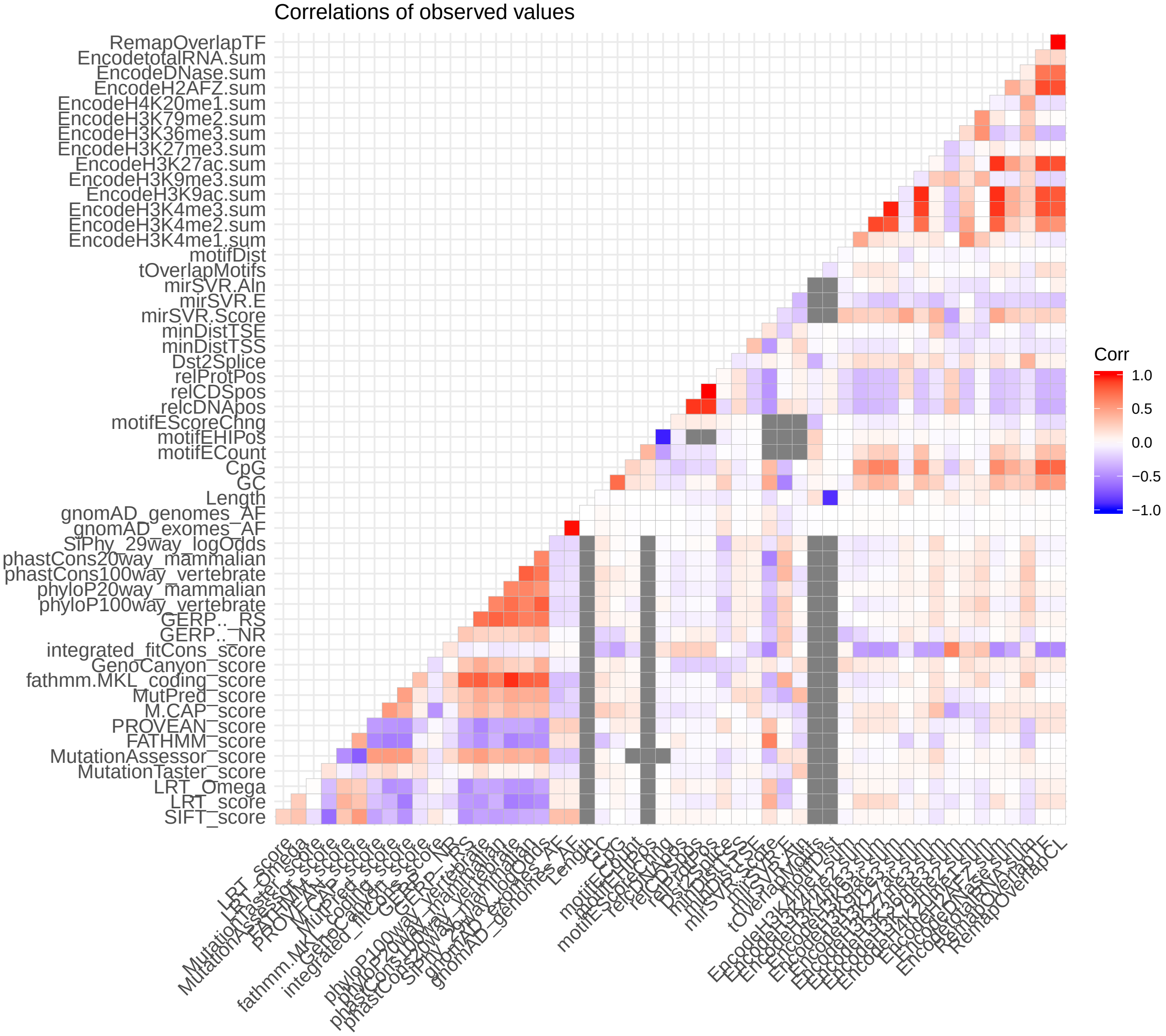
```
plot_missingness_correlations(negative_data, numeric_features, "Missingness indicator correlations (nega
```

```
## Warning in cor(miss_data[, features]): the standard deviation is zero
```



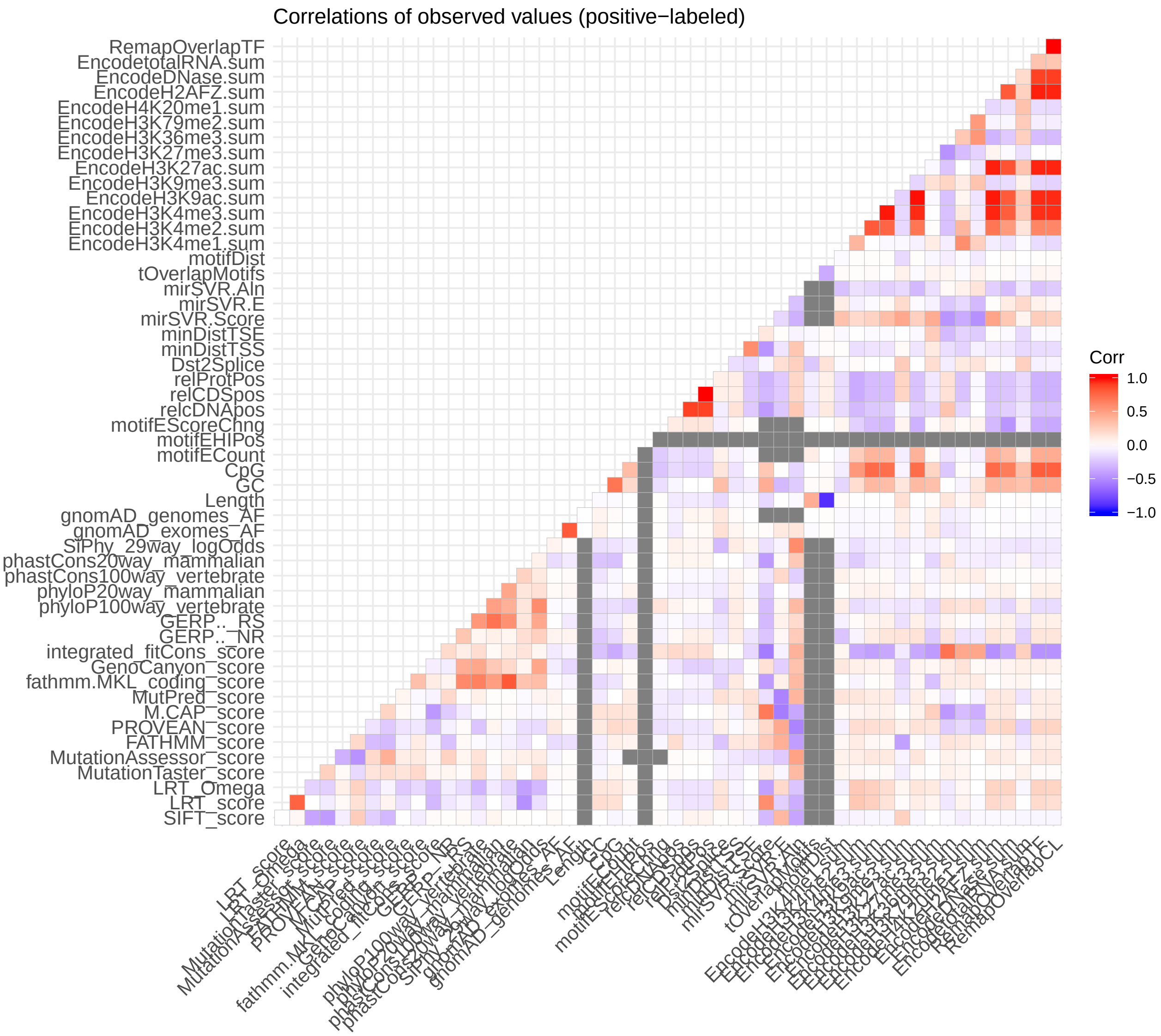
```r
# Observed value correlations
plot_observed_correlations(training_set, numeric_features, "Correlations of observed values")
```

```
## Warning in cor(data[, features], use = "pairwise.complete.obs"): the
## standard deviation is zero
```



```
plot_observed_correlations(positive_data, numeric_features, "Correlations of observed values (positive-
```

