

# Package ‘ggmsa’

November 21, 2024

**Title** Plot Multiple Sequence Alignment using 'ggplot2'

**Version** 1.12.0

**Description** A visual exploration tool for multiple sequence alignment and associated data. Supports MSA of DNA, RNA, and protein sequences using 'ggplot2'. Multiple sequence alignment can easily be combined with other 'ggplot2' plots, such as phylogenetic tree Visualized by 'ggtree', boxplot, genome map and so on. More features: visualization of sequence logos, sequence bundles, RNA secondary structures and detection of sequence recombinations.

**Depends** R (>= 4.1.0)

**Imports** Biostrings, ggplot2, magrittr, tidyr, utils, stats, aplot, RColorBrewer, ggalt, ggforce, dplyr, R4RNA, grDevices, seqmagick, grid, methods, statebins, ggtree (>= 1.17.1)

**Suggests** ggtreeExtra, ape, cowplot, knitr, BiocStyle, rmarkdown, readxl, ggnewscale, kableExtra, gggenes, testthat (>= 3.0.0)

**License** Artistic-2.0

**Encoding** UTF-8

**URL** [https://doi.org/10.1093/bib/bbac222\(paper\)](https://doi.org/10.1093/bib/bbac222(paper)),  
<https://www.amazon.com/Integration-Manipulation-Visualization-Phylogenetic-Computational-ebook/dp/B0B5NLZR1Z/>  
(book)

**BugReports** <https://github.com/YuLab-SMU/ggmsa/issues>

**biocViews** Software, Visualization, Alignment, Annotation,  
MultipleSequenceAlignment

**RoxygenNote** 7.1.2

**VignetteBuilder** knitr

**Config/testthat/edition** 3

**git\_url** <https://git.bioconductor.org/packages/ggmsa>

**git\_branch** RELEASE\_3\_20

**git\_last\_commit** c09df2f

**git\_last\_commit\_date** 2024-10-29

**Repository** Bioconductor 3.20

**Date/Publication** 2024-11-20

**Author** Lang Zhou [aut, cre],

Guangchuang Yu [aut, ths] (<<https://orcid.org/0000-0002-6485-8781>>),

Shuangbin Xu [ctb],

Huina Huang [ctb]

**Maintainer** Lang Zhou <nyzhoulang@gmail.com>

## Contents

adjust_ally . . . . .	3
assign_dms . . . . .	3
available_colors . . . . .	4
available_fonts . . . . .	4
available_msa . . . . .	5
extract_seq . . . . .	5
facet_msa . . . . .	6
geom_GC . . . . .	6
geom_helix . . . . .	7
geom_msa . . . . .	8
geom_msaBar . . . . .	10
geom_seed . . . . .	10
geom_seqlogo . . . . .	11
gghelix . . . . .	12
ggmaf . . . . .	13
ggmsa . . . . .	13
ggSeqBundle . . . . .	15
Gram-negative_AKL.fasta . . . . .	17
Gram-positive_AKL.fasta . . . . .	17
GVariation . . . . .	17
LeaderRepeat_All.fa . . . . .	18
merge_seq . . . . .	18
plot . . . . .	19
readSSfile . . . . .	20
read_maf . . . . .	20
reset_pos . . . . .	21
Rfam . . . . .	21
sample.fasta . . . . .	22
seedSample.fa . . . . .	22
seqdiff . . . . .	22
seqlogo . . . . .	23
sequence-link-tree.fasta . . . . .	24
show . . . . .	24
simplify_hdata . . . . .	25
simplplot . . . . .	25
theme_msa . . . . .	26
tidy_hdata . . . . .	27
tidy_maf_df . . . . .	27
tidy_msa . . . . .	28
tp53.fa . . . . .	28
TP53_genes.xlsx . . . . .	29
treeMSA_plot . . . . .	29

**Index****31**

---

adjust_ally	<i>adjust_ally</i>
-------------	--------------------

---

**Description**

adjust the tree branch position after assigning ancestor node

**Usage**

```
adjust_ally(tree, node, sub = FALSE, seq_colname = "mol_seq")
```

**Arguments**

tree	ggtree object
node	internal node in tree
sub	logical value.
seq_colname	the colname of MSA on tree\$data

**Value**

tree

**Author(s)**

Lang Zhou

---

assign_dms	<i>assign_dms</i>
------------	-------------------

---

**Description**

assign dms value to alignments.

