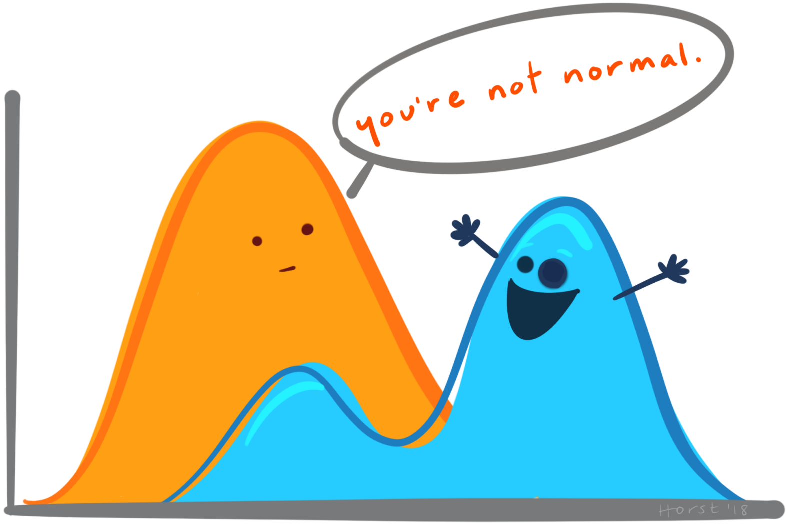
COMPASS

Community Platform for Agricultural Sciences

Visualizing biological data



Viviana Ortiz · Paulo Izquierdo

20 Feb 2021

Artwork by Allison Horst,
some slides modified from Jessica Minnier · Meike Niederhausen. bit.ly/berd_ggplot

Did you get the dataset “gene_expr_pasilla_results.csv” and the “VizBiological.Rmd” file?

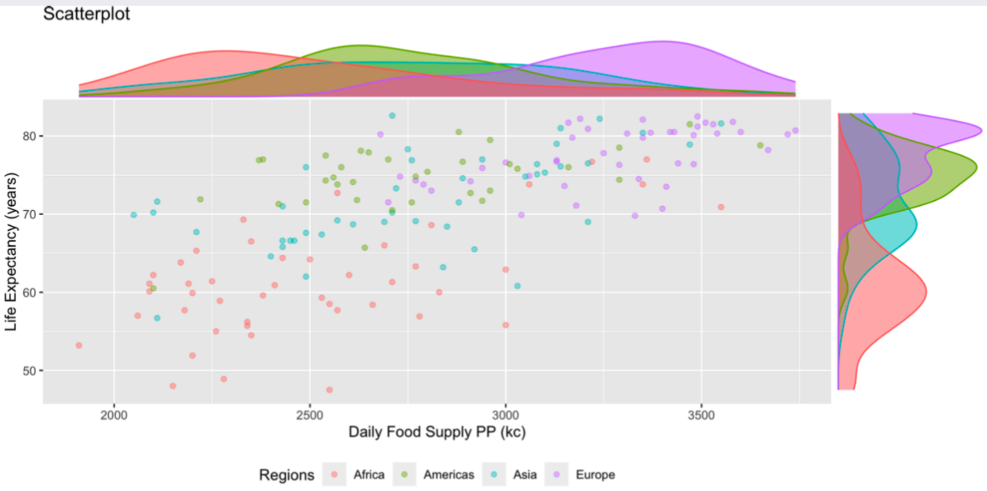
✋ 👍

ggmarginal

```
# library(ggExtra)

p <- ggplot(data = gapminder2011,
        aes(x = FoodSupplykcPPD,
            y = LifeExpectancyYrs,
            color = four_regions)
        ) +
  geom_point(alpha = .4) +
  scale_color_discrete(
    name = "Regions",
    labels = c("Africa", "Americas",
               "Asia", "Europe")
    ) +
  theme(legend.position="bottom") +
  labs(
    x = "Daily Food Supply PP (kc)",
    y = "Life Expectancy (years)",
    title = "Scatterplot"
    )
```