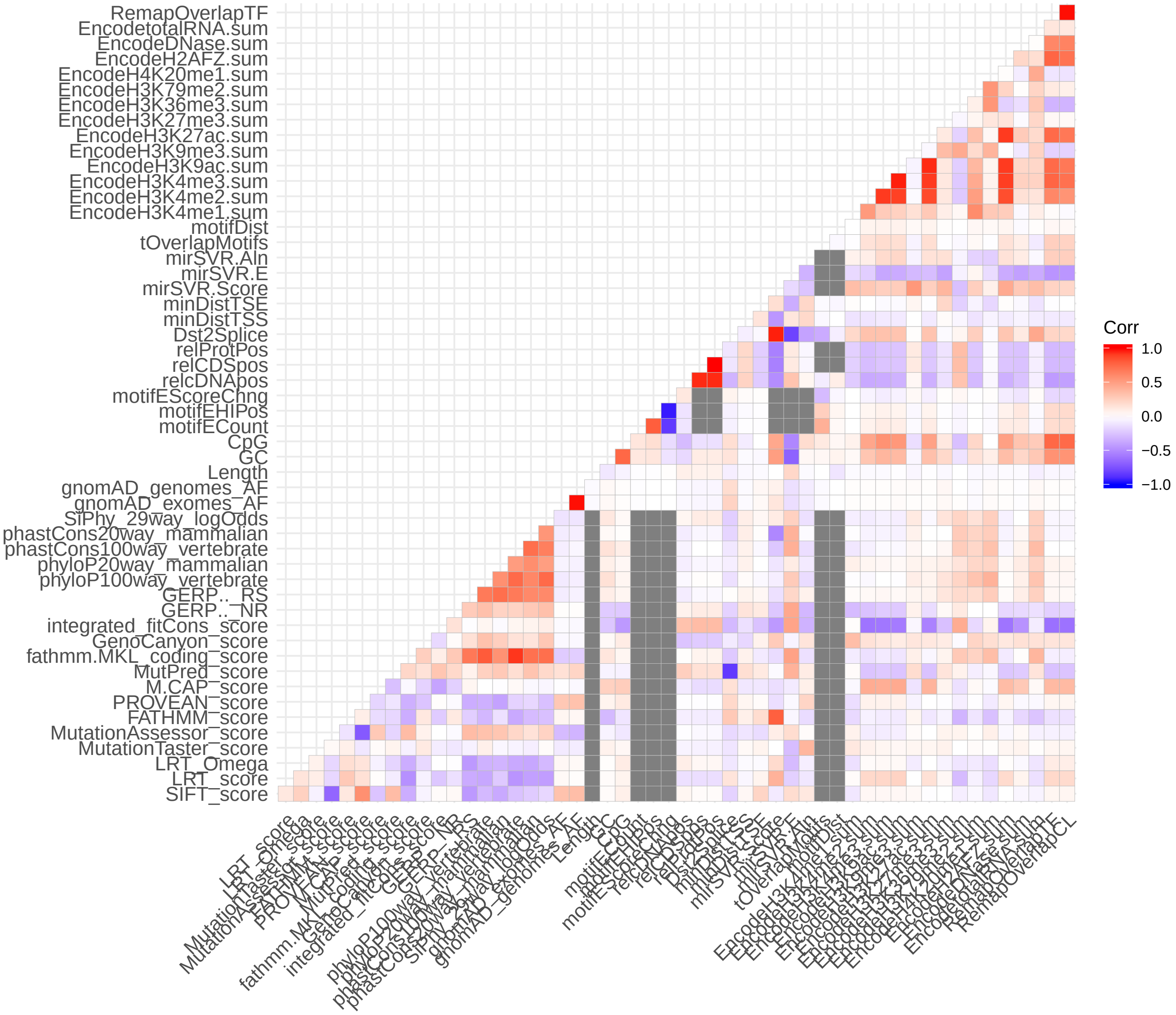
```
## Warning in cor(data[, features], use = "pairwise.complete.obs"): the
## standard deviation is zero
```



```r
plot_observed_correlations(negative_data, numeric_features, "Correlations of observed values (negative-
```

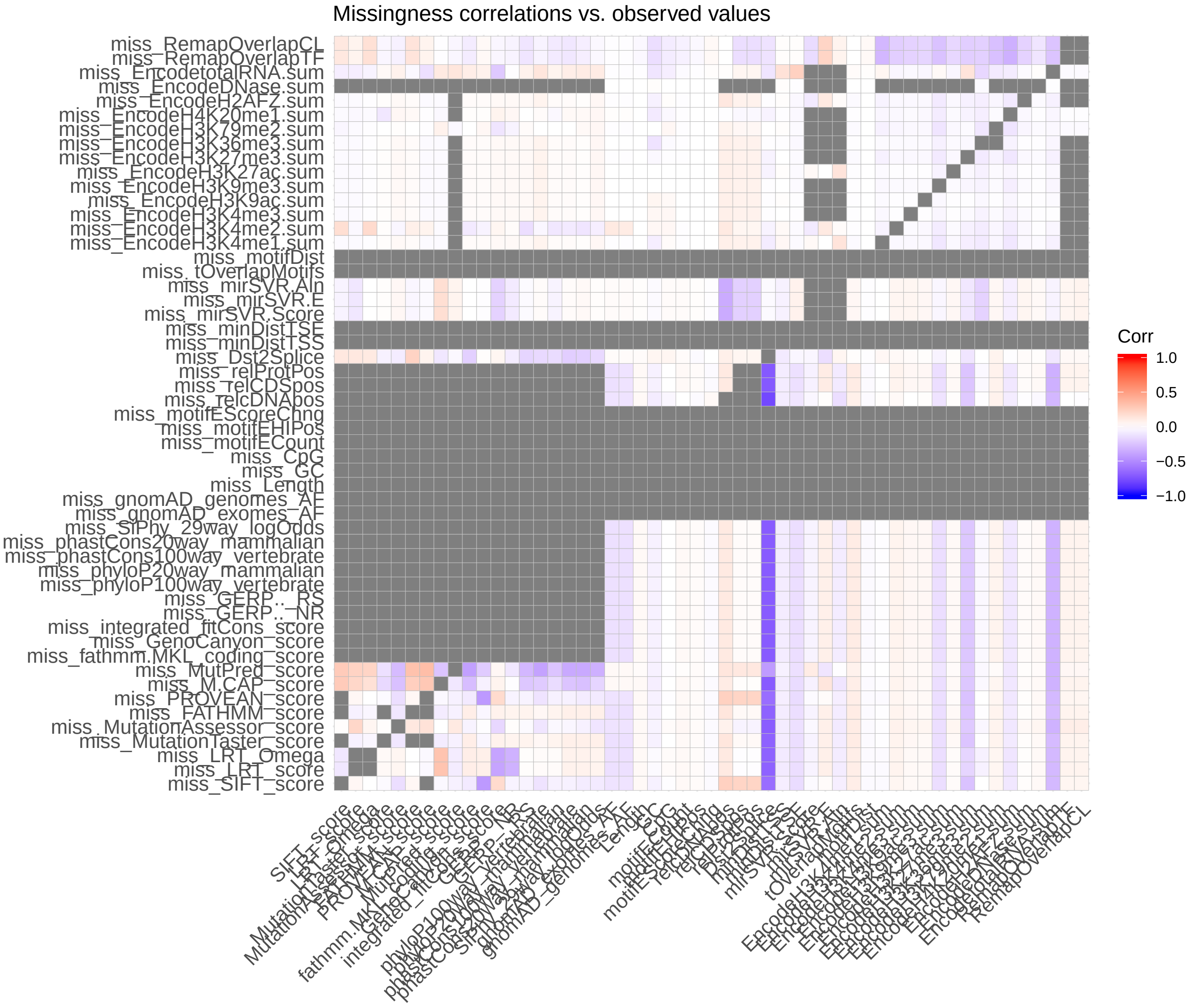
```
## Warning in cor(data[, features], use = "pairwise.complete.obs"): the
## standard deviation is zero
```

Correlations of observed values (negative−labeled)



