FUSION

(Family-level Unique Small RNA Integration)

Version 1.0.2

User Manual

By:

Hukam C. Rawal, Qi Chen, Tong Zhou

October, 2025

Table of Contents

1. Introduction

FUSION (Family-level Unique Small RNA Integration) is a computational tool for detecting sncRNA family-level expression across samples from expression matrix of unique sncRNA species using R packages. First, it quantifies unique sncRNA species and then aggregates them into their respective parental RNA families. This family-level integration captures the contributions of individual sncRNA species while enhancing statistical power and robustness for differential expression analysis. Two modules, *FUSION_ms* and *FUSION_ps*, are proposed in the *FUSION* framework. By effectively reducing noise and amplifying collective signals from unique sncRNA species, *FUSION_ms* enables reliable detection of sncRNA family-level expression changes even when sample size is limited. In comparison, *FUSION_ps* is powered by paired-sample analysis, which enables “1-on-1” sncRNA differential expression analysis and is optimized for single-case research setting as well.

2. Installation and usage

FUSION is a package written for the R computing environment; therefore, first install R and Rstudio (https://rstudio.com).

2.1. Pre-requisites:

Rtools, R packages (read.delim (from utilis), use_package from usethis), lm, p.adjust and wilcox.test (from stats))

2.2. Installation:

To install it directly from GitHub:

```
remotes::install_github("cozyrna/FUSION")
```

Or, using the downloaded R package:

Download FUSION_*.tar.gz file

Open Rstudio or R and type as below:

```
install.packages("~/FUSION_*.tar.gz", repos = NULL, type = "source")
```

2.3. Running FUSION:

After installing the package FUSION, call the library as:

```
library(FUSION)
```

There are three functions in this package :

FUSION_ps - Differential expression analysis of sncRNA families in paired data samples using expression matrix

FUSION_ms - Differential expression analysis of sncRNA families in multiple samples data using expression matrix

FUSION_msmc - Differential expression analysis of sncRNA families in multiple samples data with multiple conditions using expression matrix

Ask help for the description and help menus of each of these functions :

```
?FUSION_ps
?FUSION_ms
?FUSION_msmc
```

2.3.1. FUSION_ps:

Differential expression analysis of sncRNA families in paired data samples using expression matrix. It will return, for each pair in the input matrix, an output data-frame with w_positive, w_negative, *P*-value, and adjusted *P*-value for each sncRNA family chosen for the analysis.

Inputs:

a: A matrix file where the first column is "Sequence" (or ID) and the second column is "Annotation". The Sequence (or ID) must be unique. The rest of the columns contain RPM or expression values from different samples under study, such as first half of the columns corresponds to the samples from Condition1 (e.g., control or healthy tissue) and the second half of the columns correspond to the samples from Condition2 (e.g., treated or infected tissue)

There are multiple example files available in the folder **"/home/..../R/.../FUSION/extdata/"** that can be referenced for proper formatting and structure of the input matrix (**a**):

example_matrix_p1.txt : This matrix file comprises 5 pairs of samples. The first 5 columns (Healthy_1, Healthy_2, Healthy_3, Healthy_4, Healthy_5) represent the expression values for samples from Condition1 (e.g., Healthy), and the last 5 columns (Infected_1, Infected_2, Infected_3, Infected_4, Infected_5) represent the expression values for samples from Condition2 (e.g., Infected).

example_matrix_p2.txt : This matrix file comprises 10 pairs of samples. The first 10 columns (Control_1, Control_2, …, Control_10) represent the expression values for samples from Condition1 (e.g., Control), and the last

10 columns (Treated_1, Treated_2, …, Treated_10) represent the expression values for samples from Condition2 (e.g., Treated).

example_matrix_p3.txt : This matrix file comprises 4 pairs of samples. The first 4 columns (Sample_1, Sample_2, Sample_3, Sample_4) represent the expression values for samples from Condition1, and the last 4 columns (Sample_5, Sample_6, Sample_7, Sample_8) represent the expression values for samples from Condition2.

example_matrix_p4.txt : This matrix file comprises 5 pairs of samples, with the samples are arranged in pairs of columns. The first pair consists of Healthy_1 and Infected_1, followed by the second pair Healthy_2 and Infected_2, and so on, with each subsequent pair representing a sample from Condition1 (e.g., Healthy) and its corresponding sample from Condition2 (e.g., Infected).

order: Use either G or P to specify the order of paired samples in the input matrix.

