Bioinformatics, 33(18), 2017, 2938–2940
doi: 10.1093/bioinformatics/btx364
Advance Access Publication Date: 22 June 2017
Applications Note

Genome analysis

UpSetR: an R package for the visualization of intersecting sets and their properties

Jake R. Conway,[1] Alexander Lex[2] and Nils Gehlenborg[1,*]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA and [2]SCI Institute, School of Computing, University of Utah, Salt Lake City, UT 84112, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 25, 2017; revised on May 19, 2017; editorial decision on May 31, 2017; accepted on June 5, 2017

Abstract

Motivation: Venn and Euler diagrams are a popular yet inadequate solution for quantitative visualization of set intersections. A scalable alternative to Venn and Euler diagrams for visualizing intersecting sets and their properties is needed.
Results: We developed UpSetR, an open source R package that employs a scalable matrix-based visualization to show intersections of sets, their size, and other properties.
Availability and implementation: UpSetR is available at https://github.com/hms-dbmi/UpSetR/ and released under the MIT License. A Shiny app is available at https://gehlenborglab.shinyapps.io/upsetr/.
Contact: nils@hms.harvard.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The visualization of sets and their intersections is a common challenge for researchers who are dealing with biological and biomedical data. For example, a researcher might need to compare multiple algorithms that identify single nucleotide polymorphisms (Xu *et al.*, 2012, Supplementary Fig. S1) or show orthologs of genes in newly sequenced species across genomes of related species (D'Hont *et al.*, 2012, Supplementary Fig. S2). Although many alternative set visualization techniques exist (Alsallakh *et al.*, 2016), such data are typically visualized using Venn and Euler diagrams. Such diagrams can be generated with R packages such as *venneuler* (Wilkinson, 2012) and *VennDiagram* (Chen and Boutros, 2011). These closely related techniques have well known shortcomings, as they are hard to generate for more than a small number of sets. The visual representation of intersection size by irregularly shaped and unaligned areas makes it hard to answer essential questions such as 'What is the biggest intersection?' or 'Is intersection *X* larger than intersection *Y*?' (Cleveland and McGill, 1984).

2 Materials and methods

Here we present an R package named 'UpSetR' based on the 'UpSet' technique (Lex *et al.*, 2014; Lex and Gehlenborg, 2014) that employs a matrix-based layout to show intersections of sets and their sizes. It is implemented using ggplot2 (Wickham, 2009) and allows data analysts to easily generate generate UpSet plots for their own data. UpSetR support three input formats: (i) a table in which the rows represent elements and columns include set assignments and additional attributes; (ii) sets of elements names; and (iii) an expression describing the size of the set intersections as introduced by the *venneuler* package (Wilkinson, 2012). UpSetR provides support for the visualization of attributes associated with the elements contained in the sets, enabling researchers to explore and characterize the intersections. UpSetR differs from the original UpSet technique as it is optimized for static plots and for integration into typical bioinformatics workflows. We also provide a Shiny app that allows researchers to create publication-quality UpSet plots directly in a web browser.

UpSetR visualizes intersections of sets as a matrix in which the rows represent the sets and the columns represent their intersections (Fig. 1 and Supplementary Figs. S1 and S2 for comparisons of Venn and Euler diagrams with UpSetR plots). For each set that is part of a given intersection, a black filled circle is placed in the corresponding matrix cell. If a set is not part of the intersection, a light gray circle is shown. A vertical black line connects the topmost black circle

