The Journal of Open Source Software

ggoncoplot: an R package for visualising somatic mutation data from cancer patient cohorts

Sam El-Kamand [1], **Julian M. W. Quinn** [1,2], **and Mark J. Cowley** [1,3¶]

1 Children's Cancer Institute, Australia **2** Skeletal Research Program, Garvan Institute of Medical Research, Australia **3** School of Clinical Medicine, UNSW Medicine & Health, Australia ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

The ggoncoplot R package generates interactive oncoplots (also called oncoprints) that visualize mutational patterns across patient cancer cohorts (Figure 1). These plots reveal patterns of mutation co-occurrence in a cohort, with marginal plots that indicate correlations between gene mutations and specific tumour characteristics. These tumour characteristics can include user-supplied annotations such as patient clinical data, cancer subtypes and histological features.

ggoncoplot offers several features to enhance utility and user experience. These include automatic colour palette selection for common mutation impact dictionaries, customizable tooltips, and automatic rendering of clinical annotations as barplots or heatmaps for quantitative or qualitative data, respectively. ggoncoplot supports visualisation of mutation-level data in tidy, tabular formats, making it easy to run on existing large somatic mutation datasets stored as MAF files or in relational databases. The ggoncoplot package is available from github at https://github.com/selkamand/ggoncoplot.

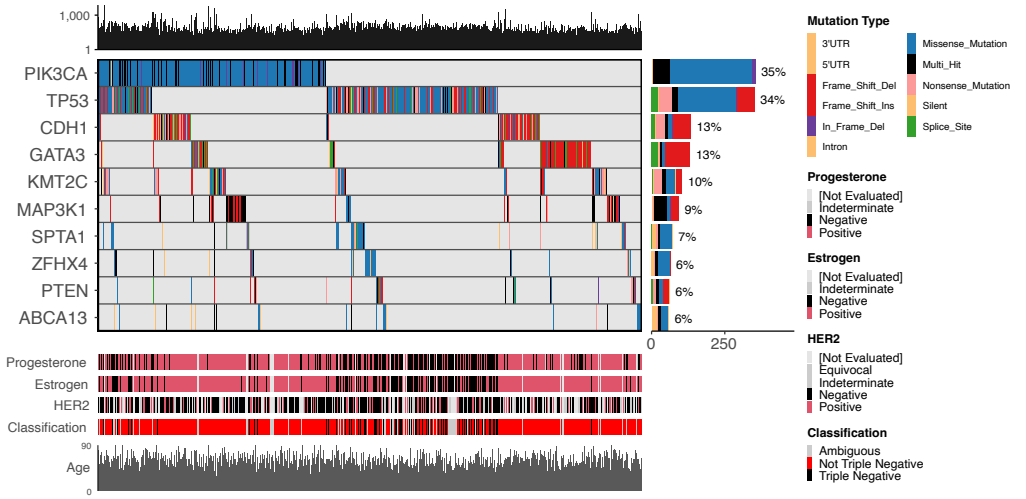

Figure 1: ggoncoplot output visualising mutational trends in the TCGA breast carcinoma cohort. Individual patient samples are plotted on the x-axis, ordered by ggoncoplot. The plot indicates (y-axis, sorted by genes mutation frequency) that PIK3CA is the most recurrently mutated gene, followed by TP53. Marginal plots indicate the total number of mutations per sample (top), and the number of samples showing mutations in each gene, coloured by mutation type (right). A range of clinical features, including progesterone and estrogen receptor status are shown on the marginal plot at the bottom.