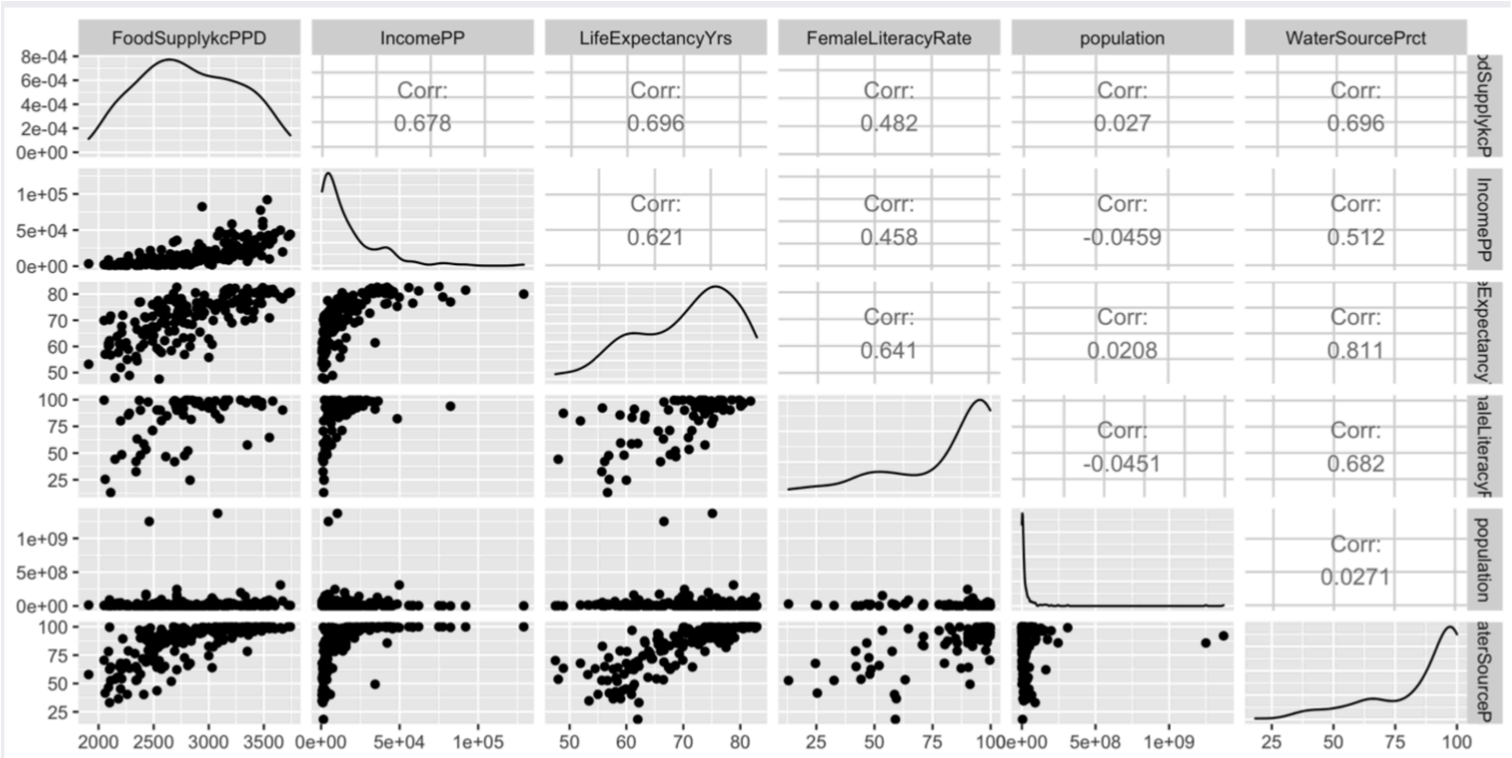
```
ggMarginal(p,
  type = "density",
  margins = "both",
  groupColour = TRUE,
  groupFill = TRUE
)
```



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

GGally::ggpairs()

```
# library(GGally)
gapminder2011 %>%
  select(FoodSupplykcPPD:WaterSourcePrct) %>% # specifying which columns to use
  ggpairs()
```



https://ggobi.github.io/ggally/index.html

ggtree()

