ggplot and ColorBrewer
Nice plots with R

November 30, 2015
ggplot

ggplot2

ggplot is an advanced plotting system for R. The current version is ggplot2 (http://ggplot2.org/). It is not in the base installation of R, so you have to install it using install.packages("ggplot2"). You can find many tutorials and help pages for ggplot2 using google! Use them!
Other than the common plot command in the R base, ggplot uses multiple functions to create one plot. The functions are concatenated using the ’+’-operator.

First, you need to provide the **what** to plot

- ggplot accepts as first argument an data frame. Use column names to specify the content of the column.
- The **aes** function (aesthetics) maps aesthetics values onto the different options for the plot. It is used in many different ggplot functions. In combination with ggplot, it is usually used to specify what goes onto the x- and y-axes.

```r
ggplot(data,aes(x,y))
```
When using the standard plot commands, you need to specify either the type (lines, points, etc.) of the plot or you use a specific plot command (boxplot, barplot, etc.). All of them have different options and names for the arguments.
ggplot tries to unify this. There are several `geom_XXX` functions. Depending which of them is used, the a scatter plot (`geom_point`) or a barplot (`geom_bar`) is created. However, the arguments are very similar across the different functions. One argument is e.g. the color or for fillable plot object the color to fill the object (`fill=color`).
ggplot(data,aes(x,y)) + geom_point(color="red")
Where to plot

When the data frame contains data for multiple plots, ggplots allows to arrange them automatically. Let’s assume the data frame contains data about the final grade of students at the university for multiple years, study area, levels (master and bachelor). We want to plot the grade distribution as boxplots for master and bachelor. We want each year in a row and each study area in a column. Furthermore, the fill color should correspond to the degree (bachelor, master). With ggplot, this is very easy using facets. They allow you to arrange your plots.
Assume the data frame has the columns 'grade', 'year', 'area', 'degree'.

```r
ggplot(data, aes(x=degree,y=grade)) + geom_boxplot(fill=degree) + facet_grid(area ~ year)
```
Where to plot

Instead of facet_grid which makes a grid with columns and row, you can use facet_wrap(variable) to arrange the plots in groups defined by 'variable'. The plots will be arrange in the a way that they fit best.

```r
ggplot(data, aes(x=degree,y=grade)) + geom_boxplot(fill=degree) + facet_grid(year)
```
Themes define how the plot looks. For example, if you want to change the default behavior for the axis, you use the themes function. The following plot parts can be changed for example with the themes function.

- **plot.title**: how the plot title look like. The `element_text` function can be used to set size, font, face, and spacing. For example:

  ```r
  ggplot(data,aes(x,y)) +
  geom_point(color="red") +
  themes(plot.title=element_text(size=12,face='''bold'''))
  ```

- **axis.text.x** and **axis.text.y** or **axis.text** for the marks at axis
- **axis.ticks.x** and **axis.ticks.y** or **axis.ticks** for the tick marks
- **axis.title.x** and **axis.title.y** or **axis.title** for the axis title
More on ggplot

There are many more features in ggplot which allow you to customize your R plot. You can find the documentation in the internet. If you are looking for some specific use google and add the key words 'R' and 'ggplot'.

Useful pages:

▶ http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/#quicksetup-the-dataset
▶ http://www.statmethods.net/advgraphs/ggplot2.html
▶ https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
ColorBrewer

If you make plots with colors, be sure that the choice of color is suited for your purpose. You may want to have printer-friendly colors or colorblind safe colors. Be aware of the fact that your color choice will influence how other people interpret your plot. Be aware that a human eye cannot distinguish more than 12 colors when the objects which should be compared are too far away. Ask yourself:

▶ Which type of data represents the color: sequential, diverging, qualitative?
▶ How many colors do I really need
▶ What is my background color
▶ do they have to be colorblind safe, printer friendly or photocopy safe?

The webpage: http://colorbrewer2.org/ will assist you in finding the best set of colors for you purpose. The corresponding R package is called RColorBrewer.
Practice time

Get familiar with ggplot2 and RColorBrewer by replotting the plots for the gene expression analysis. Use the gene expression data set or the lymphoma data set from last week. Make nice plots with ggplot for the corresponding data sets.

For example:

- Boxplots for the read counts before and after normalization using DESeq or EdgeR for the chimp and human and bowtie and segemehl.
- Plot the p-value distribution (with and without adjustment) for chimp and human in brain and B-cells.
- Compare the brain data and B-cell data and depict the result in appropriated way. (Do they have the same number of diff. exp. genes? Are the fold enrichment distributions different? Do the intensity distributions differ across the samples?)
More Practice

There are many datasets directly in R available. You can use them to get practice with statistical analysis or as toy data sets for, e.g., ggplot.

Have a look at https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html and choose one dataset. Use descriptive statistics to get an idea what is in the data. Generate the corresponding plots unsing ggpolt. If you still have time, use more than one dataset.