G: Samples are in Group order (i.e., all samples from Condition1 followed by all samples from Condition2). Refer to the file 'example_matrix_p1.txt' for this format.

P: Samples are in Pairs order (where each pair consists of one sample from Condition1 and one from Condition2). Refer to the file 'example_matrix_p4.txt' for this format.

By default, the order is considered as Group (G).

row_mean: This parameter specifies the mean RPM (default value is **0.1**) threshold used to retain the sncRNA species (rows) in the matrix. Rows with a mean RPM value below the specified threshold will be excluded from the analysis.

sncrna_family: This parameter specifies the list of sncRNA families to be analyzed in the expression analysis study. Use the following options:

- **"tsrna"** for **tsRNAs** (transfer RNA-derived small RNAs),
- **"rsrna"** for **rsRNAs** (ribosomal RNA-derived small RNAs),
- **"ysrna"** for **ysRNAs** (Y RNA-derived small RNAs),
- **"mirna"** for **miRNAs** (microRNAs),
- **"other"** for a combination of **pRNA**, **snRNA**, and **snoRNA**.

For all families, you can use any letter or number, e.g., "a","b", "c", 1, 2, 3. By default (i.e., if no option is specified), it will search and analyze for tryRNAs (tsRNAs, rsRNAs, and ysRNAs).

padj_method: This parameter specifies the adjustment method for correcting *P*-values. You can choose from the following options:

- "bonferroni" for the the Bonferroni correction method,
- "BH" for the Benjamini & Hochberg method.

By default (i.e., if no option isspecified), the Bonferroni correction will applied.

Note: If you want to save the terminal/console output to a file, use sink() command.

e.g., options(max.print = 1e6);
sink("~/output.txt");
FUSION_ps(a = "./extdata/example_matrix_p1.txt"); sink()

unique_anno: This parameter specifies whether to filter (TRUE) or not (FALSE) the input matrix to consider only uniquely mapped reads (sncRNA species), meaning only the sncRNA species that map to a single parent RNA. By default, it is FALSE (i.e., all relevant sncRNA species in the matrix are considered, irrespective of whether they map to a single or multiple parent RNAs).

Example runs:

Note: After installation, one can find the example files in "../FUSION/extdata/".

If want to run examples straight as in the Help documentation, it is necessary to first set working directory to the base folder of the installed package FUSION. Such as : setwd("/home/..../R/.../FUSION/").

Or, simply provide the exact path of the appropriate example matrix files.

To run on your own matrix file, provide the full path as :

FUSION(a = "/path/to/your_matrix.txt")

Example 1:

FUSION_ps(a = "./extdata/example_matrix_p1.txt")

Run differential expression analysis on example_matrix_p1.txt (5 pairs of samples) at default row_mean threshold (i.e., 0.1) for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default).

Example 2:

FUSION_ps(a = "./extdata/example_matrix_p1.txt", padj_method = "BH")

Run differential expression analysis on example_matrix_p1.txt (5 pairs of samples) at default row_mean threshold (i.e., 0.1) for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default) using BH (Benjamini & Hochberg) method for correcting or adjusting p-values.

Example 3:

FUSION_ps(a = "./extdata/example_matrix_p1.txt", row_mean = 0.5)

Run differential expression analysis on example_matrix_p1.txt (5 pairs of samples) at row_mean threshold of 0.5 for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default).

Example 4:

```
FUSION_ps(a = "./extdata/example_matrix_p2.txt", sncrna_family = "a")
```

Run differential expression analysis on example_matrix_p2.txt (10 pairs of samples) at default row_mean threshold (i.e., 0.1) for all sncRNA families.

Example 5:

```
FUSION_ps(a = "./extdata/example_matrix_p2.txt", sncrna_family = 0)
```

Run differential expression analysis on example_matrix_p2.txt (10 pairs of samples) at default row_mean threshold (i.e., 0.1) for all sncRNA families;

Example 6:

```
FUSION_ps(a = "./extdata/example_matrix_p2.txt", sncrna_family = "mirna")
```

Run differential expression analysis on example_matrix_p2.txt (10 pairs of samples) at default row_mean threshold (i.e., 0.1) for miRNA families.