The Journal of Open Source Software

Statement of Need

Oncoplots effectively visualize cohort-level mutation but are challenging to generate with the major R plotting systems (base, lattice, or ggplot2) due to their algorithmic and graphical complexity. Simplifying the process would make oncoplots more accessible to researchers. Packages like ComplexHeatmap (Gu, 2022), maftools (Mayakonda et al., 2018), and genVisR (Skidmore et al., 2016) make static oncoplots easier to create, but there remains a significant and unmet need to easily create oncoplots with:

- **Interactive features**: Customizable tooltips, cross-selection of samples across different plots, and auto-copying of sample identifiers on click.
- **Support for tidy datasets**: Compatibility with tidy, tabular mutation-level formats (MAF files or relational databases), typical of cancer cohort datasets.
- **Auto colouring**: Automatic selection of color palettes for datasets with consequence annotations aligned with standard variant effect dictionaries (PAVE, SO, or MAF).
- **Versatility**: The ability to visualize entities beyond gene mutations, including noncoding features (e.g., enhancers) and non-genomic entities (e.g., microbial presence in microbiome datasets).

We developed ggoncoplot as the first R package that addresses all these challenges simultaneously (Figure 2). Examples of all key features are available in the ggoncoplot manual.

Property	complexheatmap	maftools	GenVisR	ggoncoplot
Sample sorting algorithm	memo sort	heirarchical sort	heirarchical sort	heirarchical sort
Plotting framework	BaseR	BaseR	ggplot2	ggplot2
Automatic rendering of clinical annotations as bar or tile plots based on datatype	No	No	No	**Yes**
Works on tabular, tidy, long-form input data as would be stored in large databases	No	**Yes**	**Yes**	**Yes**
Interactive	Yes[1]	No	No	**Yes**
Customisable tooltips	No	No	No	**Yes**
Allows any mutation dictionary to be used	**Yes**	No	**Yes**	**Yes**
Automatic colour palette selection when mutation impact dictionary conforms to known ontologies	No	Yes (MAF only)	No	**Yes (MAF, SO, or PAVE)**
Approach for resolving genes with multiple mutations	Different Visualisation on Plot[2]	**Flags as Multi-Hit**	Picks more severe consequence or leaves to user[3]	**Flags as Multi-Hit**
Supports a mutation level dataset as input	No[4]	**Yes**	**Yes**	**Yes**
Native support for faceting by pathway	No	**Yes**	No	**Yes**
Supports marginal plots describing TMB, gene mutation recurrence, and clinical annotations	**Yes**	**Yes**	**Yes**	**Yes**

Figure 2: Comparison of R packages for creating oncoplots. [1]Requires the shiny and interactiveComplexHeatmap packages. [2]Requires the user to first summarise mutations at the gene level as a sample by gene character matrix with mutations separated by semicolons (wide format). [3]For MAF inputs the most severe consequence is chosen, however for non-MAF datasets users must manually define the mutation impact hierarchy. [4]Non-unique mutation types are treated as one observation, however if different mutation types affect one gene, the indiviual mutations can be plotted with different shapes/configurations in a user-configured manner.

The Journal of Open Source Software

Acknowledgements

We thank the developers of the packages integral to ggoncoplot, especially David Gohel for ggiraph (Gohel & Skintzos, 2024), which enables its interactivity, and Thomas Lin Pedersen for patchwork (Pedersen, 2024) and ggplot2 maintenance. We also acknowledge Hadley Wickham and all contributors to ggplot2 (Wickham, 2016). Additionally, we thank Dr. Marion Mateos for her insightful feedback during the development of ggoncoplot.

References

Gohel, D., & Skintzos, P. (2024). *Ggiraph: Make 'ggplot2' graphics interactive*. https://davidgohel.github.io/ggiraph/

Gu, Z. (2022). Complex heatmap visualization. *iMeta*, *1*(3), e43. https://doi.org/10.1002/imt2.43

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*, 1747–1756. https://doi.org/10.1101/gr.239244.118

Pedersen, T. L. (2024). *Patchwork: The composer of plots*. https://patchwork.data-imaginist.com

Skidmore, Z. L., Wagner, A. H., Lesurf, R., Campbell, K. M., Kunisaki, J., Griffith, O. L., & Griffith, M. (2016). GenVisR: Genomic visualizations in r. *Bioinformatics*, *32*, 3012–3014. https://doi.org/10.1093/bioinformatics/btw325

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