**Usage**

```
assign_dms(x, dms)
```

**Arguments**

x	data frame from tidy_msa()
dms	dms data frame

**Value**

tree

**Author(s)**

Lang Zhou

---

available\_colors      *List Color Schemes currently available*

---

**Description**

This function lists color schemes currently available that can be used by 'ggmsa'

**Usage**

```
available_colors()
```

**Value**

A character vector of available color schemes

**Author(s)**

Lang Zhou

**Examples**

```
available_colors()
```

---

available\_fonts      *List Font Families currently available*

---

**Description**

This function lists font families currently available that can be used by 'ggmsa'

**Usage**

```
available_fonts()
```

**Value**

A character vector of available font family names

**Author(s)**

Lang Zhou

**Examples**

```
available_fonts()
```

---

available_msa	<i>List MSA objects currently available</i>
---------------	---

---

**Description**

This function lists MSA objects currently available that can be used by 'ggmsa'

**Usage**

```
available_msa()
```

**Value**

A character vector of available objects

**Author(s)**

Lang Zhou

**Examples**

```
available_msa()
```

---

extract_seq	<i>extract_seq</i>
-------------	--------------------

---

**Description**

extract ancestor sequence from tree data

**Usage**

```
extract_seq(tree_adjust, seq_colname = "mol_seq")
```

**Arguments**

tree_adjust	ggtree object
seq_colname	the colname of MSA on tree\$data

**Value**

character

**Author(s)**

Lang Zhou

---

`facet_msa`*segment MSA*

---

**Description**

The MSA would be plot in a field that you set.

**Usage**

```
facet_msa(field)
```

**Arguments**

`field` a numeric vector of the field size.

**Value**

ggplot layers

**Author(s)**

Lang Zhou

**Examples**

```
library(ggplot2)
f <- system.file("extdata/sample.fasta", package="ggmsa")
# 2 fields
ggmsa(f, end = 120, font = NULL, color="Chemistry_AA") +
  facet_msa(field = 60)
# 3 fields
ggmsa(f, end = 120, font = NULL, color="Chemistry_AA") +
  facet_msa(field = 40)
```

---

`geom_GC`*geom\_GC*

---

**Description**

Multiple sequence alignment layer for ggplot2. It plot points of GC content.

**Usage**

```
geom_GC(show.legend = FALSE)
```

**Arguments**

`show.legend` logical. Should this layer be included in the legends?

**Value**

a ggplot layer

**Author(s)**

Lang Zhou

**Examples**

```
#plot GC content
f <- system.file("extdata/LeaderRepeat_All.fa", package="ggmsa")
ggmsa(f, font = NULL, color="Chemistry_NT") + geom_GC()
```

geom\_helix

*geom\_helix***Description**

The layer of helix plot

**Usage**

```
geom_helix(helix_data, color_by = "length", overlap = FALSE, ...)
```

**Arguments**

helix_data	a data frame. The file of nucleotide secondary structure and then read by readSSfile().
color_by	generate colors for helices by various rules, including integer counts and value ranges one of "length" and "value"
overlap	Logicals. If TRUE, two structures data called predict and known must be given(eg:helix_data = list(known = data1, predicted = data2)), plots the predicted helices that are known on top, predicted helices that are not known on the bottom, and finally plots unpredicted helices on top in black.
...	additional parameter

**Value**

ggplot2 layers

**Author(s)**

Lang Zhou

**Examples**

```
RF03120 <- system.file("extdata/Rfam/RF03120_SS.txt", package="ggmsa")
RF03120_fas <- system.file("extdata/Rfam/RF03120.fasta", package="ggmsa")
SS <- readSSfile(RF03120, type = "Vienna")
ggmsa(RF03120_fas, font = NULL, border = NA,
      color = "Chemistry_NT", seq_name = FALSE) +
geom_helix(SS)
```

geom\_msa

*geom\_msa***Description**

Multiple sequence alignment layer for ggplot2. It creates background tiles with/without sequence characters.