Example 7:

```
FUSION_ps(a = "./extdata/example_matrix_p3.txt", row_mean = 0.5, sncrna_family = "tsrna")
```

Run differential expression analysis on example_matrix_p2.txt (4 pairs of samples) at row_mean threshold of 0.5 for tsRNA families.

Example 8:

```
FUSION_ps(a = "./extdata/example_matrix_p3.txt", row_mean = 0.1, sncrna_family = "ysrna")
```

Run differential expression analysis on example_matrix_p2.txt (4 pairs of samples) at row_mean threshold of 0.1 for for ysRNA families.

Example 9:

```
FUSION_ps(a = "./extdata/example_matrix_p3.txt", sncrna_family = "other")
```

Run differential expression analysis on example_matrix_p2.txt (4 pairs of samples) at default row_mean threshold (i.e., 0.1) for other (pRNA,snRNA and snoRNA) sncRNA families.

Example 10:

FUSION_ps(a = "./extdata/example_matrix_p4.txt", order = "P", unique_anno = TRUE)

Run differential expression analysis on example_matrix_p4.txt (5 pairs of samples) considering only uniquely mapped reads (sncRNA species that map to a single parent RNA) at default row_mean threshold (i.e., 0.1) for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default) and samples are arranged in as pairs of columns.

2.3.2. FUSION_ms:

Differential expression analysis of sncRNA families in multiple samples data using expression matrix. It will return a final output in a data-frame with t-statistics, *P*-value, and adjusted *P*-value for each sncRNA family chosen for analysis.

Inputs:

a: A matrix file where the first column is "Sequence" (or ID) and the second column is "Annotation". The Sequence (or ID) must be unique. The rest of the columns contain RPM or expression values from different samples under study, such as first set of the columns (S1) corresponds to the samples from Condition1 (e.g., control or healthy tissue) and the second set of the columns (S2) correspond to the samples from Condition2 (e.g., treated or infected tissue)

S1: Number of samples from Condition1 (e.g., control or healthy tissue)

S2: Number of samples from Condition2 (e.g., treated or infected tissue)

There are multiple example files available in the folder **"/home/..../R/.../FUSION/extdata/"** that can be referenced for proper formatting and structure of the input matrix (**a**):

example_matrix1.txt : This matrix file contains a total of 26 samples, comprising 10 (S1) samples (Sample_1, Sample_2, Sample_3, …, Sample_10) from Condition1, and 16 (S2) samples (Sample_17, Sample_18, Sample_19, …, Sample_26) from Condition2

example_matrix2.txt : a matrix file with total 10 samples comprising 5 (S1) samples (Healthy_1, Healthy_2, Healthy_3, Healthy_4, Healthy_5) from Condition1 and 5 (S2) samples (Infected_1, Infected_2, Infected_3, Infected_4, Infected_5) from Condition2

example_matrix3.txt : a matrix file with total 18 samples comprising 10 (S1) samples (Control_1, Control_2, Control_3, …, Control_10) from Condition1 and 8 (S2) samples (Treated_1, Treated_2, Treated_3, …, Treated_8) from Condition2

row_mean: This parameter specifies the mean RPM (default value is **0.1**) threshold used to retain the sncRNA species (rows) in the matrix. Rows with a mean RPM value below the specified threshold will be excluded from the analysis.

sncrna_family: This parameter specifies the list of sncRNA families to be analyzed in the expression analysis study. Use the following options:

- **"tsrna"** for **tsRNAs** (transfer RNA-derived small RNAs),
- **"rsrna"** for **rsRNAs** (ribosomal RNA-derived small RNAs),
- **"ysrna"** for **ysRNAs** (Y RNA-derived small RNAs),
- **"mirna"** for **miRNAs** (microRNAs),
- **"other"** for a combination of **pRNA**, **snRNA**, and **snoRNA**.

For all families, you can use any letter or number, e.g., "a","b", "c", 1, 2, 3. By default (i.e., if no option is specified), it will search and analyze for tryRNAs (tsRNAs, rsRNAs, and ysRNAs).

padj_method: This parameter specifies the adjustment method for correcting *P*-values. You can choose from the following options:

- "bonferroni" for the the Bonferroni correction method,
- "BH" for the Benjamini & Hochberg method.

