Complex heatmaps reveal patterns and correlations in multidimensional genomic data

Zuguang Gu[1,2], Roland Eils[1,2,3] and Matthias Schlesner[1,*]

[1]Division of Theoretical Bioinformatics, [2]Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center (DKFZ), Heidelberg, Germany and [3]Department for Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular Biotechnology (IPMB) and BioQuant, Heidelberg University, Heidelberg, Germany

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Abstract

Summary: Parallel heatmaps with carefully designed annotation graphics are powerful for efficient visualization of patterns and relationships among high dimensional genomic data. Here we present the *ComplexHeatmap* package that provides rich functionalities for customizing heatmaps, arranging multiple parallel heatmaps and including user-defined annotation graphics. We demonstrate the power of *ComplexHeatmap* to easily reveal patterns and correlations among multiple sources of information with four real-world datasets.

Availability and Implementation: The *ComplexHeatmap* package and documentation are freely available from the Bioconductor project: http://www.bioconductor.org/packages/devel/bioc/html/ComplexHeatmap.html.

Contact: m.schlesner@dkfz.de

1 Introduction

Heatmaps are a fundamental visualization method that is broadly used to unravel patterns hidden in genomic data. They are especially popular for gene expression analysis (Eisen *et al.*, 1998) and methylation profiling (Sturm *et al.*, 2012). With the increasing availability of genomic datasets, visualization methods that effectively show relations within multidimensional data are urgently needed. In this paper, we demonstrate how heatmaps with carefully designed annotation graphics can give great enhancement for revealing underlying data structure and how utilization of parallel heatmaps which focus on different sources of information gives a quick and comprehensive overview of the data.

In the R programming environment, traditional tools for drawing heatmaps, like the basic *heatmap* function or add-on packages such as *pheatmap* or *heatmapplus*, only provide limited functionality to display annotation graphics and do not support plotting of multiple parallel heatmaps. The *ComplexHeatmap* package has been designed to overcome these limitations. It provides a general solution to juxtapose different sources of information in multiple parallel heatmaps. Each heatmap can be enhanced by multiple annotation graphics to complete the portrayal of the dataset. Both column and row annotation graphics supported in *ComplexHeatmap* can either be predefined graphics, e.g. points, bar plots or boxplots or more general user-defined graphics. Other features of *ComplexHeatmap* include: (i) flexible support for clustering. For example, rendered dendrograms (Galili, 2015) or a user-defined distance function that accepts two paired vectors are supported; (ii) separating of heatmap rows into slices to support visualization of subgroups, where splitting on rows can be done either by a partitioning method, e.g. *k*-means clustering or a data frame that contains classifications; (iii) user-customization of the heatmap grids for more advanced visualization of complex information, e.g. the enhanced OncoPrint (Gao *et al.*, 2013) in Figure 1A ; (iv) interactive selection on heatmaps to obtain subset of rows and columns if heatmaps are drawn on an interactive device (e.g. X11) and (v) the ability to add more customized graphics after heatmaps are generated.

ComplexHeatmap has a modular design with a user-friendly application programming interface (API). The flexibility and

extensibility of the package enables fast and easy generation of novel views of multidimensional datasets and greatly facilitates discoveries from genomic and other omics data.

2 Implementation

ComplexHeatmap is implemented in an object-oriented way. The main classes are:

Heatmap: Representation of a single heatmap. The class processes the data matrix and provides methods for drawing heatmap components, i.e. the heatmap body, row/column names, titles, dendrograms and column annotations.

HeatmapList: Representation of a list of heatmaps. The class adjusts graphical settings for the list of heatmaps, creates the layout and provides methods for drawing components like the heatmap legends. The '+' operator is used to concatenate parallel heatmaps (i.e. heatmaps in which corresponding rows in all matrices corresponds to the same object):

```
Heatmap(matrix1, ...) + Heatmap(matrix2, ...)
```

HeatmapAnnotation: Heatmap annotation is a general and flexible concept. The only requirement for heatmap annotations is that the graphics should be aligned to columns or rows in the heatmap, respectively. The most commonly used annotation graphics are a list of grids that show different groups of the data. However, other types of graphics can also be used, e.g. it can be boxplots which visualize data distribution in corresponding rows or columns. *ComplexHeatmap* provides several fixed types of annotation graphics, e.g. points, bar plots and boxplots. The package also provides an API that allows users to design their own annotation graphics.

Column annotations and row annotations are both encapsulated by *HeatmapAnnotation* class. Column annotations are components of a single heatmap:

```
ha = HeatmapAnnotation(...)
Heatmap(matrix, top_annotation = ha, ...)
```

Row annotations are concatenated to the heatmap list by the '+' operator and can correspond to multiple parallel heatmaps:

```
ha = HeatmapAnnotation(..., which = 'row')
Heatmap(matrix, ...) + ha
```

There is no limitation for the amount and order of heatmap annotations in the list of heatmaps.

3 Application

Figure 1A visualizes genomic alterations for 38 selected genes in 134 patients from the TCGA lung adenocarcinoma cohort (Collisson *et al.*, 2014) as an enhanced OncoPrint. Genes are split into two groups based on the amplification rate among patients. Bar plots are added to rows and columns to show the numbers of alterations across patients and genes, respectively. An additional heatmap on the right indicates the biological functions of the mutated genes. The combination of OncoPrint and this heatmap reveals that highly amplified genes are relatively enriched in cellular response processes, while mutated genes are enriched in processes related to development and metabolism.

Figure 1B illustrates heterogeneity of mouse T-cells analyzed by single cell RNA-Seq data (Buettner *et al.*, 2015). Expression profiles of 721 selected genes classify the cells into two subpopulations. The left subpopulation (highlighted in light red in the column dendrogram) is characterized by relatively high expression of a subset of cell cycle genes, however, the absolute expression level is low. The subpopulation on the right side exhibits relatively low expression of this subset of cell cycle genes, while the other selected genes, including other cell cycle genes, are expressed at high levels. A group of ribonucleoprotein genes show very strong coexpression. Gene names of highly expressed cell cycle genes are indicated. The heatmap on the right visualizes pairwise correlations between the 721 genes. The correlation values were hierarchically clustered and the resulting row order used to define the row order of all parallel heatmaps.

Supplementary File S3 visualizes correlations between methylation, gene expression, enhancers and gene-related information. The data is randomly generated based on patterns found in an unpublished work. In the heatmaps, each row corresponds to a differentially methylated region (DMR) or other objects that are associated with the DMR. These objects are, for example, the nearest gene with expression negatively correlating to the methylation level in the associated DMR, or enhancers that overlap with the DMR. The complex heatmaps reveal that highly methylated DMRs are enriched in intergenic and intragenic regions and rarely overlap with

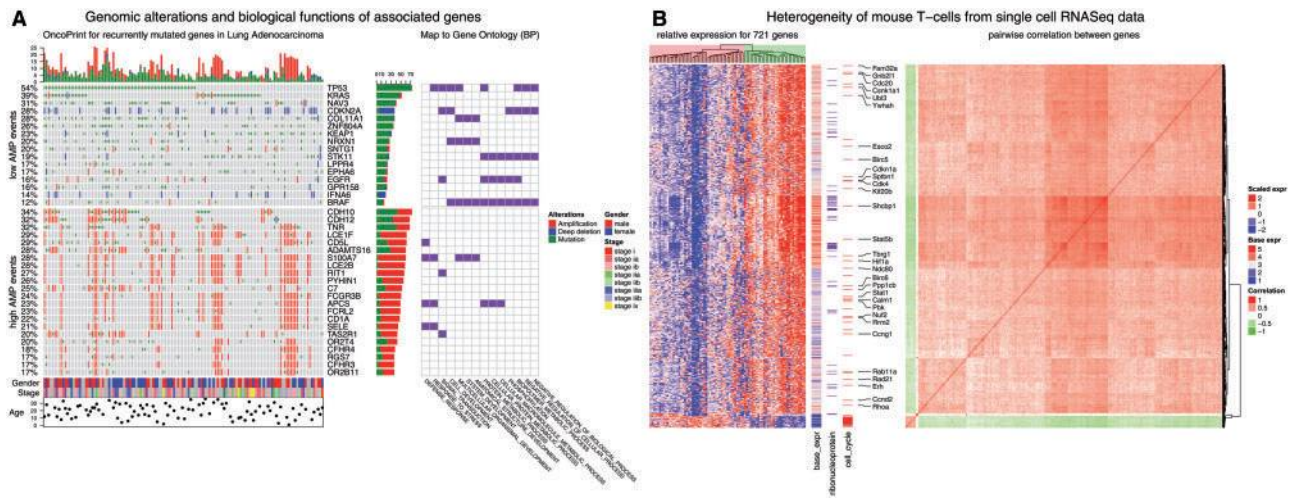

Fig. 1. Examples by *ComplexHeatmap*. (**A**) Genomic alterations and biological functions of the associated mutated genes. (**B**) Heterogeneity of mouse T-cells from single cell RNA-Seq data. Data sources and R code for the plots can be found at Supplementary S1 and S2

enhancers. In contrast, lowly methylated DMRs are enriched for transcription start sites (TSS) and enhancers.

Supplementary File S4 reimplements Figure 1 in Sturm *et al.* (2012) to demonstrate the ability of *ComplexHeamtap* to make complex annotations. Compared to the original figure, two new heatmaps are added, one visualizing the distance between CpG sites and the nearest TSS, and the second visualizing annotations to CpG Islands (CGI). The heatmap is split according to the CGI annotation, revealing that the CpG sites that are in CGI Shelf and open sea have higher methylation levels and higher distance to the nearest TSS.

Funding

This work was supported by the German Cancer Research Center-Heidelberg Center for Personalized Oncology (DKFZ-HIPO) and the BMBF-funded de.NBI HD-HuB network (#031A537A, #031A537C).

Conflict of Interest: none declared.

References

Buettner,F. *et al.* (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.

Collisson,E.A. *et al.* (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci*, **95**, 14863–14868.

Galili,T. (2015) dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*, **31**, 3718–3720.

Gao,J. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.

Sturm,D. *et al.* (2012) Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell*, **22**, 425–437.