**Usage**

```
geom_msa(
  data,
  font = "helvetica",
  mapping = NULL,
  color = "Chemistry_AA",
  custom_color = NULL,
  char_width = 0.9,
  none_bg = FALSE,
  by_conservation = FALSE,
  position_highlight = NULL,
  seq_name = NULL,
  border = NULL,
  consensus_views = FALSE,
  use_dot = FALSE,
  disagreement = TRUE,
  ignore_gaps = FALSE,
  ref = NULL,
  position = "identity",
  show.legend = FALSE,
  dms = FALSE,
  position_color = FALSE,
  ...
)
```

**Arguments**

data	sequence alignment with data frame, generated by tidy_msa().
font	font families, possible values are 'helvetica', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'helvetica'.
mapping	aes mapping If font = NULL, only plot the background tile.
color	A Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6', 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
custom_color	A data frame with two column called "names" and "color".Customize the color scheme.
char_width	a numeric vector. Specifying the character width in the range of 0 to 1. Defaults is 0.9.
none_bg	a logical value indicating whether background should be displayed. Defaults is FALSE.



by_conservation	a logical value. The most conserved regions have the brightest colors.
position_highlight	A numeric vector of the position that need to be highlighted.
seq_name	a logical value indicating whether sequence names should be displayed. Defaults is 'NULL' which indicates that the sequence name is displayed when 'font = null', but 'font = char' will not be displayed. If 'seq_name = TRUE' the sequence name will be displayed in any case. If 'seq_name = FALSE' the sequence name will not be displayed under any circumstances.
border	a character string. The border color.
consensus_views	a logical value that opening consensus views.
use_dot	a logical value. Displays characters as dots instead of fading their color in the consensus view.
disagreement	a logical value. Displays characters that disagreement to consensus(excludes ambiguous disagreements).
ignore_gaps	a logical value. When selected TRUE, gaps in column are treated as if that row didn't exist.
ref	a character string. Specifying the reference sequence which should be one of input sequences when 'consensus_views' is TRUE.
position	Position adjustment, either as a string, or the result of a call to a position adjustment function, default is 'identity' meaning 'position_identity()'.
show.legend	logical. Should this layer be included in the legends?
dms	logical.
position_color	logical.
...	additional parameter

**Value**

A list

**Author(s)**

Guangchuang Yu, Lang Zhou seq\_name' work position\_highlight' work border' work none\_bg' work

**Examples**

```
library(ggplot2)
aln <- system.file("extdata", "sample.fasta", package = "ggmsa")
tidy_aln <- tidy_msa(aln, start = 150, end = 170)
ggplot() + geom_msa(data = tidy_aln, font = NULL) + coord_fixed()
```

geom\_msaBar

*geom\_msaBar*

---

**Description**

Multiple sequence alignment layer for ggplot2. It plot sequence conservation bar.

**Usage**

```
geom_msaBar()
```

**Value**

A list

**Author(s)**

Lang Zhou

**Examples**

```
#plot multiple sequence alignment and conservation bar.  
f <- system.file("extdata/sample.fasta", package="ggmsa")  
ggmsa(f, 221, 280, font = NULL, seq_name = TRUE) + geom_msaBar()
```

---

geom\_seed

*geom\_seed*

---

**Description**

Highlighting the seed in miRNA sequences

**Usage**

```
geom_seed(seed, star = FALSE)
```

**Arguments**

seed	a character string.Specifying the miRNA seed sequence like 'GAGGUAG'.
star	a logical value indicating whether asterisks should be displayed.

**Value**

a ggplot layer

**Author(s)**

Lang Zhou

**Examples**

```
miRNA_sequences <- system.file("extdata/seedSample.fa", package="ggmsa")
ggmsa(miRNA_sequences, font = 'DroidSansMono',
      color = "Chemistry_NT", none_bg = TRUE) +
geom_seed(seed = "GAGGUAG", star = FALSE)
ggmsa(miRNA_sequences, font = 'DroidSansMono',
      color = "Chemistry_NT") +
geom_seed(seed = "GAGGUAG", star = TRUE)
```

geom\_seqlogo

*geom\_seqlogo***Description**

Multiple sequence alignment layer for ggplot2. It plot sequence motifs.

**Usage**

```
geom_seqlogo(
  font = "DroidSansMono",
  color = "Chemistry_AA",
  adaptive = TRUE,
  top = TRUE,
  custom_color = NULL,
  show.legend = FALSE,
  ...
)
```

**Arguments**

font	font families, possible values are 'helvetica', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'DroidSansMono'.
color	A Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6', 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
adaptive	A logical value indicating whether the overall height of seqlogo corresponds to the number of sequences.If is FALSE, seqlogo overall height = 4, fixedly.
top	A logical value. If TRUE, seqlogo is aligned to the top of MSA.
custom_color	A data frame with two cloumn called "names" and "color".Customize the color scheme.
show.legend	logical. Should this layer be included in the legends?
...	additional parameter