By default (i.e., if no option isspecified), the Bonferroni correction will applied.

top_species: This parameter specifies the number (default is 1000) of top species for each sncRNA family to be considered for analysis. It is useful for reducing the runtime of the analysis, especially for families (like rsrna families) with a large number of species. If time is not a concern, higher values such as 5000, 10000, etc., can be used.

unique_anno: This parameter specifies whether to filter (TRUE) or not (FALSE) the input matrix to consider only uniquely mapped reads (sncRNA species), meaning only the sncRNA species that map to a single parent RNA. By default, it is FALSE (i.e., all relevant sncRNA species in the matrix are considered, irrespective of whether they map to a single or multiple parent RNAs).

Example runs:

Note: After installation, one can find the example files in "../FUSION/extdata/".

If want to run examples straight as in the Help documentation, it is necessary to first set working directory to the base folder of the installed package FUSION. Such as : setwd("/home/..../R/.../FUSION/").

Or, simply provide the exact path of the appropriate example matrix files.

To run on your own matrix file, provide the full path as :
FUSION(a = "/path/to/your_matrix.txt")

Example 1:
FUSION_ms(a = "./extdata/example_matrix1.txt", S1 = 10, S2 = 16, row_mean = 1, top_species = 5000)
Run differential expression analysis on example_matrix1.txt with 10 healthy samples (S1) and 16 patients (S2) at row_mean threshold of 1 for 5000 top_species for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default).

Example 2:
FUSION_ms(a = "./extdata/example_matrix1.txt", S1 = 10, S2 = 16)
Run differential expression analysis on example_matrix1.txt with 10 samples from Condition1 (S1) and 16 samples from Condition2 (S2) at default row_mean (i.e., 0.1) and top_species (i.e., 1000) threshold for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default).

Example 3:
FUSION_ms(a = "./extdata/example_matrix1.txt", S1 = 10, S2 = 16, padj_method = "BH")
Run differential expression analysis on example_matrix1.txt with 10 samples from Condition1 (S1) and 16 samples from Condition2 (S2) at default row_mean (i.e., 0.1) and top_species (i.e., 1000) threshold for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default) using BH (Benjamini & Hochberg) method for correcting or adjusting p-values.

Example 4:
FUSION_ms(a = "./extdata/example_matrix1.txt", S1 = 10, S2 = 16, sncrna_family = "a")

Run differential expression analysis on example_matrix1.txt with 10 samples from Condition1 (S1) and 16 samples from Condition2 (S2) at default row_mean threshold (i.e., 0.1) and top_species (i.e., 1000) for all sncRNA families.

Example 5:

FUSION_ms(a = "./extdata/example_matrix2.txt", S1 = 5, S2 = 5, sncrna_family = 0)

Run differential expression analysis on example_matrix2.txt with 5 healthy samples (S1) and 5 patients (S2) at default row_mean threshold (i.e., 0.1) and top_species (i.e., 1000) for all sncRNA families;

Example 6:

FUSION_ms(a = "./extdata/example_matrix2.txt", S1 = 5, S2 = 5, sncrna_family = "mirna", top_species = 1000)

Run differential expression analysis on example_matrix2.txt with 5 samples from Condition1 (S1) and 5 samples from Condition2 (S2) at default row_mean threshold (i.e., 0.1) for 1000 top_species for miRNA families.

Example 7:

FUSION_ms(a = "./extdata/example_matrix2.txt", S1 = 5, S2 = 5, row_mean = 10, top_species = 2000, sncrna_family = "rsrna")

Run differential expression analysis on example_matrix2.txt with 5 samples from Condition1 (S1) and 5 samples from Condition2 (S2) at row_mean threshold of 10 for 2000 top_species for rsRNA families.

Example 8:

FUSION_ms(a = "./extdata/example_matrix3.txt", S1 = 10, S2 = 8, row_mean = 10, sncrna_family = "tsrna")

Run differential expression analysis on example_matrix3.txt with 10 control samples (S1) and 8 treated samples (S2) at row_mean threshold of 10 for default (i.e., 1000) top_species for tsRNA families.