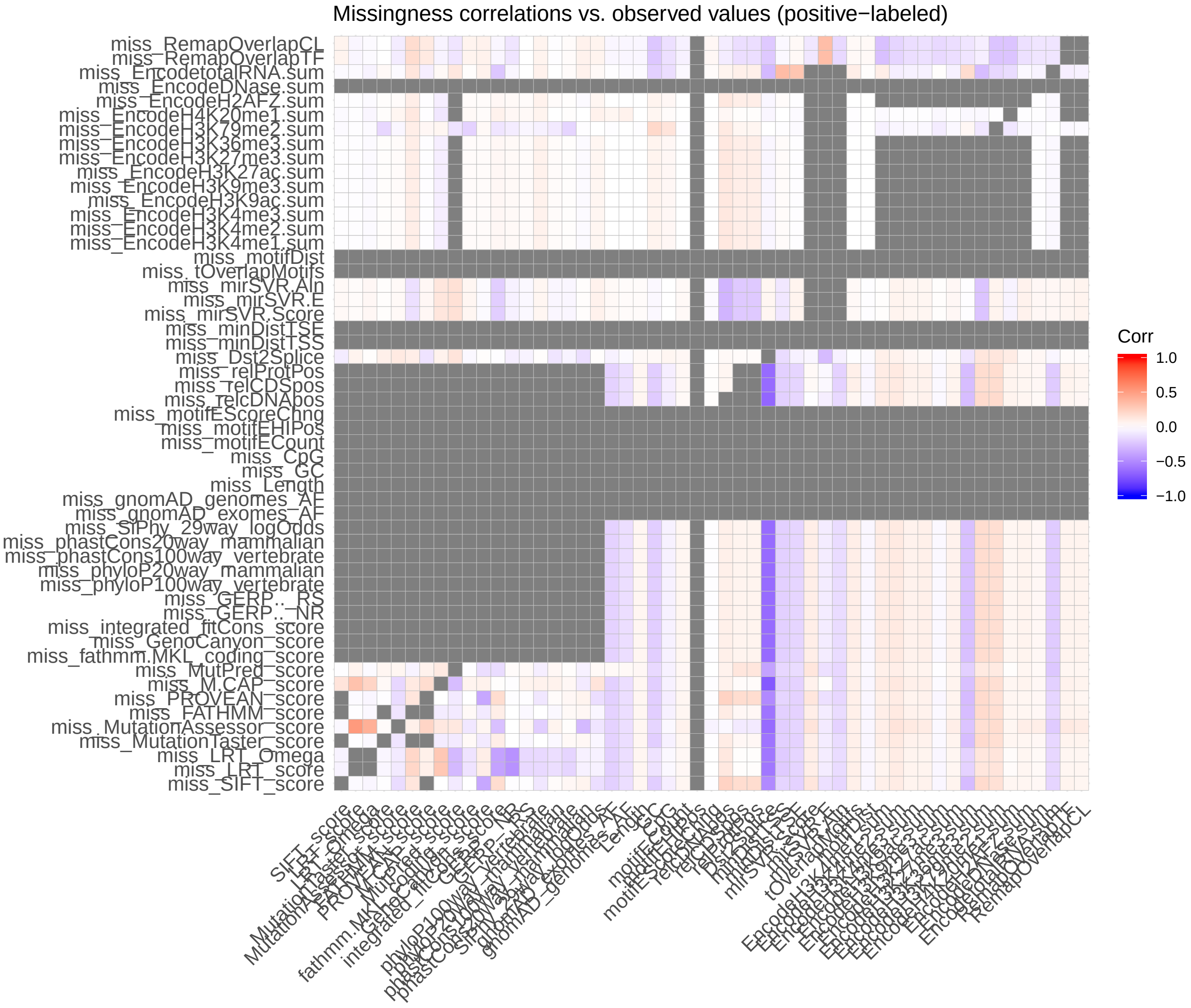
```
# Missingness vs. observed correlations
plot_missingness_vs_observed_correlations(training_set, numeric_features, "Missingness correlations vs.
```

```
## Warning in cor(data, miss_data, use = "pairwise.complete.obs"): the
## standard deviation is zero
```



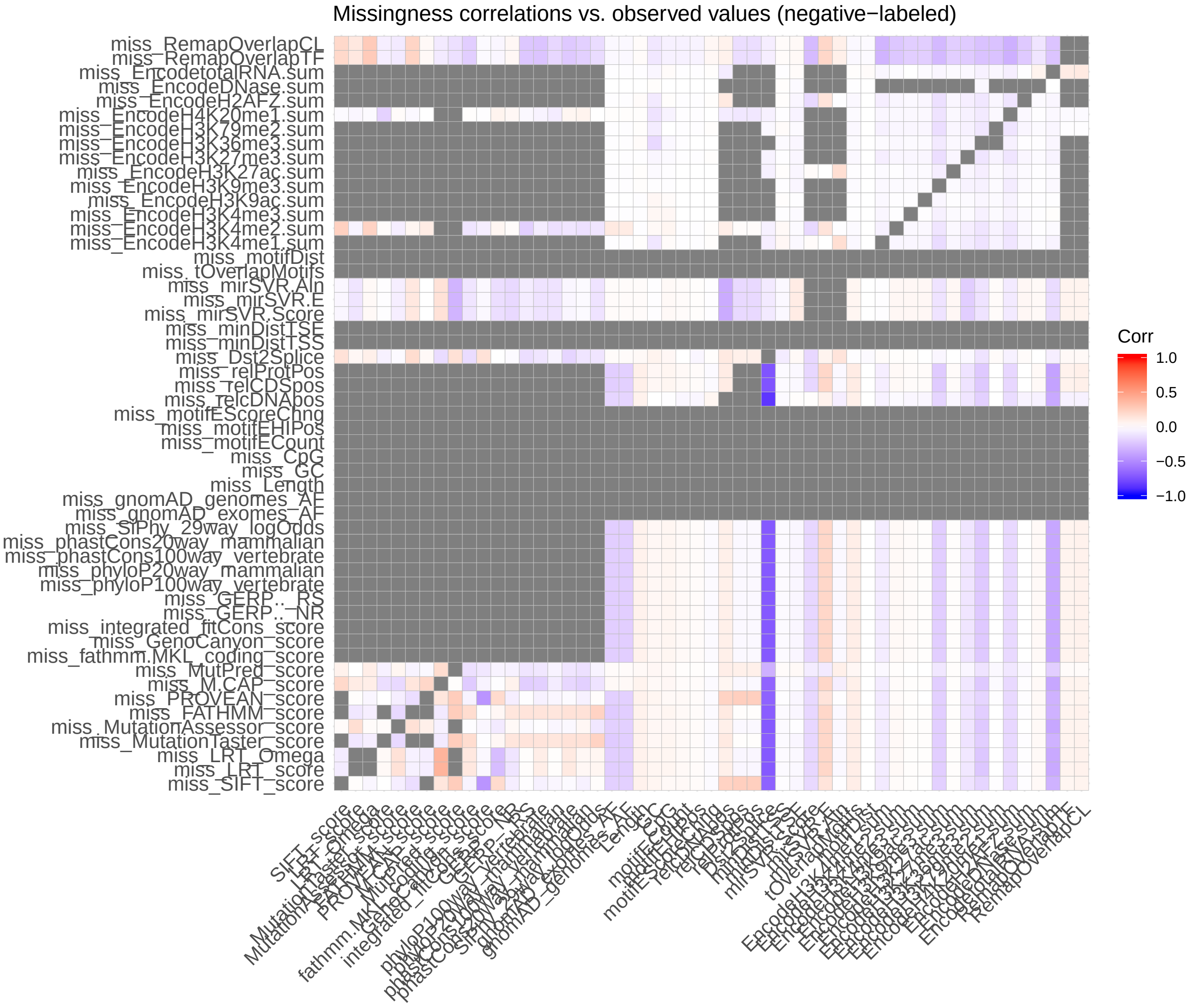
```
plot_missingness_vs_observed_correlations(positive_data, numeric_features, "Missingness correlations vs
```

```
## Warning in cor(data, miss_data, use = "pairwise.complete.obs"): the
## standard deviation is zero
```



```
plot_missingness_vs_observed_correlations(negative_data, numeric_features, "Missingness correlations vs
```

```
## Warning in cor(data, miss_data, use = "pairwise.complete.obs"): the
## standard deviation is zero
```

