The Journal of Open Source Software

HCP: A Matlab package to create beautiful heatmaps with richly annotated covariates

Manuela Salvucci[1] **and Jochen H. M. Prehn**[1]

1 Centre for Systems Medicine, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, Dublin, Ireland

DOI: 10.21105/joss.01291

Software

- Review
- Repository
- Archive

Submitted: 26 January 2019
Published: 12 June 2019

Summary

A heatmap is a graphical technique that maps 2-dimensional matrices of numerical values to colors to provide an immediate and intuitive visualization of the underlying patterns (Eisen, Spellman, Brown, & Botstein, 1998). Heatmaps are often used in conjunction with cluster analysis to re-order observations and/or features by similarity and thus, rendering common and distinct patterns more apparent. When generating these visualizations, it is often of interest to interpret the underlying patterns in the context of other data sources. In the field of bioinformatics, heatmaps are frequently used to visualize high-throughput and high-dimensional datasets, such as those derived from profiling biological samples with *-omic* technologies (whole genome sequencing, transcriptomics and proteomics). Often, biological samples (for example, patient tumour samples) are characterized at multiple *-omic* level and it is of interest to contrast and compare patterns captured at the different molecular layers along with their associations with other observable features (covariates). The concurrent display of continuous or categorical covariates enriches the visualization with additional information such as group membership.

The R-language features powerful packages to create annotated heatmaps with complex layouts such as *pheatmap* (Kolde, 2019), *ComplexHeatmap* (Gu, Eils, & Schlesner, 2016) and *Superheat* (Barter & Yu, 2018). However, to the best of our knowledge, Matlab only features the *HeatMap* and *Clustergram* classes (from the *Bioinformatics Toolbox*) to draw simple heatmaps with/without cluster analysis with basic annotations and there is no package with the breath of those in the R ecosystem. The HCP package fills this gap by enabling bioinformaticians and data scientists from other fields to create richly annotated visualizations for data exploration in Matlab.

HCP (*HeatmapCovariatePlot*) provides a simple high level application programming interface (API) to design elaborated visualizations in a modular fashion. The user can select which elements to include, covariate row annotations and/or heatmaps, by invoking the `AddCovariateRow` or the `AddHeatmap` methods. Elements can be vertically stacked and also grouped in functionally related sub-blocks encapsulated by the `AddSubBlock` method to adjust the figure layout. The plotting options in *HCP* are chosen sensibly to create production-quality out-of-the-box visualizations in most use-case. *HCP* features several plotting options to adjust the plot aesthetics to cater for the user preferences in terms of colormaps, labelling, legends and layouts (margins and positions). *HCP* ease-of-use and rapidity enables the users to iterate through multiple visualization alternatives while focusing on the message conveyed by the data rather than the technicalities involved in generating the plot.

HCP plotting functionality has been applied in a scholarly manuscript currently undergoing revisions and in exploratory analyses in several other ongoing research projects.

Figure 1 and Figure 2 showcase visualizations generated with *HCP* for a case study drawn from a bioinformatic application in cancer research (cutaneous melanoma) and a more general data science application (occupation by US state).

Case study 1: Clinical and molecular characterization of a cohort of cutaneous melanoma patients