Example 9:

```
FUSION_ms(a = "./extdata/example_matrix3.txt", S1 = 10, S2 = 8, row_mean = 0.1, sncrna_family = "ysrna")
```

Run differential expression analysis on example_matrix3.txt with 10 samples from Condition1 (S1) and 8 samples from Condition2 (S2) at row_mean threshold of 0.1 for default (i.e., 1000) top_species for ysRNA families.

Example 10:

```
FUSION_ms(a = "./extdata/example_matrix3.txt", S1 = 10, S2 = 8, top_species = 100, sncrna_family = "other", unique_anno = TRUE)
```

Run differential expression analysis on example_matrix3.txt with 10 samples from Condition1 (S1) and 8 samples from Condition2 (S2) considering only uniquely mapped reads (sncRNA species that map to a single parent RNA) at default row_mean threshold (i.e., 0.1) for 100 top_species for other (pRNA, snRNA and snoRNA) sncRNA families.

2.3.3. FUSION_msmc:

Differential expression analysis of sncRNA families in multiple samples data with multiple conditions using expression matrix. It will return a final output in a data-frame with t-statistics, *P*-value, and adjusted *P*-value for each sncRNA family chosen for the analysis.

Inputs:

a: A matrix file where the first column is "Sequence" (or ID) and the second column is "Annotation". The Sequence (or ID) must be unique. The rest of

the columns contain RPM or expression values from different samples under study.

cl: a file that contains multiple sample conditions in a comma-separated format. For example, if there are 3 different conditions with 6 samples each, the file would contains the input: 1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3.
Each number represents a sample's corresponding condition.

There are multiple example files available in the folder **"/home/..../R/.../FUSION/extdata/"** that can be referenced for proper formatting and structure of the input matrix (**a**):

Example files for running FUSION_msmc: There are examples files in the folder: **"/home/..../R/.../FUSION/extdata/"** that can be referenced for proper formatting and structure of the input matrix (**a**) and condition specifying file (cl):

example_matrix_cl.txt : a matrix file with total 18 samples

example_condition1.txt: a file specifying three conditions as 1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3 i.e., a comma separated format showing 3 different conditions with 6 samples each

example_condition2.txt: a file specifying four conditions as 1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4 i.e., a comma separated format showing 4 different conditions with condition 1, 2, and 3 having five samples each, while last three samples are representing the condition 4.

Note: one can specify the condition in any order depending on the order of the samples in the matrix, such as : "2,2,2,1,1,1,3,3,3" or "1,2,1,2,1,2,1,2,1,2" or "1,1,1,2,2,2,1,1,1,3,3,3"

row_mean: This parameter specifies the mean RPM (default value is **0.1**) threshold used to retain the sncRNA species (rows) in the matrix. Rows with a mean RPM value below the specified threshold will be excluded from the analysis.

sncrna_family: This parameter specifies the list of sncRNA families to be analyzed in the expression analysis study. Use the following options:

- **"tsrna"** for **tsRNAs** (transfer RNA-derived small RNAs),
- **"rsrna"** for **rsRNAs** (ribosomal RNA-derived small RNAs),
- **"ysrna"** for **ysRNAs** (Y RNA-derived small RNAs),
- **"mirna"** for **miRNAs** (microRNAs),
- **"other"** for a combination of **pRNA**, **snRNA**, and **snoRNA**.

For all families, you can use any letter or number, e.g., "a","b", "c", 1, 2, 3. By default (i.e., if no option is specified), it will search and analyze for tryRNAs (tsRNAs, rsRNAs, and ysRNAs).

padj_method: This parameter specifies the adjustment method for correcting *P*-values. You can choose from the following options:

- "bonferroni" for the the Bonferroni correction method,
- "BH" for the Benjamini & Hochberg method.

By default (i.e., if no option isspecified), the Bonferroni correction will applied.

top_species: This parameter specifies the number (default is 1000) of top species for each sncRNA family to be considered for analysis. It is useful for reducing the runtime of the analysis, especially for families (like rsrna families) with a large number of species. If time is not a concern, higher values such as 5000, 10000, etc., can be used.

unique_anno: This parameter specifies whether to filter (TRUE) or not (FALSE) the input matrix to consider only uniquely mapped reads (sncRNA

species), meaning only the sncRNA species that map to a single parent RNA. By default, it is FALSE (i.e., all relevant sncRNA species in the matrix are considered, irrespective of whether they map to a single or multiple parent RNAs).