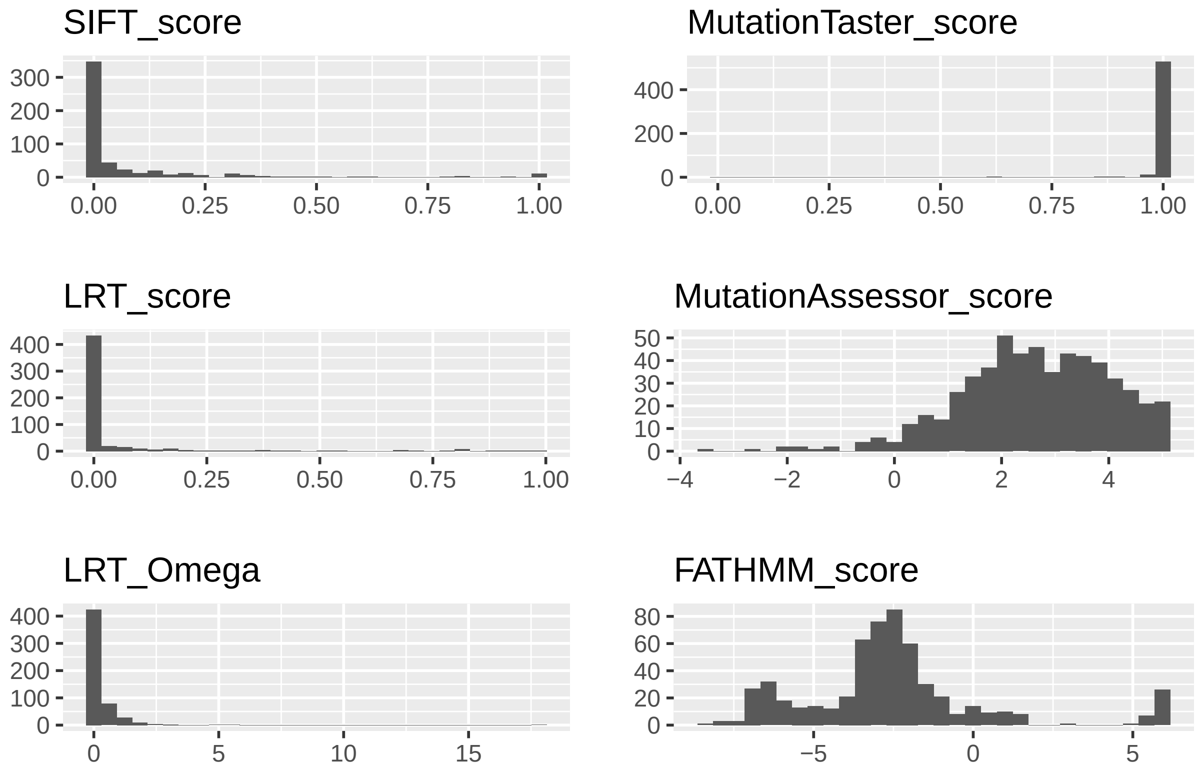


Feature value distributions

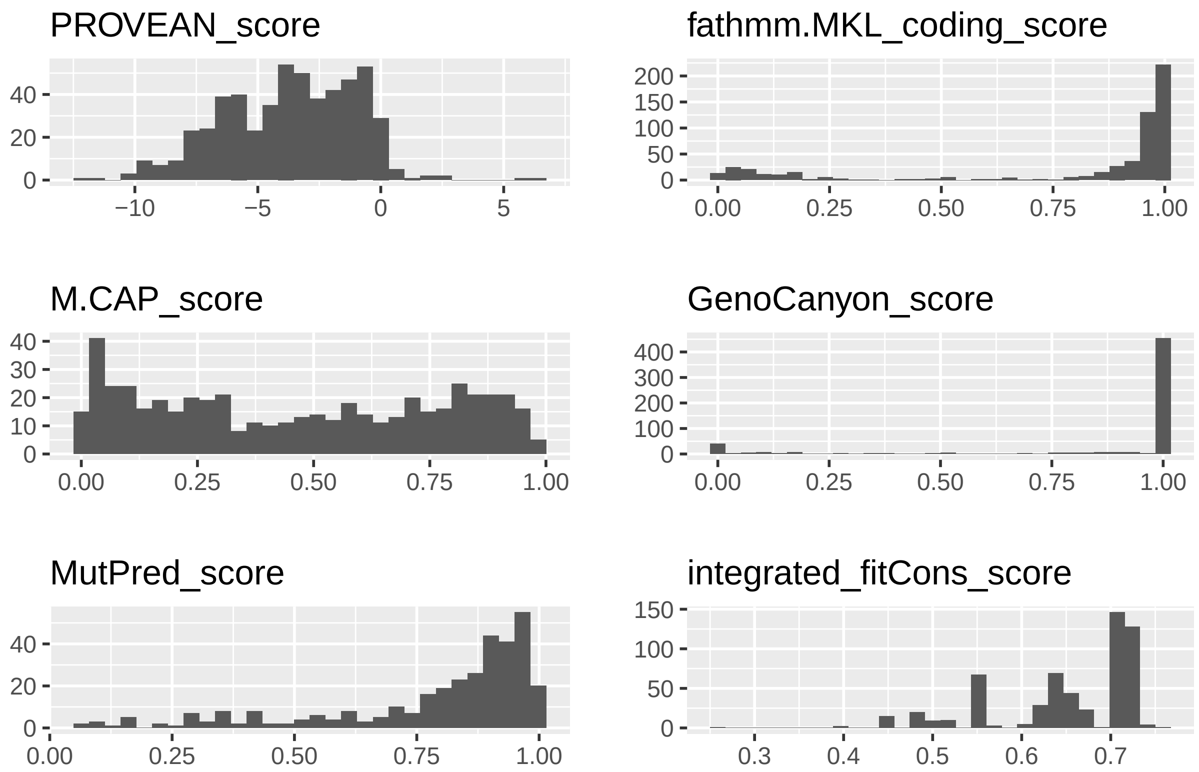
Next, plot distributions of each feature. Are they normal or linear?

```
feature_distribution_plots <- lapply(numeric_features,
                                     function(column) {
                                       ggplot2::quickplot(
                                         na.omit(training_set[,column]),
                                         main = column,
                                         xlab = "",
                                         bins = 30
                                       )
                                     })

marrangeGrob(
  ncol = 2, nrow = 3,
  grobs = feature_distribution_plots
)
```