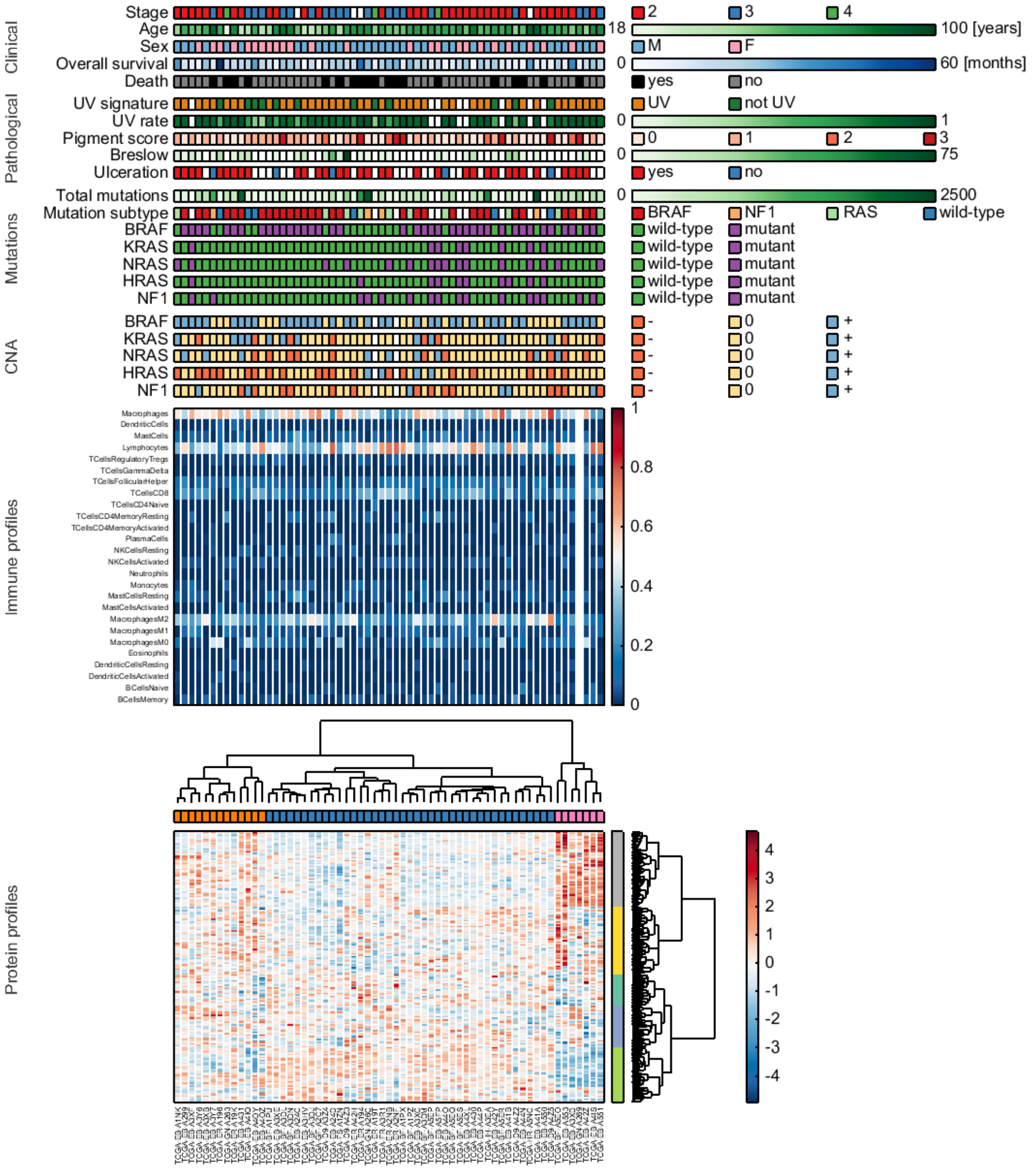

Figure 1. *Case study 1.* Integrated visualization of clinico-pathological features (Liu et al., 2018) and molecular alterations in mutational status, copy number alterations (CNAs) (Akbani et al., 2015), immune cell composition (Thorsson et al., 2018) and protein profiles (Grossman et al., 2016; TCGA Network, 2018) determined in primary tumour samples of n=61 cutaneous melanoma patients from the The Cancer Genome Atlas network (TCGA) cohort. Source data for the figure are further detailed in tcga_cutaneous_melanoma_cancer.md.

Case study 2: Geographical, demographic, socio-economic and work occupations by US state