**Value**

A list

**Author(s)**

Lang Zhou

## Examples

```
#plot multiple sequence alignment and sequence motifs
f <- system.file("extdata/LeaderRepeat_All.fa", package="ggmsa")
ggmsa(f, font = NULL, color = "Chemistry_NT") + geom_seqlogo()
```

---

gghelix

*gghelix*

---

## Description

Plots nucleotide secondary structure as helices in arc diagram

## Usage

```
gghelix(helix_data, color_by = "length", overlap = FALSE)
```

## Arguments

helix_data	a data frame. The file of nucleotide secondary structure and then read by readSSfile().
color_by	generate colors for helices by various rules, including integer counts and value ranges one of "length" and "value"
overlap	Logicals. If TRUE, two structures data called predict and known must be given(eg:helix_data = list(known = data1, predicted = data2)), plots the predicted helices that are known on top, predicted helices that are not known on the bottom, and finally plots unpredicted helices on top in black.

## Value

ggplot object

## Author(s)

Lang Zhou

## Examples

```
RF03120 <- system.file("extdata/Rfam/RF03120_SS.txt", package="ggmsa")
helix_data <- readSSfile(RF03120, type = "Vienna")
gghelix(helix_data)
```

ggmaf

*ggmaf***Description**

plot MAF

**Usage**

```
ggmaf(
  data,
  ref,
  block_start = NULL,
  block_end = NULL,
  facet_field = NULL,
  heights = c(0.4, 0.6),
  facet_heights = NULL
)
```

**Arguments**

data	a tidy MAF data frame. You can get it by tidy_maf_df()
ref	character, the name of reference genome. eg:"hg38.chr1_KI270707v1_random"
block_start	a numeric vector(>0). The start block to plot.
block_end	a numeric vector(< max block). The end block to plot.
facet_field	a numeric vector. The field in a facet panel.
heights	two numeric vector. The plot proportion between "Genomic location" panel(upon) and "Alignment" panel(down). Default:c(0.4,0.6)
facet_heights	Numeric vectors. The facet proportion.

**Value**

ggplot object

**Author(s)**

Lang Zhou

ggmsa

*ggmsa***Description**

Plot multiple sequence alignment using ggplot2 with multiple color schemes supported.

**Usage**

```
ggmsa(
  msa,
  start = NULL,
  end = NULL,
  font = "helvetica",
  color = "Chemistry_AA",
  custom_color = NULL,
  char_width = 0.9,
  none_bg = FALSE,
  by_conservation = FALSE,
  position_highlight = NULL,
  seq_name = NULL,
  border = NULL,
  consensus_views = FALSE,
  use_dot = FALSE,
  disagreement = TRUE,
  ignore_gaps = FALSE,
  ref = NULL,
  show.legend = FALSE
)
```

**Arguments**

<code>msa</code>	Multiple aligned sequence files or objects representing either nucleotide sequences or AA sequences.
<code>start</code>	a numeric vector. Start position to plot.
<code>end</code>	a numeric vector. End position to plot.
<code>font</code>	font families, possible values are 'helvetica', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'helvetica'. If font = NULL, only plot the background tile.
<code>color</code>	a Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6', 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
<code>custom_color</code>	A data frame with two column called "names" and "color".Customize the color scheme.
<code>char_width</code>	a numeric vector. Specifying the character width in the range of 0 to 1. Defaults is 0.9.
<code>none_bg</code>	a logical value indicating whether background should be displayed. Defaults is FALSE.
<code>by_conservation</code>	a logical value. The most conserved regions have the brightest colors.
<code>position_highlight</code>	A numeric vector of the position that need to be highlighted.
<code>seq_name</code>	a logical value indicating whether sequence names should be displayed. Defaults is 'NULL' which indicates that the sequence name is displayed when 'font = null', but 'font = char' will not be displayed. If 'seq_name = TRUE' the sequence name will be displayed in any case. If 'seq_name = FALSE' the sequence name will not be displayed under any circumstances.

border	a character string. The border color.
consensus_views	a logical value that opening consensus views.
use_dot	a logical value. Displays characters as dots instead of fading their color in the consensus view.
disagreement	a logical value. Displays characters that disagreement to consensus(excludes ambiguous disagreements).
ignore_gaps	a logical value. When selected TRUE, gaps in column are treated as if that row didn't exist.
ref	a character string. Specifying the reference sequence which should be one of input sequences when 'consensus_views' is TRUE.
show.legend	logical. Should this layer be included in the legends?