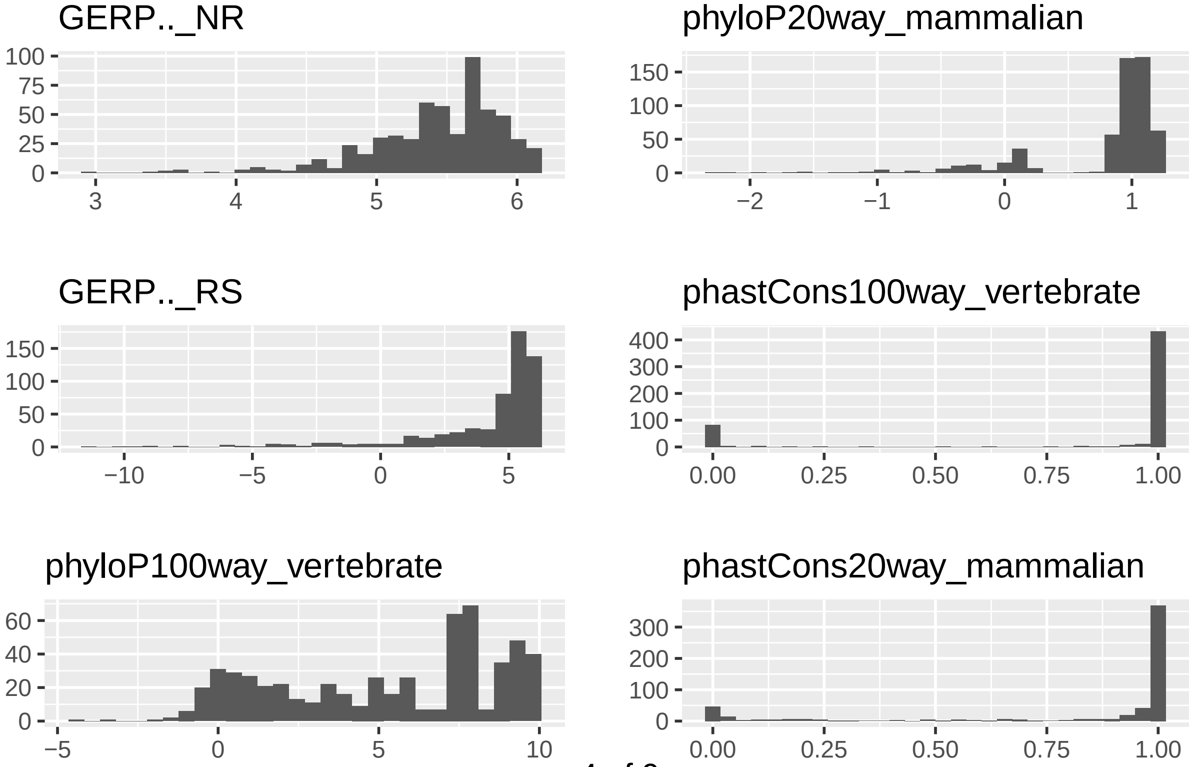
page 1 of 9



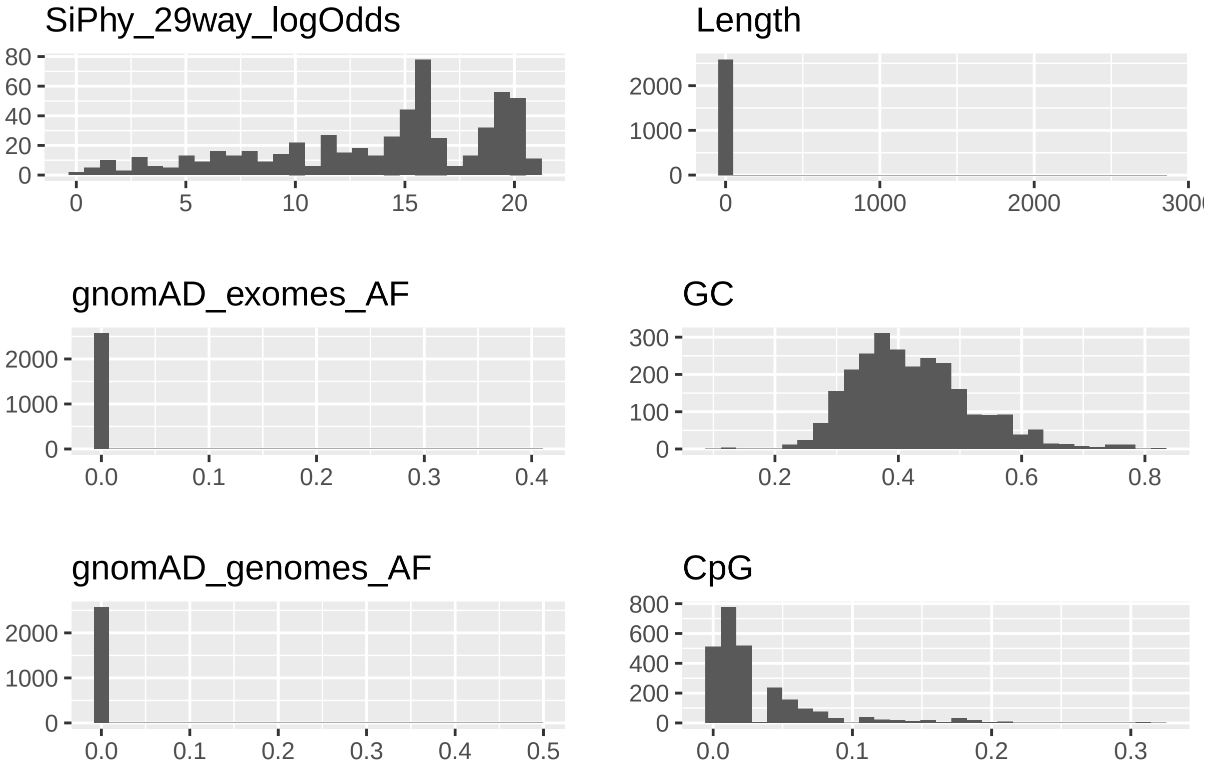
page 2 of 9



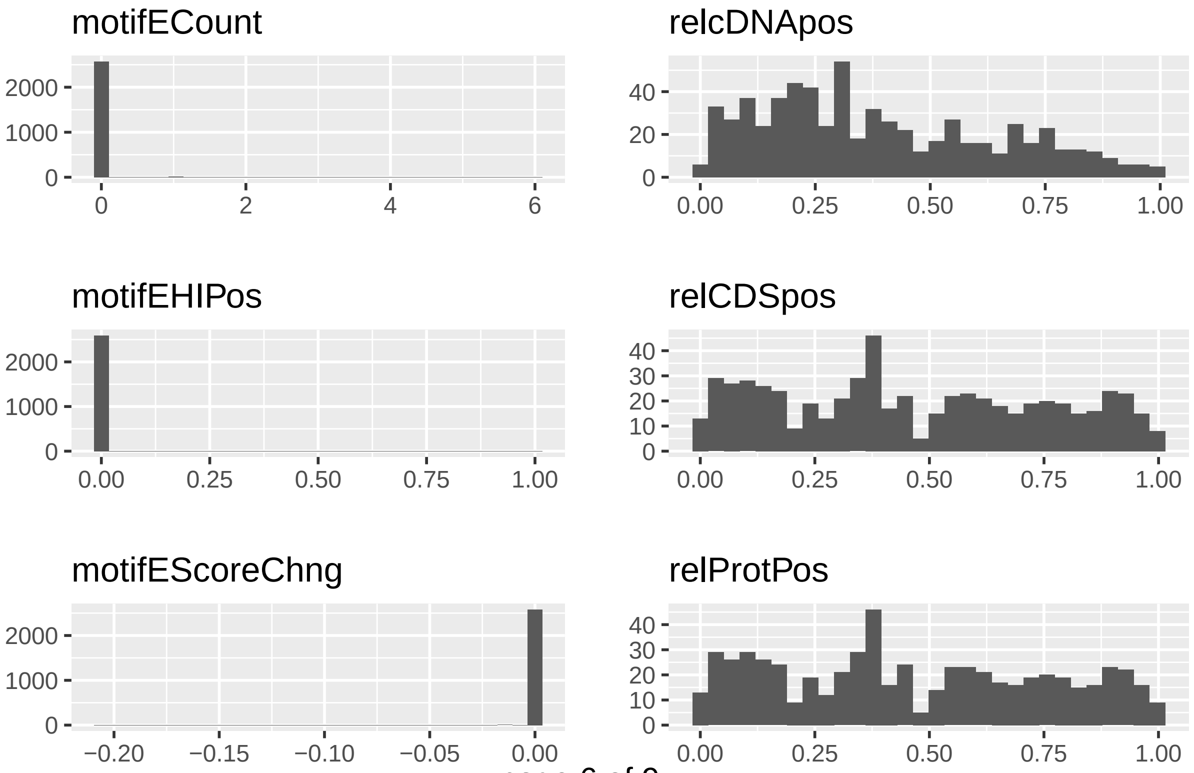
page 3 of 9



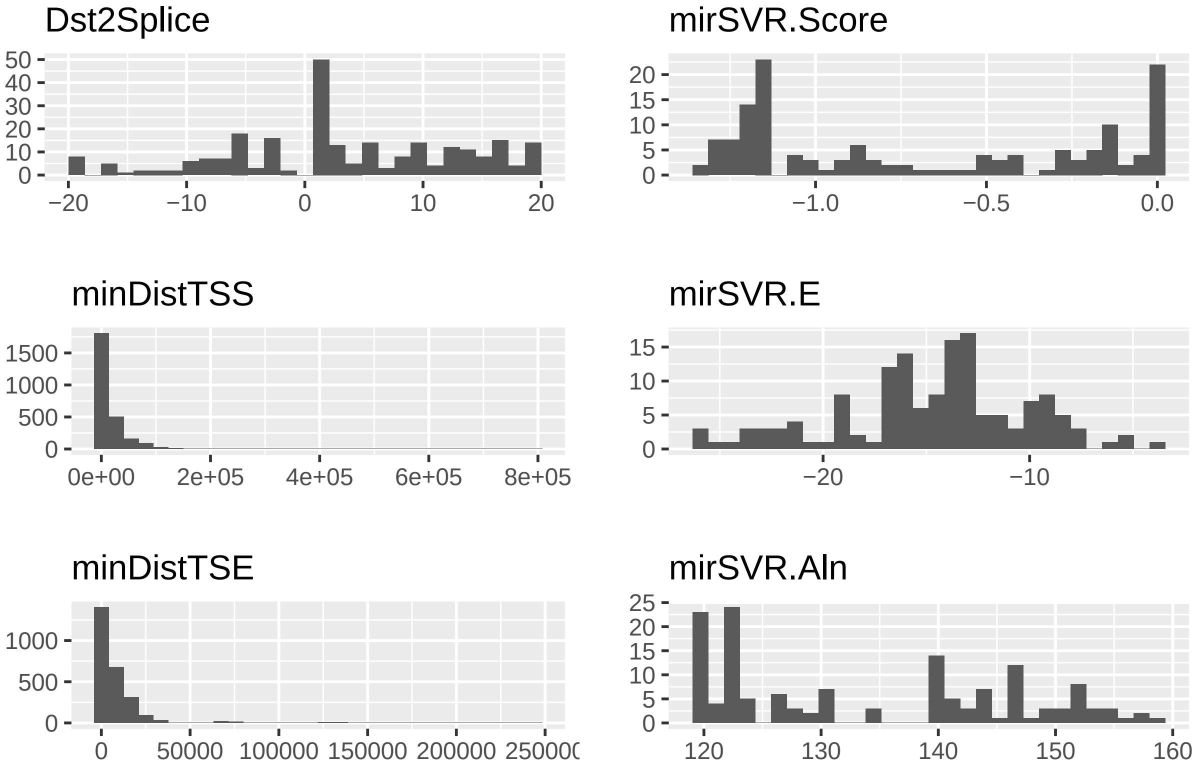
page 4 of 9



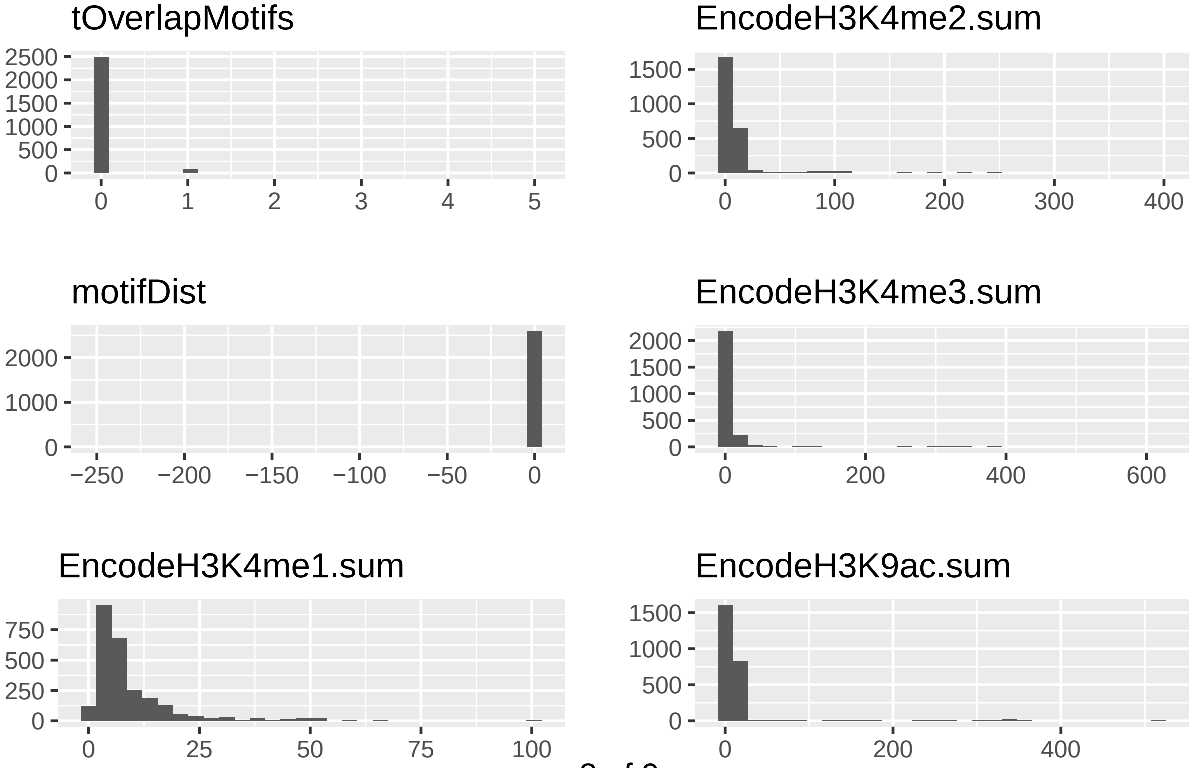
page 5 of 9



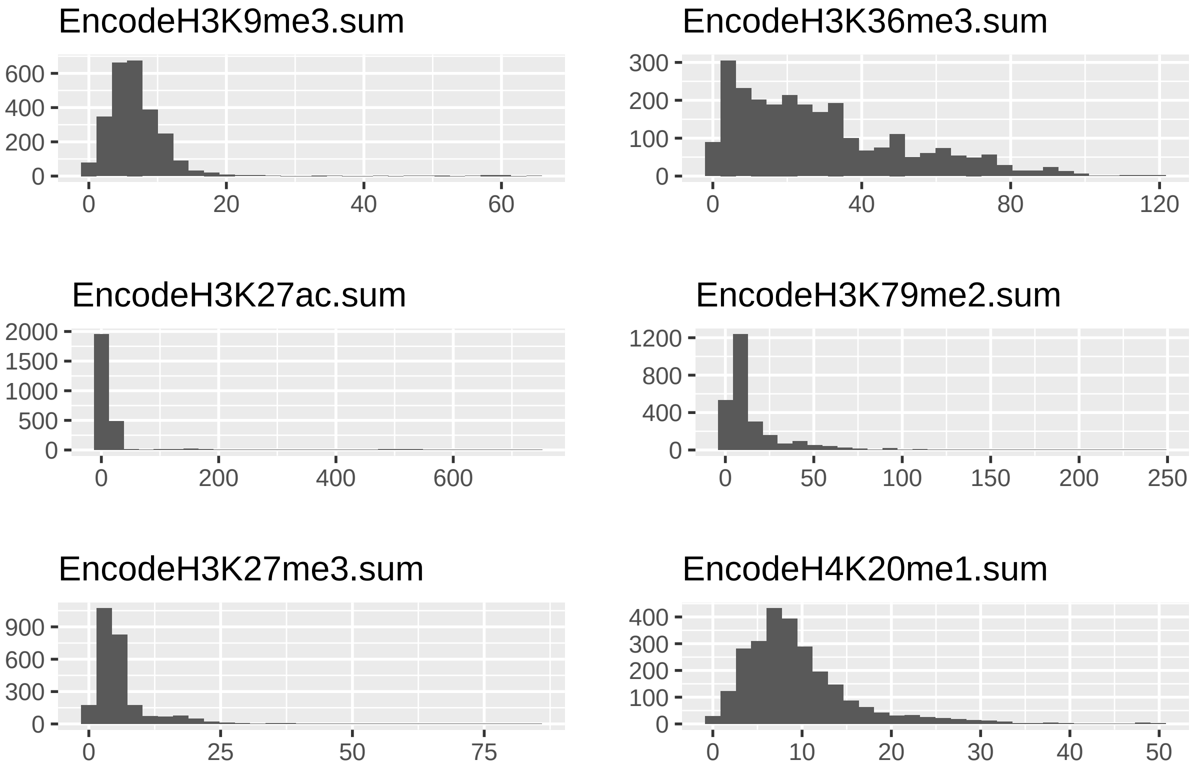
page 6 of 9

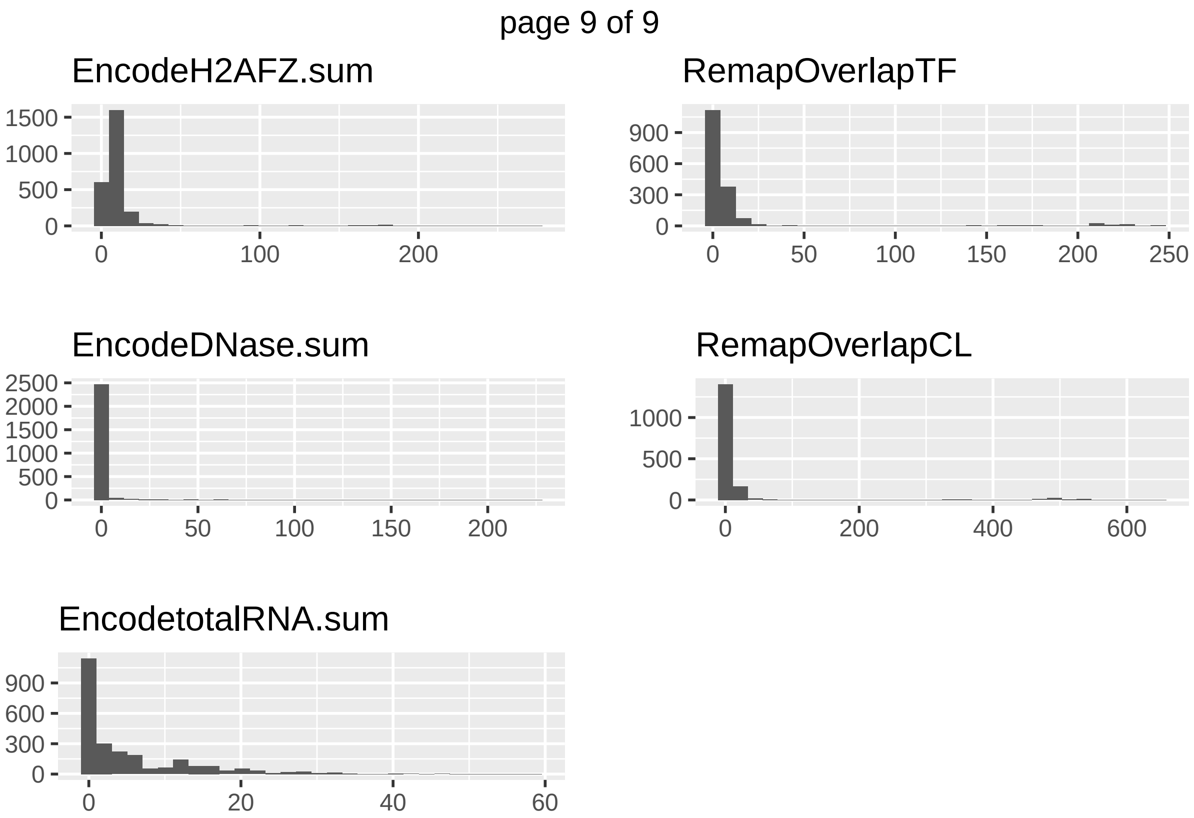


page 7 of 9



page 8 of 9





They are not, and thus it might be worth considering data transformations. In the case of random forest, however, monotone transformations should have no effect.

Categorical level occurence counts

Print (one-dimensional) contingency tables, i.e. occurence counts of each level of categorical variables.

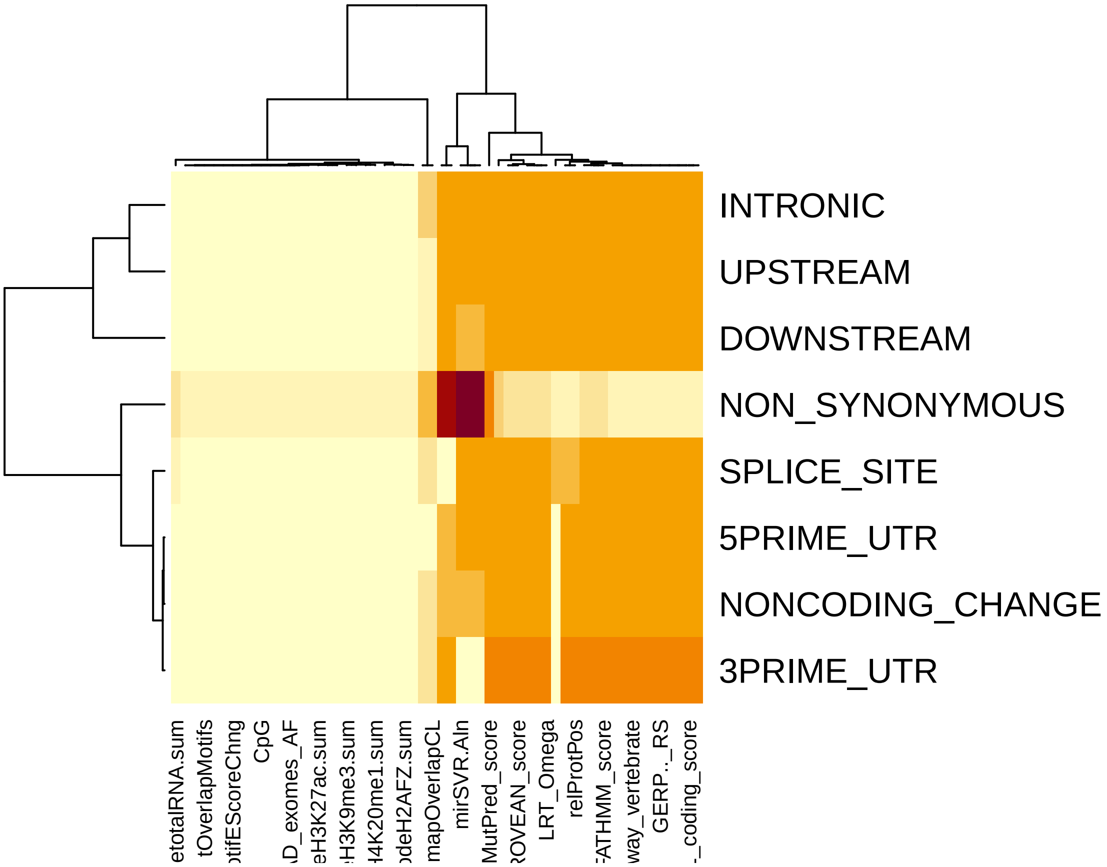
```