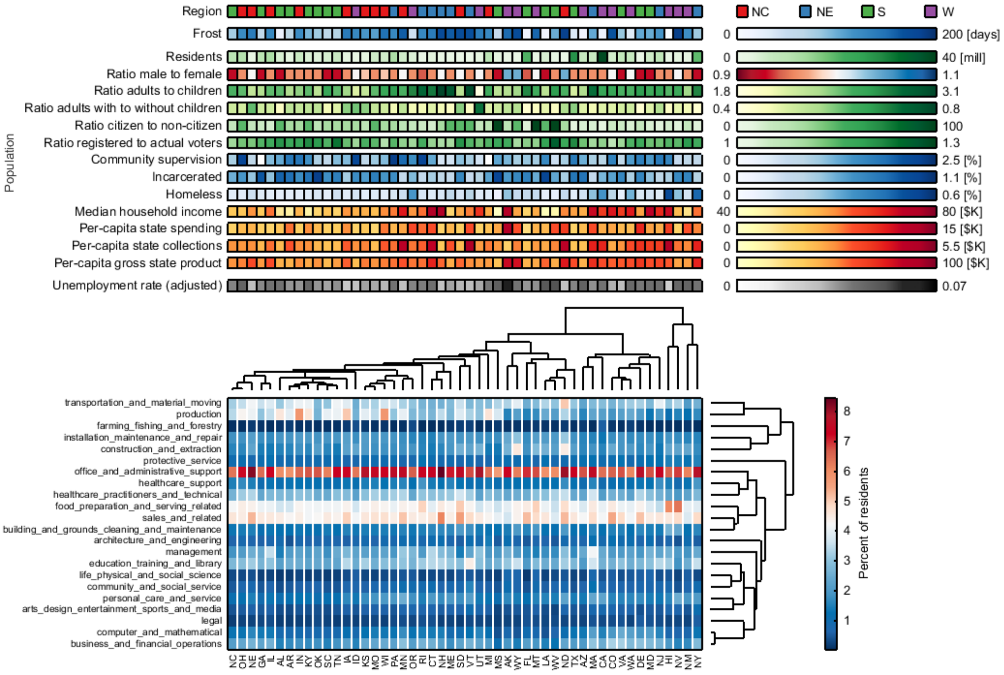

Figure 2. *Case study 2.* Integrated visualization of geographical (Bureau of the Census U.S. Department of Commerce, 1977a, 1977b; R Core Team, 2018), demographic and socio-economic metadata (The Kaiser Family Foundation, 2019) and work occupations by US state ("Occupational employment statistics," 2019). Source data for the figure are further detailed in occupation_by_us_state.md.

Acknowledgements

HCP uses extensively the *panel* package (Mitch, 2019). ColorBrewer colormaps (Brewer, 2019) are supported *via* the *brewermap* package (Cobeldick, 2019).

The visualization for case study 1 shown here is based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. The visualization for case study 2 is based largely upon data generated by the Henry J Kaiser Family Foundation: https://www.kff.org/ and by the Bureau of Labor Statistics: https://www.bls.gov/oes/.

This package was developed to support research funded by Science Foundation Ireland (13/IA/1881 and 14/IA/2582).

References

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., Arachchi, H., et al. (2015). Genomic Classification of Cutaneous Melanoma. *Cell*, *161*(7), 1681–1696. doi:10.1016/j.cell.2015.05.044

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

Brewer, C. A. (2019). ColorBrewer. Retrieved from http://www.colorbrewer.org

Bureau of the Census U.S. Department of Commerce. (1977a). County and City Data Book. Retrieved from https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/state.html

Bureau of the Census U.S. Department of Commerce. (1977b). Statistical Abstract of the United States. Retrieved from https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/state.html

Cobeldick, S. (2019). ColorBrewer: Attractive and Distinctive Colormaps. Retrieved from https://bit.ly/2Wqlbg0

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(25), 14863–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9843981

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, *375*(12), 1109–1112. doi:10.1056/NEJMp1607591

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

Kolde, R. (2019). *pheatmap: Pretty Heatmaps*. Retrieved from https://cran.r-project.org/package=pheatmap

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, *173*(2), 400–416.e11. doi:10.1016/j.cell.2018.02.052

Mitch, B. (2019). Panel. Retrieved from https://uk.mathworks.com/matlabcentral/fileexchange/20003-panel

Occupational employment statistics. (2019). Bureau of Labor Statistics, U.S. Department of Labor. Retrieved from https://www.bls.gov/oes/special.requests/oesm17st.zip

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

TCGA Network. (2018). TCGA PanCanAtlas. Retrieved from http://api.gdc.cancer.gov/data/fcbb373e-28d4-4818-92f3-601ede3da5e1

The Kaiser Family Foundation. (2019). KFF's State Health Facts. Data sources: Census Bureau's American Community Survey, 2008-2017; U.S. Census Bureau, Current Population Survey, November 2014 and 2016; Voting and Registration Tables; U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Correctional Populations in the United States Series, Accessed June, 28, 2018; U.S. Department of Housing and Urban Development, Point in Time Estimates of Homelessness, 2017, December 2017; U.S. Census Bureau, 2017 Current Population Survey, Annual Social and Economic Supplements Data Tables. Historical Household Income; KCMU calculations based on adjusted data collected in the National Association of State Budget Officers (NASBO) State Expenditure Report: Examining Fiscal 2015-2017 State Spending, November 2017; and the U.S. Census Bureau Resident Population Data, 2016; KFF calculation based on data from the U.S. Census Bureau, 2017 Annual Survey of State Government Tax Collections and the U.S. Census Bureau Population Estimates, accessed April 10, 2018; U.S. Bureau of Economic Analysis (BEA), Broad Growth Across

States in 2014: Advance Statistics of GDP by State, Table 4. Current-Dollar GDP by State, 2011-2014, June 10, 2015; Bureau of Labor Statistics (BLS), Regional and State Employment and Unemployment (Monthly), Civilian labor force and unemployment by state and selected area, seasonally adjusted, September report for 2013-2018. Retrieved from https://www.kff.org/

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., Porta-Pardo, E., et al. (2018). The Immune Landscape of Cancer. *Immunity*, *48*(4), 812–830.e14. doi:10.1016/j.immuni.2018.03.023