Example runs:

Note: After installation, one can find the example files in "../FUSION/extdata/".

If want to run examples straight as in the Help documentation, it is necessary to first set working directory to the base folder of the installed package FUSION. Such as : setwd("/home/..../R/.../FUSION/").

Or, simply provide the exact path of the appropriate example matrix files.

To run on your own matrix file, provide the full path as :

FUSION(a = "/path/to/your_matrix.txt")

Example 1:

FUSION_msmc(a = "./extdata/example_matrix_cl.txt", cl = "./extdata/example_condition1.txt", row_mean = 1, top_species = 5000)

Run differential expression analysis on example_matrix_cl.txt with 18 samples as per the conditions specified (i.e., 3 different conditions with 6 samples each) in example_condition1.txt at row_mean threshold of 1 for 5000 top_species for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default).

Example 2:

FUSION_msmc(a = "./extdata/example_matrix_cl.txt", cl = "./extdata/example_condition2.txt", row_mean = 1, top_species = 5000)

Run differential expression analysis on example_matrix_cl.txt with 18 samples as per the conditions specified (i.e., 4 different conditions with

condition 1, 2, and 3 having five samples each, while last three samples are representing the condition 4) in example_condition2.txt at row_mean threshold of 1 for 5000 top_species for sncRNA families (tsRNAs, rsRNAs and ysRNAs) (default).

Example 3 (This example is same as the Example 8 for FUSION_ms):

```
FUSION_msmc(a = "./extdata/example_matrix_cl.txt", cl = "./extdata/example_condition3.txt", row_mean = 10, sncrna_family = "tsrna")
## Run differential expression analysis on example_matrix_cl.txt with 18 samples as per the conditions specified (i.e., 2 different conditions with condition 1, and 2 having 10 and 8 samples, respectively in the example_condition3.txt at row_mean threshold of 10 for default (1000) top_species for tsRNA families.
```

3. Understanding output

Sample output files in the "extdata" folder:

- FUSION_ms_sample_output.txt
- FUSION_ps_sample_output.txt

FUSION_ms* and *FUSION_msmc:

The generated output file (**FUSION_ms_sample_output.txt**) contains four columns:

1. sncrna_family
2. t (*t*-statistic)
3. p (*P*-value)
4. adjusted_p (Adjusted *P*-value)

In this output:

- The magnitude of the differential expression of the given sncRNA family (sncrna_family) is represented by the *t*-statistic (t).

- The significance of this differential expression can be evaluated using the *P*-value (p) and the adjusted *P*-value (adjusted_p).

FUSION_ps:

The generated output file (**FUSION_ps_sample_output.txt**) contains six columns:

1. Pair (Paired sample identifiers, e.g., Pair_1 (for sample s_1 and s_2))
2. sncrna_family
3. w_pos (Positive-rank sum)
4. w_neg (Negative-rank sum
5. p (*P*-value)
6. adjusted_p (Adjusted *P*-value)

In this output:

- The significance of the differential expression of the given sncRNA family (sncrna_family) between the paired samples (*i.e.*, s_1 and s_2) (Pair_1) can be evaluated using the *P*-value (P) and the adjusted *P*-value (adjusted_p).
- The positive-rank sum (w_pos) and negative-rank sum (w_neg) help determine the direction of dysregulation between the two samples:
    - If w_pos > w_neg, the given sncRNA family is upregulated in s_1 compared to s_2
    - If w_pos < w_neg, the given sncRNA family is downregulated in s_1 compared to s_2.

Note: Outputs with all NAs for **w_pos**, **w_neg**, **p**, and **adjusted_p** (as shown below) indicate that the sncrna_family is present in the input matrix but **did not meet the** row_mean **threshold** (default = 0.1). These entries are therefore excluded from statistical testing.