**Value**

ggplot object

**Author(s)**

Guangchuang Yu

**Examples**

```
#plot multiple sequences by loading fasta format
fasta <- system.file("extdata", "sample.fasta", package = "ggmsa")
ggmsa(fasta, 164, 213, color="Chemistry_AA")

## Not run:
#XMultipleAlignment objects can be used as input in the 'ggmsa'
AAMultipleAlignment <- readAAMultipleAlignment(fasta)
ggmsa(AAMultipleAlignment, 164, 213, color="Chemistry_AA")

#XStringSet objects can be used as input in the 'ggmsa'
AAStringSet <- readAAStringSet(fasta)
ggmsa(AAStringSet, 164, 213, color="Chemistry_AA")

#Xbin objects from 'seqmagick' can be used as input in the 'ggmsa'
AAbin <- fa_read(fasta)
ggmsa(AAbin, 164, 213, color="Chemistry_AA")

## End(Not run)
```

---

ggSeqBundle

*ggSeqBundle*


---

**Description**

plot Sequence Bundles for MSA based 'ggolot2'

**Usage**

```
ggSeqBundle(
  msa,
  line_widch = 0.3,
  line_thickness = 0.3,
  line_high = 0,
  spline_shape = 0.3,
  size = 0.5,
  alpha = 0.2,
  bundle_color = c("#2ba0f5", "#424242"),
  lev_molecule = c("-", "A", "V", "L", "I", "P", "F", "W", "M", "G", "S", "T", "C",
    "Y", "N", "Q", "D", "E", "K", "R", "H")
)
```

**Arguments**

<code>msa</code>	Multiple sequence alignment file(FASTA) or object for representing either nucleotide sequences or peptide sequences. Also receives multiple MSA files. eg: <code>msa = c("Gram-negative_AKL.fasta", "Gram-positive_AKL.fasta")</code> .
<code>line_widch</code>	The width of bundles at each site, default is 0.3.
<code>line_thickness</code>	The thickness of bundles at each site, default is 0.3.
<code>line_high</code>	The high of bundles at each site, default is 0.
<code>spline_shape</code>	A numeric vector of values between -1 and 1, which control the shape of the spline relative to the control points. From <code>geom_xspline()</code> in <code>ggalt</code> package.
<code>size</code>	A numeric vector of values between 0 and 1, which control the size of each lines.
<code>alpha</code>	A numeric vector of values between 0 and 1, which control the alpha of each lines.
<code>bundle_color</code>	The colors of each sequence bundles. eg: <code>bundle_color = c("#2ba0f5", "#424242")</code> .
<code>lev_molecule</code>	Reassigning the Y-axis and displaying letter-coded amino acids/nucleotides arranged by physiochemical properties or others. eg: amino acids hydrophobicity <code>lev_molecule = c("-", "A", "V", "L", "I", "P", "F", "W", "M", "G", "S", "T", "C", "Y", "N", "Q", "D", "E", "K", "R", "H")</code> .

**Value**

ggplot object

**Author(s)**

Lang Zhou

**Examples**

```
aln <- system.file("extdata", "Gram-negative_AKL.fasta", package = "ggmsa")
ggSeqBundle(aln)
```



---

Gram-negative\_AKL.fasta

*Gram-negative\_AKL*

---

**Description**

Amino acids in the adenylate kinase lid (AKL) domain from Gram-negative bacteria.

**Format**

A MSA fasta with 100 sequences and 36 positions.

**Source**

<http://biovis.net/year/2013/info/redesign-contest>

---

Gram-positive\_AKL.fasta

*Gram-positive\_AKL*

---

**Description**

Amino acids in the adenylate kinase lid (AKL) domain from Gram-positive bacteria.

**Format**

A MSA fasta with 100 sequences and 36 positions.

**Source**

<http://biovis.net/year/2013/info/redesign-contest>

---

GVariation

*GVariation*

---

**Description**

A folder containing 4 MAS files as a sample data set to identify the sequence recombination event.

**Format**

a folder

**Details**

- A.Mont.fas MSA with sequences of 'Mont' and 'CF\_YL21'
- B.Oz.fas MSA with sequences of 'Oz' and 'CF\_YL21'
- C.Wilga5.fas MSA with sequences of 'Wilga5' and 'CF\_YL21'
- sample\_alignment.fa MSA with sequences of 'Mont', 'CF\_YL21', 'Oz', and 'Wilga5'

**Source**

<https://link.springer.com/article/10.1007/s11540-015-9307-3>

---

LeaderRepeat\_All.fa    *A sample DNA alignment sequences*

---

**Description**

DNA alignment sequences with 24 sequences and 56 positions.