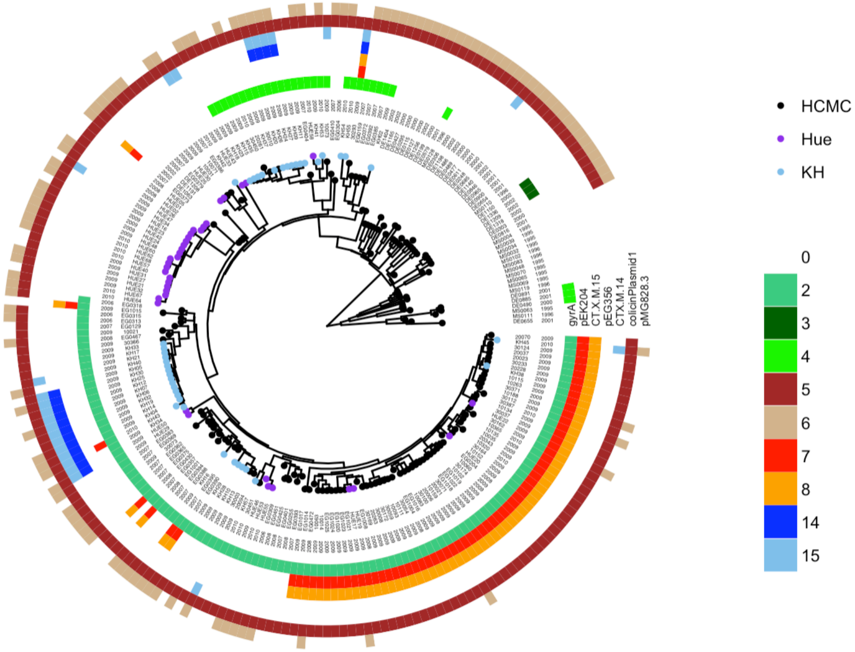
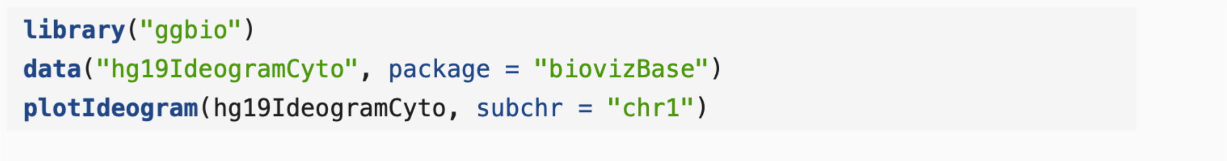


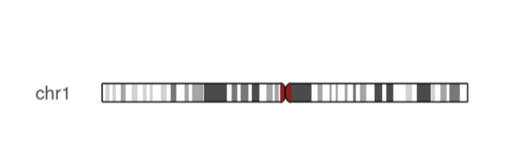
Figure 4.17: **Example of annotating a tree with diverse associated data.** Circle symbols are colored by strain sampling location. Taxa names and sampling years are aligned to the tips. Curated gene information were visualized as a heatmap (colored boxed on the outer circles).

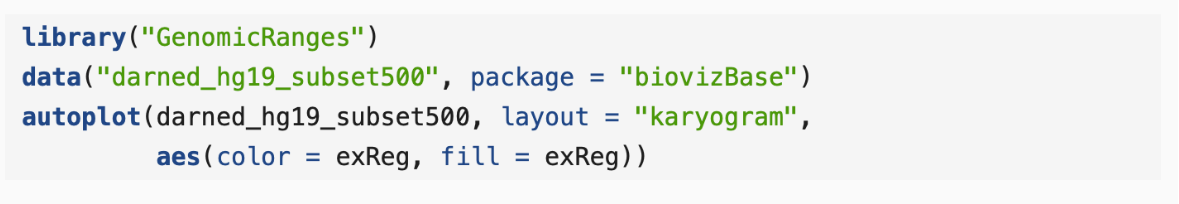
https://guangchuangyu.github.io/ggtree-book/chapter-ggtree.html

Genomic data with `ggbio`



```
library("ggbio")
data("hg19IdeogramCyto", package = "biovizBase")
plotIdeogram(hg19IdeogramCyto, subchr = "chr1")
```





```
library("GenomicRanges")
data("darned_hg19_subset500", package = "biovizBase")
autoplot(darned_hg19_subset500, layout = "karyogram",
         aes(color = exReg, fill = exReg))
```