for (cat_feat in categorical_features) {
  table(training_set[, cat_feat, drop = FALSE], dnn = cat_feat, useNA = "always") %>% as.data.frame %>%
}
```

```
##   LRT_pred Freq
## 1        D  382
## 2        N  162
## 3        U    4
## 4     <NA> 2044
##   Dst2SplType Freq
## 1    ACCEPTOR  115
## 2       DONOR  139
## 3        <NA> 2338
##      Consequence.x Freq
## 1       3PRIME_UTR   22
## 2       5PRIME_UTR   17
## 3       DOWNSTREAM  458
## 4         INTRONIC  794
## 5   NON_SYNONYMOUS  580
## 6 NONCODING_CHANGE   13
## 7      SPLICE_SITE   73
## 8         UPSTREAM  635
## 9             <NA>    0
```

Heatmap of feature missingness against consequence

It is likely that missing values are more or less common in some variables depending on the predicted consequence. This can be visualized by a heatmap:

```
missing_value_sum_per_consequence <- lapply(training_set[, c(numeric_features, categorical_features), d
                                             function(column) {
                                               sapply(
                                                 split(is.na(column), training_set$Consequence.x),
                                                 sum
                                               )
                                             })
missing_value_sum_per_consequence %<>% data.frame %>% as.matrix
heatmap(missing_value_sum_per_consequence)
```



Non-synonymous variants have much less missingness in certain variables and more in others (as expected).