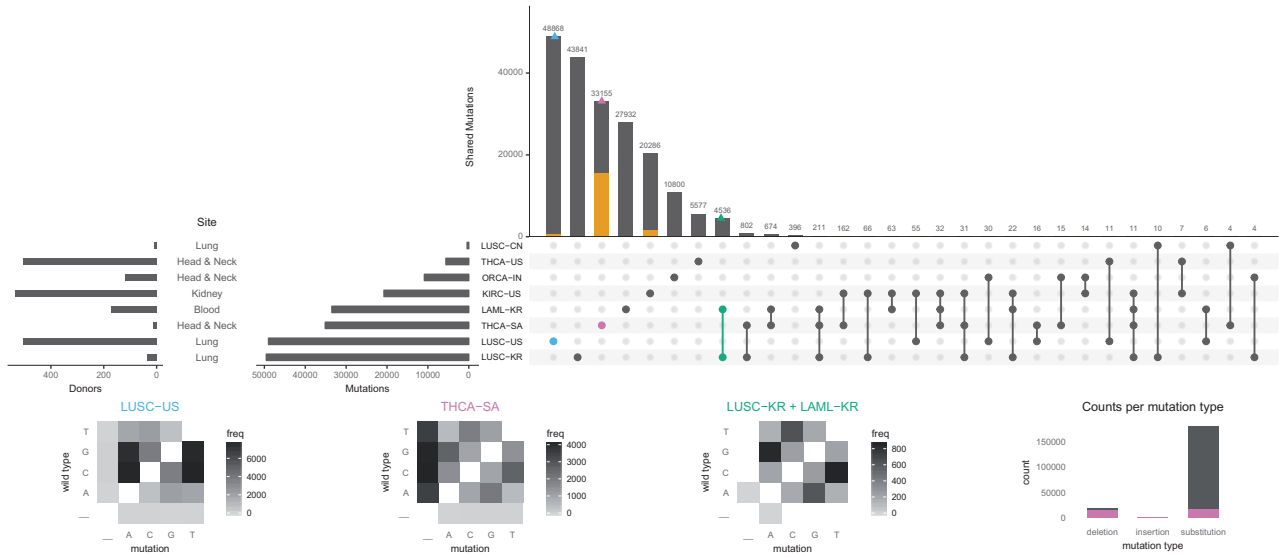

Fig. 1. An UpSetR plot of variants across eight ICGC cancer studies with three intersection queries, one element query, four attribute plots, and two set metadata plots. Data for LUSC-KR and LAML-KR are based on whole-genome sequencing, all others on whole-exome sequencing. The three intersection queries are the one-way intersections of LUSC-US (blue) and THCA-SA (purple), and the two-way intersection of LUSC-KR and LAML-KR (green). The element query (yellow) selects mutations classified as deletions. Three custom transition/transversion plots display the relative frequency of substitution events for the intersection queries. The bar plot attribute plot displays the contribution of variants unique to the THCA-SA cohort (purple) to each mutation type. Set metadata is plotted to the left of the set size bar (charts)

with the bottommost black circle in each column to emphasize the column-based relationships. The size of the intersections is shown as a bar chart placed on top of the matrix so that each column lines up with exactly one bar. A second bar chart showing the size of the each set is shown to the left of the matrix.

3 Usage scenario

To illustrate the utility and features of UpSetR, we retrieved variant calls for eight cancer studies from from the ICGC Data Portal (see Supplementary Material). Each cancer study represents a set and each variant represents an element that is contained in one or more sets (Supplementary Fig. S3). UpSetR supports queries on the data to highlight features. *Intersection queries* can be used to select subsets of elements in the dataset defined by an intersection. Queries are assigned a unique color and their results are plotted on top of the intersection size bar chart. For example, this can be used to select elements in particular intersections (Supplementary Fig. S4). Additionally, UpSetR supports queries for the selection of elements based on attributes associated with the elements in the sets. Attributes can be numerical, Boolean or categorical. In our example, element attributes are chromosome, genomic location, and variant type (deletion, insertion, substitution) associated with each variant. UpSetR *element queries* select elements across intersections and sets based on particular attribute values. Basic built-in queries can be extended to arbitrarily complex queries by providing a custom query function that operates on any combination of attributes. Element queries can be used to select variants of a particular type, such as deletions, and to view them across intersections (Supplementary Fig. S5).

UpSetR provides integration of additional *attribute plots* that visualize attributes of elements selected by an intersection or element query. Support for scatter plots and histograms is built into UpSetR. Additional plot types can be integrated by providing in a function that returns a *ggplot* object to visualize the data. When attribute or intersection queries are applied, query results can also be overlaid on attribute plots in addition to the intersection size bar plot. Figure 1 demonstrates how these features, including the visualization of metadata about the sets, can be combined into a plot that among other issues, reveals a notable over-representation of unique deletions among the variants in the THCA-SA study.

4 Conclusion

UpSetR is a highly customizable tool for data exploration and generation of set visualizations. By making UpSetR compatible with the input formats of the existing popular Venn and Euler diagram packages and by offering a Shiny web interface, we incentivize the use of UpSet diagrams and enable users without programming skills to generate effective set visualizations. Through its seamless integration with ggplot2 and its ability to apply virtually any query, it is possible to customize and explore data in ways not supported by any other set visualization package. In addition, the integration of UpSetR with ggplot2 allows developers to extend UpSetR for use in their own software packages.

Acknowledgements

We acknowledge Megan Paul for her contributions and the National Institutes of Health for funding (R00HG007583, U54HG007963, U01CA198935).

References

Alsallakh,B. *et al.* (2016) The State-of-the-Art of Set Visualization. *Computer Graphics Forum*, **35**: 234–260.

Chen,H. and Boutros,P.C. (2011) VennDiagram: a package for the generation of highly-customizable Venn & Euler diagrams in R. *BMC Bioinformatics*, **12**, 35.

Cleveland,W.S. and McGill,R. (1984) Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Am. Stat. Assoc*, **79**, 531–554.

D'Hont,A. *et al.* (2012) The banana (musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, **488**, 213–217.

Lex,A., and Gehlenborg,N. (2014) Sets & intersections. *Nat. Methods*, **11**, 779.

Lex,A. *et al.* (2014) UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, **20**, 1983–1992.

Wickham,H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

Wilkinson,L. (2012) Exact and approximate area-proportional circular Venn and Euler diagrams. *IEEE Trans. Vis. Comput. Graph.*, **18**, 321–331.

Xu,F. *et al.* (2012) A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nat. Commun.*, **3**, 1258.