Pair	sncrna_family	w_pos	w_neg	p	adjusted_p
Pair_1	mature-tRNA-SeC-TCA	NA	NA	NA	NA
Pair_1	mature-tRNA-Val-AAC	NA	NA	NA	NA

4. Possible errors and solutions

There is a possibility that users may encounter errors if the input matrix does not adhere to the specified format. Additionally, errors may arise even when there are no sncRNA species for the mentioned sncrna_family with the specified threshold (row_mean). Below, we discuss some common errors users may face, along with their potential reasons and solutions for correction.

(i) Warning in file(file, "rt") : cannot open file 'xxxxxxxxxxxxxxx': No such file or directory
Error in file(file, "rt") : cannot open the connection Called from: file(file, "rt")

Solution: Ensure that the correct file path is provided for the input matrix file. Check that the file exists at the specified location and that you have provided the correct path for the parameter “a”.

(ii) "Error: column numbers are not in pair in the matrix. Please check"

Solution: This error may occur when running FUSION_ps and indicates that a column is missing in the input matrix. Ensure that the matrix has the correct number of columns. The total number of columns must be even, and the first two columns should be reserved for Sequence (or ID) and Annotation. Double-check the matrix structure to make sure all required columns are present.

(iii) Warning: non-unique values when setting 'row.names': …… [... truncated]
Error in `.rowNamesDF<-`(x, value = value) :
duplicate 'row.names' are not allowed
Called from: `.rowNamesDF<-`(x, value = value)

Solution: This error occurs when the first column (Sequence or ID) contains duplicate

values. The Sequence (or ID) must be unique for each entry. Check the first column of your input matrix and remove any duplicate values to resolve this issue.

(iv) Error in rowMeans(e) : 'x' must be numeric
Called from: rowMeans(e)

Solution: This error occurs when there are non-numeric values in the columns containing expression values for the samples. To resolve this, you can either:

- Remove the rows that contain non-numeric values, or
- Replace the non-numeric values with suitable numbers, such as 0, depending on the context of your analysis.

(v) "Error: annotation cannot match the input data"

Solution: This error occurs when there are "NA" values in the columns for the expression values of the samples. To resolve this issue, you can either:

- Remove the rows containing "NA" values, or
- Replace the "NA" values with suitable numbers, such as 0, depending on your analysis needs.

(vi) Error in wilcox.test.default(e1, e2, paired = T, exact = FALSE) :
not enough (non-missing) 'x' observations
Called from: wilcox.test.default(e1, e2, paired = T, exact = FALSE)

Solution: This error occurs when there are no sncRNA species for the specified sncrna_family that meet the given threshold (row_mean). To resolve this issue, you can:

- Try using more lenient thresholds, or
- Choose a different sncrna_family that has sufficient data.

(vii) "Error: column number doesn't match. Please correct the sample numbers."

Solution: This error may occur when running FUSION_ms or FUSION_msmc if the values for S1 and S2 are not correctly specified in FUSION_ms, or if the conditions in the provided condition file (with the -cl option) do not correctly match the sample columns in the input matrix file. To resolve this issue:

- Verify that the input matrix has the correct number of columns. The total number of columns must be S1 + S2 + 2, where:
    - The first two columns must be for Sequence (or ID) and Annotation.
- Double-check that the conditions in the condition file match the columns for S1 and S2 in the matrix.

(viii) Error in `contrasts<-`(`*tmp*`, value = contr.funs[1 + isOF[nn]]) :
contrasts can be applied only to factors with 2 or more levels
Called from: `contrasts<-`(`*tmp*`, value = contr.funs[1 + isOF[nn]])

Solution: This error may occur when running FUSION_ms or FUSION_msmc if there are no sncRNA species for the specified sncrna_family that meet the given threshold (row_mean). To resolve this issue, you can:

- Try using more lenient thresholds, or
- Choose a different sncrna_family that has available data.

5. Visualizing Dysregulated RNA Species in Paired Samples

The R script plot_fusion.R can be used to generate a plot visualizing the positions of dysregulated RNA species along the length of a parental RNA sequence. Vertical dashed lines indicate paired RNA species from normal (blue) and tumor (red) tissues.