**Format**

A MSA fasta

---

merge\_seq                    *merge\_seq*

---

**Description**

merge two MSA

**Usage**

```
merge_seq(previous_seq, gap, subsequent_seq, adjust_name = TRUE)
```

**Arguments**

previous_seq	previous MSA
gap	gap length
subsequent_seq	subsequent MSA
adjust_name	logical value. merge seq name or not

**Value**

tidy MSA data frame

**Author(s)**

Lang Zhou

---

plot	<i>plot method for SeqDiff object</i>
------	---------------------------------------

---

**Description**

plot method for SeqDiff object

**Usage**

```
## S4 method for signature 'SeqDiff,ANY'  
plot(  
  x,  
  width = 50,  
  title = "auto",  
  xlab = "Nucleotide Position",  
  by = "bar",  
  fill = "firebrick",  
  colors = c(A = "#ff6d6d", C = "#769dcc", G = "#f2be3c", T = "#74ce98"),  
  xlim = NULL  
)
```

**Arguments**

x	SeqDiff object
width	bin width
title	plot title
xlab	xlab
by	one of 'bar' and 'area'
fill	fill color of upper part of the plot
colors	color of lower part of the plot
xlim	limits of x-axis

**Value**

plot

**Author(s)**

guangchuang yu

**Examples**

```
fas <- list.files(system.file("extdata", "GVariation", package="ggmsa"),  
                 pattern="fas", full.names=TRUE)  
x1 <- seqdiff(fas[1], reference=1)  
plot(x1)
```

---

readSSfile	<i>readSSfile</i>
------------	-------------------

---

**Description**

Read secondary structure file

**Usage**

```
readSSfile(file, type = NULL)
```

**Arguments**

file	A text file in connect format
type	file type. one of "Helix", "Connect", "Vienna" and "Bpseq"

**Value**

data frame

**Author(s)**

Lang Zhou

**Examples**

```
RF03120 <- system.file("extdata/Rfam/RF03120_SS.txt", package="ggmsa")  
helix_data <- readSSfile(RF03120, type = "Vienna")
```

---

read_maf	<i>read_maf</i>
----------	-----------------

---

**Description**

read 'multiple alignment format' (MAF) file

**Usage**

```
read_maf(multiple_alignment_format)
```

**Arguments**

multiple_alignment_format	a multiple alignment format(MAF) file
---------------------------	---------------------------------------

**Value**

data frame

**Author(s)**

Lang Zhou

---

reset_pos	<i>reset_pos</i>
-----------	------------------

---

**Description**

reset MSA position

**Usage**

```
reset_pos(seq_df)
```

**Arguments**

seq_df	MSA data
--------	----------

**Value**

data frame

**Author(s)**

Lang Zhou

---

Rfam	<i>Rfam</i>
------	-------------

---

**Description**

A folder containing seed alignment sequences and corresponding consensus RNA secondary structure.

**Format**

a folder

**Details**

- RF00458.fasta seed alignment sequences of Cripavirus internal ribosome entry site (IRES)
- RF03120.fasta seed alignment sequences of Sarbecovirus 5'UTR
- RF03120\_SS.txt consensus RNA secondary structure of Sarbecovirus 5'UTR

**Source**

<https://rfam.xfam.org/>

sample.fasta                    *A sample data used in ggmsa*

---

**Description**

A dataset containing the alignment sequences of the phenylalanine hydroxylase protein (PH4H) within nine species

**Format**

A MSA fasta with 9 sequences and 456 positions.

---

seedSample.fa                    *microRNA data used in ggmsa*

---

**Description**

Fasta format sequences of mature miRNA sequences from miRBase

**Format**

A MSA fasta with 6 sequences and 22 positions.

