The Journal of Open Source Software

# tidyHeatmap: an R package for modular heatmap production based on tidy principles

**Stefano Mangiola**[*1, 2] **and Anthony T. Papenfuss**[†1, 2, 3, 4]

**1** Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia **2** Department of Medical Biology, University of Melbourne, Melbourne, Victoria, Australia. **3** Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia. **4** School of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia.

**DOI:** 10.21105/joss.02472

**Software**

- Review
- Repository
- Archive

**Editor:** Karthik Ram

**Reviewers:**

- @rlbarter
- @AliciaSchep

**Submitted:** 08 July 2020
**Published:** 03 August 2020



## Background

The heatmap is a powerful tool for visualising multi-dimensional data, where individual values can be organised in a two-dimensional matrix and their values expressed as colours. Rows and columns can be ordered according to their reciprocal similarity using hierarchical clustering, and dendrograms can be added to the plot to facilitate the interpretation. Row- and column-wise visual annotations, such as coloured tiles, can also be included. Within the R environment, several packages have been developed to produce heatmaps. The simplest and most readily available tool, `heatmap`, is provided within the `stats` package (R Core Team, 2013) and offers basic heatmaps with simple tile annotations. The versatile package `ggplot2` can be also used to produce basic heatmaps (Wickham, 2016). More powerful software exists for producing fully annotated and/or multi-panel heatmaps, such as `Pheatmap` (Kolde, 2012), `superheat` (Barter & Yu, 2018) and `ComplexHeatmap` (Gu, Eils, & Schlesner, 2016). The versatility of these packages comes at the cost of adding complexity to the user interface, characterised by many parameters and annotation functions that introduce a steep learning curve to produce complex, clear, and good-looking graphics.

Recently, efforts have been made toward the harmonisation of data frame structures and data analysis workflows using the concept of tidiness. Tidy data frames are characterised by having a specific structure where each variable is a column and each observation is a row. They provide ease of manipulation, modelling, and visualisation. The `tidyverse` is a suite of R libraries that define the standard for tidy data and APIs (Wickham et al., 2019). The unique correspondence between quantities and annotations, characteristic of tidy data frames, allows complex operations to be performed from simple user inputs, such as a list of column names.

## Statement of need

Considering (i) the utility and complexity of creating information-rich heatmaps, and (ii) the opportunity of increased coding efficiency and robustness offered by the tidy paradigm, a bridge between the two is very much needed. Recently, many tools for data science have been implemented according with tidy principles, this package aims to fill the gap for one of the most used data explorations tools.

*Corresponding author
†Corresponding author

## Tidy paradigm for visualisation

`tidyHeatmap` is a R package that introduces tidy principles to the creation of information-rich heatmaps. It is available in the CRAN R repository. This package currently uses `ComplexHeatmap` as its graphical engine; however, due to its modular design it can be readily expanded to interface other engines. The command-line user interface is organised into (i) a main plotting utility; (ii) annotation layer utilities; and (iii) file-output utilities. The input is a tidy data frame with element (*e.g.*, person), feature (*e.g.*, physical characteristics) and value columns, with additional columns for independent variables for either elements (*e.g.*, number of sport medals) or features (*e.g.*, macroscopic or molecular characteristics). In this data structure, each observation is an element-feature pair.

| element | feature | value | annotation | group |
|---|---|---|---|---|
| `chr` or `fctr` | `chr` or `fctr` | `numeric` | ... | ... |

The input data frame streams along the utility path using the pipe operator from `magrittr`, allowing high modularity. The main utility allows the user to plot a base heatmap with dendrograms. The annotation utilities allow to serially add tile, point, bar and/or line annotation boxes to the side of the heatmap. The orientation of the annotations (row- or column-wise) is inferred based on the input data frame. The file-output utility allows the user to save vector or bitmap images directly from the R object, in the style of `ggplot2`. Row- or column-wise clusters can be defined effortlessly by applying the `group_by` function from `dplyr` (Wickham, François, Henry, & Müller, 2020) to the input data frame. Data transformation and row/column scaling is done internally. Together, this leads to a decrease of coding burden of 3 and 5 folds for lines and characters respectively compared to `ComplexHeatmap` (*e.g.*, for Figure 1). Besides offering a modular and user-friendly interface, `tidyHeatmap` applies publication-ready aesthetics such as `viridis` (Garnier, 2018) and `brewer` (Neuwirth, 2014) colour palettes and automatic sizing of row and column labels to avoid overlapping (Figure 1).