Compute number of observed missingness patterns

```
missingness_patterns <- training_set[, c(numeric_features, categorical_features)] %>% is.na
unique_missingness_patterns <- missingness_patterns %>% unique
num_missingness_patterns <- unique_missingness_patterns %>% nrow
print(paste(num_missingness_patterns, "out of", 2^length(c(numeric_features, categorical_features)), "p
```

```
## [1] "118 out of 72057594037927936 possible missingness patterns."
```

```r
missingness_pattern_factor <- apply(missingness_patterns, MARGIN = 1, function(x) paste0(as.integer(x),
rows_per_missingness_pattern <- table(missingness_pattern_factor)
rows_per_missingness_pattern <- rows_per_missingness_pattern %>% as.data.frame
rows_per_missingness_pattern[order(rows_per_missingness_pattern$Freq, decreasing = TRUE),]
```

```
##                                    missingness_pattern_factor Freq
## 85  11111111111111111110000000011110011100000000000000000110 1042
## 86  11111111111111111110000000011110011100000000000000011110  629
## 7   00000000000000000000000000000100111000000000000000000010  151
## 18  00000000100000000000000000000100111000000000000000000010   86
## 81  11111111111111111110000000011110000000000000000000000110   65
## 3   00000000000000000000000000000000111000000000000000000000   54
## 74  11111111111111111110000000011100011100000000000000000100   44
## 87  11111111111111111110000000011110011100000000000000100110   32
## 8   00000000000000000000000000000100111000000000000000011010   28
## 32  00000001100000000000000000000100111000000000000000000010   24
## 12  00000000100000000000000000000000111000000000000000000000   21
## 19  00000000100000000000000000000100111000000000000000011010   21
## 4   00000000000000000000000000000000111000000000000000011000   19
## 75  11111111111111111110000000011100011100000000000000011100   19
## 88  11111111111111111110000000011110011100000000000000111110   19
## 71  11111111111111111110000000001110011100000000000000000110   18
## 6   00000000000000000000000000000100000000000000000000000010   16
## 62  11111111111111111110000000000000011100000000000000000100   14
## 33  00000001100000000000000000000100111000000000000000011010   12
## 68  11111111111111111110000000001110000000000000000000000110   12
## 96  11111111111111111110000000011110011100000000100000011110   12
## 17  00000000100000000000000000000100000000000000000000000010   11
## 52  10000010100000000000000000000100111000000000000000000010   11
## 69  11111111111111111110000000001110000000000000000000011110    9
## 82  11111111111111111110000000011110000000000000000000011110    9
## 9   00000000000000000000000000000100111000000000000000100010    8
## 45  01100000000000000000000000000100111000000000000000000110    8
## 46  01100000000000000000000000000100111000000000000000011110    7
## 77  11111111111111111110000000011100011100000000000000111100    7
## 25  00000001000000000000000000000000111000000000000000000000    6
## 48  01100001000000000000000000000100111000000000000000000110    6
## 50  10000010100000000000000000000000111000000000000000000000    6
## 63  11111111111111111110000000000000011100000000000000011100    6
## 76  11111111111111111110000000011100011100000000000000100100    6
## 26  00000001000000000000000000000000111000000000000000011000    5
## 56  10010110000000000000000000000100111000000000000000000010    5
## 72  11111111111111111110000000001110011100000000000000011110    5
## 89  11111111111111111110000000011110011100000000000010011110    5
## 13  00000000100000000000000000000000111000000000000000011000    4
## 39  00001000100000000000000000000100111000000000000000000010    4
## 57  10010110000000000000000000000100111000000000000000011010    4
## 66  11111111111111111110000000001100011100000000000000000100    4
## 92  11111111111111111110000000011110011100000000001000000110    4
## 106 11111111111111111110000000011110011100100000000000011110    4
## 5   00000000000000000000000000000000111000000000000000100000    3
## 11  00000000100000000000000000000000000000000000000000000000    3
## 28  00000001100000000000000000000000111000000000000000000000    3
## 38  00001000000000000000000000000100111000000000000000000010    3
## 41  00001001100000000000000000000010011100000000000000000010      3
## 51  10000010100000000000000000000010000000000000000000000010      3
## 53  10000011100000000000000000000010011100000000000000000010      3
## 93  11111111111111111110000000111100111000000000001000011110      3
## 99  11111111111111111110000000111100111000000100000000011110      3
## 110 11111111111111111110000000111100111001000001111100011110      3
## 1   00000000000000000000000000000000000000000000000000000000      2
## 2   00000000000000000000000000000000000000000000000000011000      2
## 20  00000000100000000000000000000010011100000000000000100010      2
## 23  00000001000000000000000000000000000000000000000000000000      2
## 35  00000001100000000000000000000010011100010000000000011010      2
## 42  01100000000000000000000000000000011100000000000000000100      2
## 43  01100000000000000000000000000000011100000000000000011100      2
## 44  01100000000000000000000000000000011100000000001000000100      2
## 58  10010110100000000000000000000000011100000000000000011000      2
## 61  11111111111111111110000000000000000000000000000000000100      2
## 64  11111111111111111110000000000010011100000000000000000110      2
## 90  11111111111111111110000000111100111000000000000100000110      2
## 105 11111111111111111110000000111100111000110000000000011110      2
## 113 11111111111111111110000000111100111001000101110100011110      2
## 114 11111111111111111110000000111100111001100000000000011110      2
## 115 11111111111111111110000000111100111001100000001000011110      2
## 10  00000000000000000000000000000010011100000000000000111010      1
## 14  00000000100000000000000000000000011100000000000000100000      1
## 15  00000000100000000000000000000000011100000000000000111000      1
## 16  00000000100000000000000000000000011100111111111110011000      1
## 21  00000000100000000000000000000010011100000000000000111010      1
## 22  00000000100000000000000000000010011100000000000100011010      1
## 24  00000001000000000000000000000000000000000000000000011000      1
## 27  00000001000000000000000000000010000000000000000000000010      1
## 29  00000001100000000000000000000000011100000000000000011000      1
## 30  00000001100000000000000000000000011100000000000000100000      1
## 31  00000001100000000000000000000010000000000000000000000010      1
## 34  00000001100000000000000000000010011100000000000100011010      1
## 36  00001000000000000000000000000000011100000000000000011000      1
## 37  00001000000000000000000000000010000000000000000000000010      1
## 40  00001001000000000000000000000010011100000000000000000010      1
## 47  01100000000000000000000000000010011100000000001000000110      1
## 49  01100001100000000000000000000010011100000000000000000110      1
## 54  10001010100000000000000000000000011100000000000000000000      1
## 55  10010110000000000000000000000000011100000000000000011000      1
## 59  10010110100000000000000000000010011100000000000000000010      1
## 60  10011110100000000000000000000010011100000000000000000010      1
## 65  11111111111111111110000000001100000000000000000000000100      1
## 67  11111111111111111110000000001100011100000000000000011100      1
## 70  11111111111111111110000000001110000000010000000010011110      1
## 73  11111111111111111110000000001110011100000000000000100110      1
## 78  11111111111111111110000000111000111000000000001000000100      1
## 79  11111111111111111110000000111000111001000000100000011100      1
## 80  11111111111111111110000000111000111001100010001000011100      1
## 83  11111111111111111110000000111100000000010000000010011110      1
## 84  11111111111111111110000000111100000000110001000010011110      1
## 91  11111111111111111110000000111100111000000000000100011110      1
## 94  11111111111111111110000000111100111000000000001100011110      1
## 95  11111111111111111110000000011110011100000000001110011110     1
## 97  11111111111111111110000000011110011100000000100010011110     1
## 98  11111111111111111110000000011110011100000001001010011110     1
## 100 11111111111111111110000000011110011100000100000000011110     1
## 101 11111111111111111110000000011110011100000100111010011110     1
## 102 11111111111111111110000000011110011100001000000000011110     1
## 103 11111111111111111110000000011110011100001000001000011110     1
## 104 11111111111111111110000000011110011100010010000000011110     1
## 107 11111111111111111110000000011110011100100000001000011110     1
## 108 11111111111111111110000000011110011100100000011100011110     1
## 109 11111111111111111110000000011110011100100000011110011110     1
## 111 11111111111111111110000000011110011100100001001100111110     1
## 112 11111111111111111110000000011110011100100010011010011110     1
## 116 11111111111111111110000000011110011100110000100000011110     1
## 117 11111111111111111110000000011110011100110001101110011110     1
## 118 11111111111111111110000000011110011100111111101111011110     1
```