**Source**

<https://www.mirbase.org/ftp.shtml>

---

seqdiff                    *seqdiff*

---

**Description**

calculate difference of two aligned sequences

**Usage**

```
seqdiff(fasta, reference = 1)
```

**Arguments**

fasta	fasta file
reference	which sequence serve as reference, 1 or 2

**Value**

SeqDiff object

**Author(s)**

guangchuang yu

**Examples**

```
fas <- list.files(system.file("extdata", "GVariation", package="ggmsa"),
                 pattern="fas", full.names=TRUE)
seqdiff(fas[1], reference=1)
```

seqlogo

*seqlogo***Description**

plot sequence logo for MSA based 'ggolot2'

**Usage**

```
seqlogo(
  msa,
  start = NULL,
  end = NULL,
  font = "DroidSansMono",
  color = "Chemistry_AA",
  adaptive = FALSE,
  top = FALSE,
  custom_color = NULL
)
```

**Arguments**

<code>msa</code>	Multiple sequence alignment file or object for representing either nucleotide sequences or peptide sequences.
<code>start</code>	Start position to plot.
<code>end</code>	End position to plot.
<code>font</code>	font families, possible values are 'helvetica', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'DroidSansMono'. If font=NULL, only the background tiles is drawn.
<code>color</code>	A Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6', 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
<code>adaptive</code>	A logical value indicating whether the overall height of seqlogo corresponds to the number of sequences. If FALSE, seqlogo overall height = 4, fixedly.
<code>top</code>	A logical value. If TRUE, seqlogo is aligned to the top of MSA.
<code>custom_color</code>	A data frame with two column called "names" and "color". Customize the color scheme.

**Value**

ggplot object

**Author(s)**

Lang Zhou

**Examples**

```
#plot sequence motif independently
nt_sequence <- system.file("extdata", "LeaderRepeat_All.fa",
                           package = "ggmsa")
seqlogo(nt_sequence, color = "Chemistry_NT")
```

---

sequence-link-tree.fasta

*sequence-link-tree*

---

**Description**

Alignment sequences used to demonstrate circular MSA layout

**Format**

A MSA fasta with 28 sequences and 480 positions.

---

show

*show method*

---

**Description**

show method

**Usage**

show(object)

**Arguments**

object            SeqDiff object

**Value**

message

**Examples**

```
fas <- list.files(system.file("extdata", "GVariation", package="ggmsa"),
                  pattern="fas", full.names=TRUE)
x1 <- seqdiff(fas[1], reference=1)
x1
```



---

simplify_hdata	<i>simplify_hdata</i>
----------------	-----------------------

---

**Description**

reset hdata data position

**Usage**

```
simplify_hdata(hdata, sim_msa)
```

**Arguments**

hdata	data from tidy_hdata()
sim_msa	MSA data frame

**Value**

data frame

**Author(s)**

Lang Zhou

---

simplot	<i>simplot</i>
---------	----------------

---

**Description**

Sequence similarity plot

**Usage**

```
simplot(  
  file,  
  query,  
  window = 200,  
  step = 20,  
  group = FALSE,  
  id,  
  sep,  
  sd = FALSE,  
  smooth = FALSE,  
  smooth_params = list(method = "loess", se = FALSE)  
)
```

**Arguments**

file	alignment fast file
query	query sequence
window	sliding window size (bp)
step	step size to slide the window (bp)
group	whether grouping sequence.(eg. For "A-seq1,A-seq-2,B-seq1 and B-seq2", using sep = "-" and id = 1 to divide sequences into groups A and B)
id	position to extract id for grouping; only works if group = TRUE
sep	separator to split sequence name; only works if group = TRUE
sd	whether display standard deviation of similarity among each group; only works if group=TRUE
smooth	FALSE(default)or TRUE; whether display smoothed spline.
smooth_params	a list that add params for geom_smooth, (default: smooth_params = list(method = "loess", se = FALSE))

**Value**

ggplot object

**Author(s)**

guangchuang yu

**Examples**

```
fas <- system.file("extdata/GVariation/sample_alignment.fa",
                  package="ggmsa")
simplot(fas, 'CF_YL21')
```

---

theme\_msa

*theme\_msa*

---

**Description**

Theme for ggmsa.

**Usage**

```
theme_msa()
```

**Author(s)**

Lang Zhou

---

tidy_hdata	<i>tidy_hdata</i>
------------	-------------------

---

**Description**

tidy protein-protein interactive position data

**Usage**

```
tidy_hdata(gap, inter, previous_seq, subsequent_seq)
```

**Arguments**

gap	gap length
inter	protein-protein interactive position data
previous_seq	previous MSA
subsequent_seq	subsequent MSA

**Value**

helix data

**Author(s)**

Lang Zhou

---

tidy_maf_df	<i>tidy_maf_df</i>
-------------	--------------------

---

**Description**

tidy MAF data frame

**Usage**

```
tidy_maf_df(maf_df, ref)
```

**Arguments**

maf_df	a MAF data frame. You can get it by read_maf()
ref	character, the name of reference genome. eg: "hg38.chr1_KI270707v1_random"

**Value**

data frame

**Author(s)**

Lang Zhou

---

tidy_msa	<i>tidy_msa</i>
----------	-----------------

---

**Description**

Convert msa file/object to tidy data frame.