To run the script, you will need:

- A FASTA file of the parental RNA sequence (e.g., human_rRNA_28S.fa)
- An expression profile file containing sRNA sequences and their RPM values in paired samples (e.g., example_visualization_data.txt)

To visualize a specific region, provide the nucleotide coordinates using the pos_coord argument (e.g., pos_coord = c(640, 880) will zoom into the region from nucleotide 640 to 880 of the parental RNA).

Running the plot_fusion.R script will generate a PDF with two pages:

1. The first page shows the distribution of the top 100 (default) dysregulated RNA species across the full length of the parental RNA.
2. The second page provides a zoomed-in view of the specified region, highlighting the dysregulated RNA species within that range.

How to run:

Option 1. For users familiar with R:

Access the plot_fusion.R script from the installed package folder (scripts/), open it in your R environment, edit the input file paths as needed, and run the code.

Option 2. With guided instructions:

Use the following steps in your R console:

```
# Load the script from the installed package
script_path <- system.file("scripts", "plot_fusion.R", package = "FUSION")

# Open and edit input paths if you want to use your own data
file.edit(script_path)

# Run the script
source(script_path)

# Make sure to update in the script:
# - The path to your FASTA file
# - The path to your expression data file
# - The desired coordinates
# - The number of dysregulated RNA species to display
```

6. Preparing query matrix from SPORTS outputs

The R script ‘prepare_matrix_from_SPORTS_outputs.R’ can be used to prepare the count-matrix and RPM-matrix using the ‘X_output.txt’ files from the SPORTS output.

Example SPORTS output folder structure for a query dataset with 4 sample fastq files (SRRRXXXX1.fastq, SRRRXXXX2.fastq, SRRRXXXX3.fastq, SRRRXXXX4.fastq):

```
SPORTS_output
  │
  ├─1_SRRRXXXX1
  │   │
  │   ├─SRRRXXXX1_fa
  │   ├─SRRRXXXX1_processed
  │   └─SRRRXXXX1_result
  │       │
  │       ├─SRRRXXXX1_output.txt
  │       ├─
  │       ├─
  │       └─SRRRXXXX1_summary.txt
  │
  ├─2_SRRRXXXX2
  │   │
  │   ├─SRRRXXXX2_fa
  │   ├─SRRRXXXX2_processed
  │   └─SRRRXXXX2_result
  │       │
  │       ├─SRRRXXXX2_output.txt
  │       ├─
  │       ├─
  │       └─SRRRXXXX2_summary.txt
```

```
│
├─3_SRRRXXXX3
│   │
│   ├─SRRRXXXX3_fa
│   ├─SRRRXXXX3_processed
│   └─SRRRXXXX3_result
│       │
│       ├─SRRRXXXX3_output.txt
│       ├─
│       ├─
│       └─SRRRXXXX3_summary.txt
│
├─4_SRRRXXXX4
│   │
│   ├─SRRRXXXX4_fa
│   ├─SRRRXXXX4_processed
│   └─SRRRXXXX4_result
│       │
│       ├─SRRRXXXX4_output.txt
│       ├─
│       ├─
│       └─SRRRXXXX4_summary.txt
└─sh_file
```

First, prepare "file_list_document" specifying the full path of different relevant '_output.txt' files, as follows:

/path_to_SPORTS_output/1_SRRXXXX1/SRRXXXX1_result/SRRXXXX1_output.txt
/path_to_SPORTS_output/2_SRRXXXX2/SRRXXXX2_result/SRRXXXX2_output.txt
/path_to_SPORTS_output/3_SRRXXXX3/SRRXXXX3_result/SRRXXXX3_output.txt
/path_to_SPORTS_output/4_SRRXXXX4/SRRXXXX4_result/SRRXXXX4_output.txt

[Note: The order in which the files are listed will determine the sample order in the matrix file. Please ensure the files are arranged accordingly.]

Access the prepare_matrix_from_SPORTS_outputs.R script from the installed package folder (scripts/), and run as follows:

Rscript prepare_matrix_from_SPORTS_outputs.R file_list_document out_prefix

It will generate two files:
out_prefix_count-matrix.txt
out_prefix_RPM-matrix.txt

out_prefix_RPM-matrix.txt will serve as the input matrix file for FUSION run
out_prefix_count-matrix.txt can be used as input for differential expression analysis with DESeq2, edgeR, etc. i.e. tools which requires a count-matrix as an input.