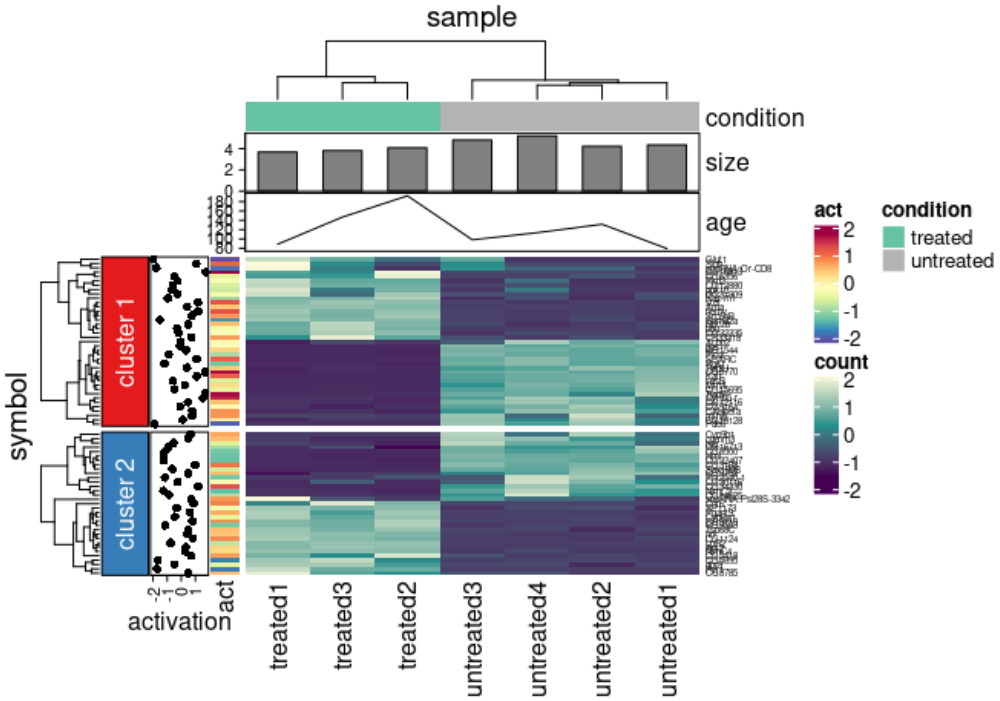


**Figure 1:** Heatmap of the pasilla dataset including grouping and multiple annotations. Some annotation data was simulated for visualisation purposes.

The code interface consists of modular functions linked through the pipe operator. Custom colour palettes can be used by passing an array of colours or a colour function (*e.g.*, circlize (Gu, Gu, Eils, Schlesner, & Brors, 2014)) to the palette argument of the annotation utilities.

```
my_heatmap =

    # Grouping
    input_df %>%
    group_by(pathway) %>%

    # Plotting
    heatmap(feature, element, value) %>%

    # Annotation
    add_tile(condition) %>%
    add_tile(act) %>%
    add_point(activation) %>%
    add_bar(size) %>%
    add_line(age)

# Saving
my_heatmap %>% save_pdf("my_file.pdf")
```

## Conclusions

In order to perform complex tasks, the use of disjointed data structures demands time consuming and bug-prone information matching. Joint, tidy data frames decrease the cost/benefit ratio for the user, automating a large part on the data manipulation. `tidyHeatmap` introduces a modular paradigm for specifying information-rich heatmaps, just requiring column names as input. Due to its intuitive user interface and its advanced default aesthetic features, `tidyHeatmap` is ideal for the quick production of publication-ready heatmaps. This software is designed for modular expandability. Future directions include the incorporation of more static and interactive heatmap visualisation engines.

## Acknowledgements

We acknowledge contributions all the Papenfuss Lab for feedback, and Maria Doyle for constant support.

## References

Barter, R. L., & Yu, B. (2018). Superheat: An R package for creating beautiful and extendable heatmaps for visualizing complex data. *Journal of Computational and Graphical Statistics*, *27*(4), 910–922. doi:10.1080/10618600.2018.1473780

Garnier, S. (2018). *viridis: Default color maps from 'matplotlib'*. Retrieved from https://CRAN.R-project.org/package=viridis

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. doi:10.1093/bioinformatics/btw313

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. doi:10.1093/bioinformatics/btu393

Kolde, R. (2012). Pheatmap: Pretty heatmaps. *R package version*.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. Retrieved from https://CRAN.R-project.org/package=RColorBrewer

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr