BioTech 76

Creating Plots, Graph and Maps using R

Xin Tian, Ph.D., Mathematical Statistician

Office of Biostatistics Research
National Heart, lung and Blood Institute, NIH
Email: tianx@nhlbi.nih.gov

GitHub: **https://github.com/npmldabook/rplots**

Agenda (break 15 mins *2, lunch 12-1 PM)

- A : R basics
- B : Graphics systems in R
- **C : Using color in R**
- **D : R Graphics Devices – Static display**
- **E: R Graphics Devices – Interactive display**
- **F: Working with Maps**
- **H: Specialty Figures**
- **G: Network plots**
- Q & A, hands on your own research data

htmlwidgets for R

https://www.htmlwidgets.org/

Dygraphs: provides rich facilities for charting time-series data in R and includes support for many interactive features including series/point highlighting, zooming, etc.

Plotly, ggplotly, heatmaply: allows you to easily translate your ggplot2 graphics to an interactive web-based version.

https://images.plot.ly/plotly-documentation/images/r_cheat_sheet.pdf

Image can save as webpage (html) format to open from browser

Websites: https://www.r-graph-gallery.com/ (browse and try the code)

F.

- Leaflet is one of the most popular open-source JavaScript libraries for interactive maps. It's used by websites ranging from The New York Times and The Washington Post to GitHub and Flickr
- This R package makes it easy to integrate and control Leaflet maps in R.
- **Features (***layers)* : **Interactive panning/zooming**

A) leaflet() returns a Leaflet map widget,

B) Compose maps using arbitrary combinations of: %>% (pipe operator)
  - Map tiles
  - Markers
  - Polygons
  - Lines
  - Popups , etc

Create maps right from the R console or Rstudio, Embed maps in R Markdown docs and Shiny apps

- See website for lots of examples: https://rstudio.github.io/leaflet/

H. Specialty Figures: **Word Cloud**

1. **Word Cloud**: text mining that highlights the most frequently used words in a text as a visual representation of text data.

The text mining package (***tm***) and the word cloud generator package (***wordcloud***)

http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know

The 5 main steps to create word clouds in R

- Step 1: Create a text file
- Step 2 : Install and load the required packages
- Step 3 : Text mining
- Step 4 : Build a term-document matrix for frequency table of words
- Step 5 : Generate the Word cloud

Principal Component Analysis (PCA)

Dimensionality reduction p → q (2-3 dim) for visualization, clustering (unknown class) and linear discriminative analysis (when the class is known).

https://tgmstat.wordpress.com/2013/11/21/introduction-to-principal-component-analysis-pca/

Assume you have n observations of p different variables. Define X to be a $(n \times p)$ matrix where the i-th column of X contains the observations of the i-th variable, $i = 1, ..., p$. Each row x_i of X can be represented as a point in a p-dimensional space. Therefore, X contains n points in a p-dimensional space.

PCA projects p-dimensional data into a q-dimensional sub-space $(q \leq p)$ in a way that If we pick the first q principal components, we have projected our p-dimensional data into a q-dimensional sub-space. We can define R^2 in this context to be the fraction of the original variance kept by the projected points,

$$R^2 = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

Genomic plots -ggbio

-Bioconductor packages, not from CRAN

http://bioconductor.org/packages/release/bioc/html/ggbio.html

```
install.packages("BiocManager")
BiocManager::install("ggbio")
```

-Read pdf plot manual in Github: page 32-41 for circular plot

The ggbio package

- Build on ggplot2: Extends and specializes the grammar of graphics for biological data.
- A programmable genome browser environment: Visualization tool for plotting different types of genomic data in separate tracks along chromosomes
- The graphics are designed to answer common genomics data questions. Most core Bioconductor data structures are supported; work well with Granges object.
- Circular genome plots : 1) Visualize somatic mutation 2) Visualize inter, intra-chromosome rearrangement; 3) Visualize mutation score as point tracks; 4) add scales/ticks/labels ; 5) comparison of multiple samples. *Add them one by one, it will be automatically created from inner circle to outside.*

Goal

ggbio is to make it easy to make the common genomic plots once you have your data in `GRanges` objects and other objects from GenomicRanges.

Some plots

- Manhattan plot (SNPs)
- Ideograms
- Tracks (emulate a genome browser)
- Circular: good for re-arrangements

How it works

- The syntax is similar to ggplot2 as ggbio builds on top of it.
- Plotting functions return `ggplot2` objects which you can then modify using `ggplot2` code.
- `ggbio` figures out how to align all your data in the genome axis for you.

Manhattan plot

A **Manhattan plot** is a type of scatter plot, usually used to display data with a large number of data-points - many of non-zero amplitude, and with a distribution of higher-magnitude values, for instance in genome-wide association studies (GWAS). In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, Y-axis= -log10(P-value), for each single nucleotide polymorphism (SNP) displayed on the Y-axis, meaning that each dot on the Manhattan plot signifies a SNP.

It gains its name from the similarity of such a plot to the Manhattan skyline: a profile of skyscrapers towering above the lower level "buildings" which vary around a lower height.

Network plots (http://kateto.net/networks-r-igraph)