**Usage**

```
tidy_msa(msa, start = NULL, end = NULL)
```

**Arguments**

msa	multiple sequence alignment file or sequence object in DNAStrngSet, RNAS-trngSet, AAStringSet, BStringSet, DNAMultipleAlignment, RNAMultipleAlign-ment, AAMultipleAlignment, DNAbin or AAbin
start	start position to extract subset of alignment
end	end position to extract subset of alignemnt

**Value**

tibble data frame

**Author(s)**

Guangchuang Yu

**Examples**

```
fasta <- system.file("extdata", "sample.fasta", package = "ggmsa")
aln <- tidy_msa(msa = fasta, start = 10, end = 100)
```

---

tp53.fa	<i>TP53 MSA</i>
---------	-----------------

---

**Description**

Alignment sequences of used to show graphical combination

**Format**

A MSA fasta with 5 sequences and 404 positions.

---

TP53_genes.xlsx	<i>genome locus</i>
-----------------	---------------------

---

**Description**

The local genome map shows the 30000 sites around the TP53 gene.

**Format**

xlsx

---

treeMSA_plot	<i>treeMSA_plot</i>
--------------	---------------------

---

**Description**

plot Tree-MSA plot

**Usage**

```
treeMSA_plot(
  p_tree,
  tidymsa_df,
  ancestral_node = "none",
  sub = FALSE,
  panel = "MSA",
  font = NULL,
  color = "Chemistry_AA",
  seq_colname = NULL,
  ...
)
```

**Arguments**

<code>p_tree</code>	tree view
<code>tidymsa_df</code>	tidy MSA data
<code>ancestral_node</code>	vector, internal node in tree. Assigning a internal node to display "ancestral sequences", If <code>ancestral_node = "none"</code> hides all ancestral sequences, if <code>ancestral_node = "all"</code> shows all ancestral sequences.
<code>sub</code>	logical value. Displaying a subset of ancestral sequences or not.
<code>panel</code>	panel name for plot of MSA data
<code>font</code>	font families, possible values are 'helvetica', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'helvetica'. If <code>font = NULL</code> , only plot the background tile.
<code>color</code>	a Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6', 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
<code>seq_colname</code>	the colname of MSA on <code>tree\$data</code>
<code>...</code>	additional parameters for 'geom_msa'

**Details**

'treeMSA\_plot()' automatically re-arranges the MSA data according to the tree structure,

**Value**

ggplot object

**Author(s)**

Lang Zhou

# Index

## \* datasets

- Gram-negative\_AKL.fasta, 17
  - Gram-positive\_AKL.fasta, 17
  - GVariation, 17
  - LeaderRepeat\_All.fa, 18
  - Rfam, 21
  - sample.fasta, 22
  - seedSample.fa, 22
  - sequence-link-tree.fasta, 24
  - tp53.fa, 28
  - TP53\_genes.xlsx, 29
- adjust\_ally, 3
- assign\_dms, 3
- available\_colors, 4
- available\_fonts, 4
- available\_msa, 5
- extract\_seq, 5
- facet\_msa, 6
- geom\_GC, 6
- geom\_helix, 7
- geom\_msa, 8
- geom\_msaBar, 10
- geom\_seed, 10
- geom\_seqlogo, 11
- gghelix, 12
- ggmaf, 13
- ggmsa, 13
- ggSeqBundle, 15
- Gram-negative\_AKL.fasta, 17
- Gram-positive\_AKL.fasta, 17
- GVariation, 17
- LeaderRepeat\_All.fa, 18
- merge\_seq, 18
- plot, 19
- plot, SeqDiff, ANY-method (plot), 19
- read\_maf, 20
- readSSfile, 20
- reset\_pos, 21
- Rfam, 21
- sample.fasta, 22
- seedSample.fa, 22
- seqdiff, 22
- SeqDiff-class (show), 24
- seqlogo, 23
- sequence-link-tree.fasta, 24
- show, 24
- show, SeqDiff-method (show), 24
- simplify\_hdata, 25
- simplot, 25
- theme\_msa, 26
- tidy\_hdata, 27
- tidy\_maf\_df, 27
- tidy\_msa, 28
- tp53.fa, 28
- TP53\_genes.xlsx, 29
- treeMSA\_plot, 29