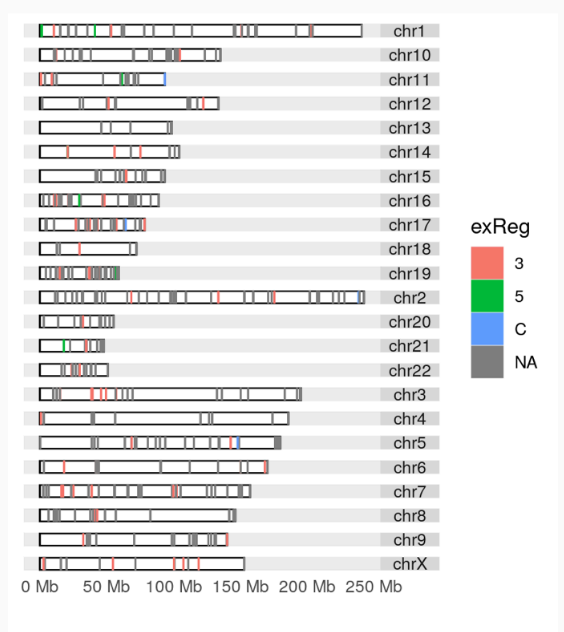
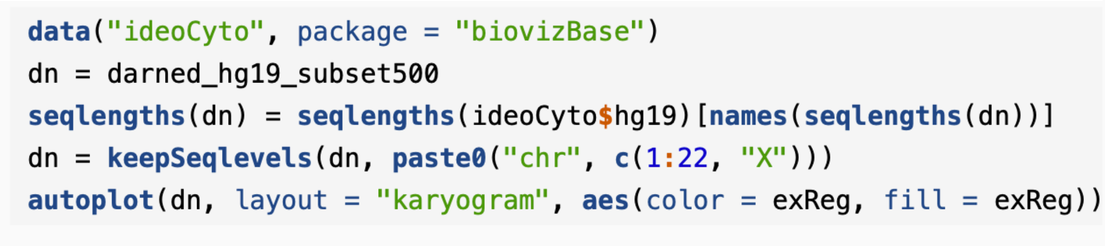


Figure 3.46: Karyogram with RNA editing sites. exReg indicates whether a site is in the coding region (C), 3’- or 5’-UTR.

http://127.0.0.1:12631/library/ggbio/doc/ggbio.pdf

https://web.stanford.edu/class/bios221/book/Chap-Graphics.html

Genomic data with ggbio



```
data("ideoCyto", package = "biovizBase")
dn = darned_hg19_subset500
seqlengths(dn) = seqlengths(ideoCyto$hg19)[names(seqlengths(dn))]
dn = keepSeqlevels(dn, paste0("chr", c(1:22, "X")))
autoplot(dn, layout = "karyogram", aes(color = exReg, fill = exReg))
```

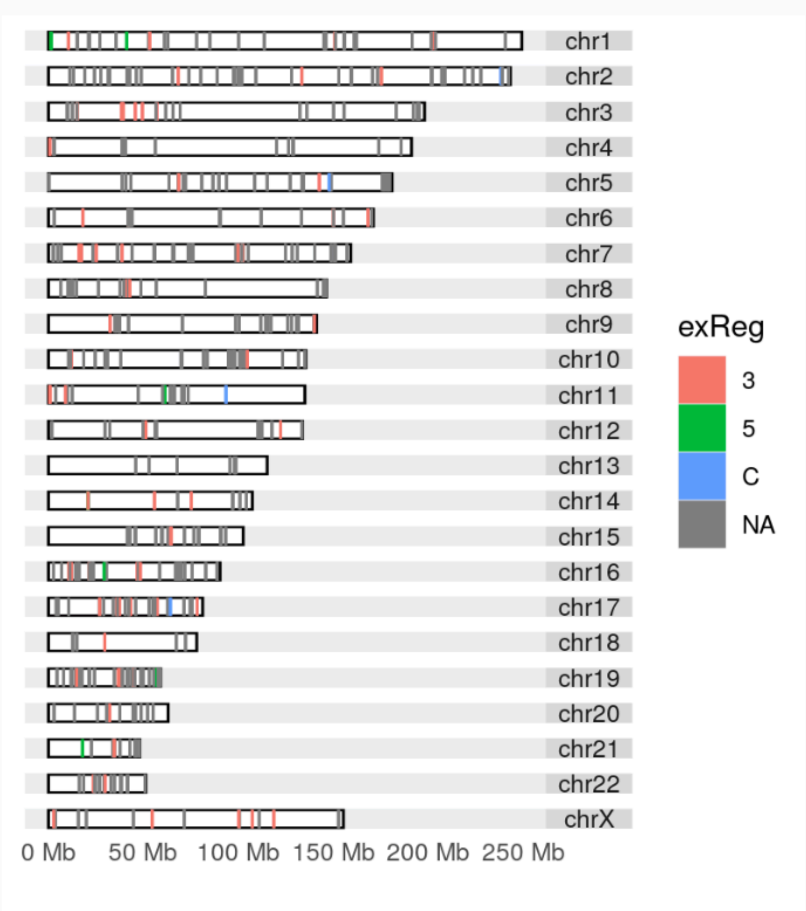


Figure 3.47: Improved version of Figure 3.46.

http://127.0.0.1:12631/library/ggbio/doc/ggbio.pdf

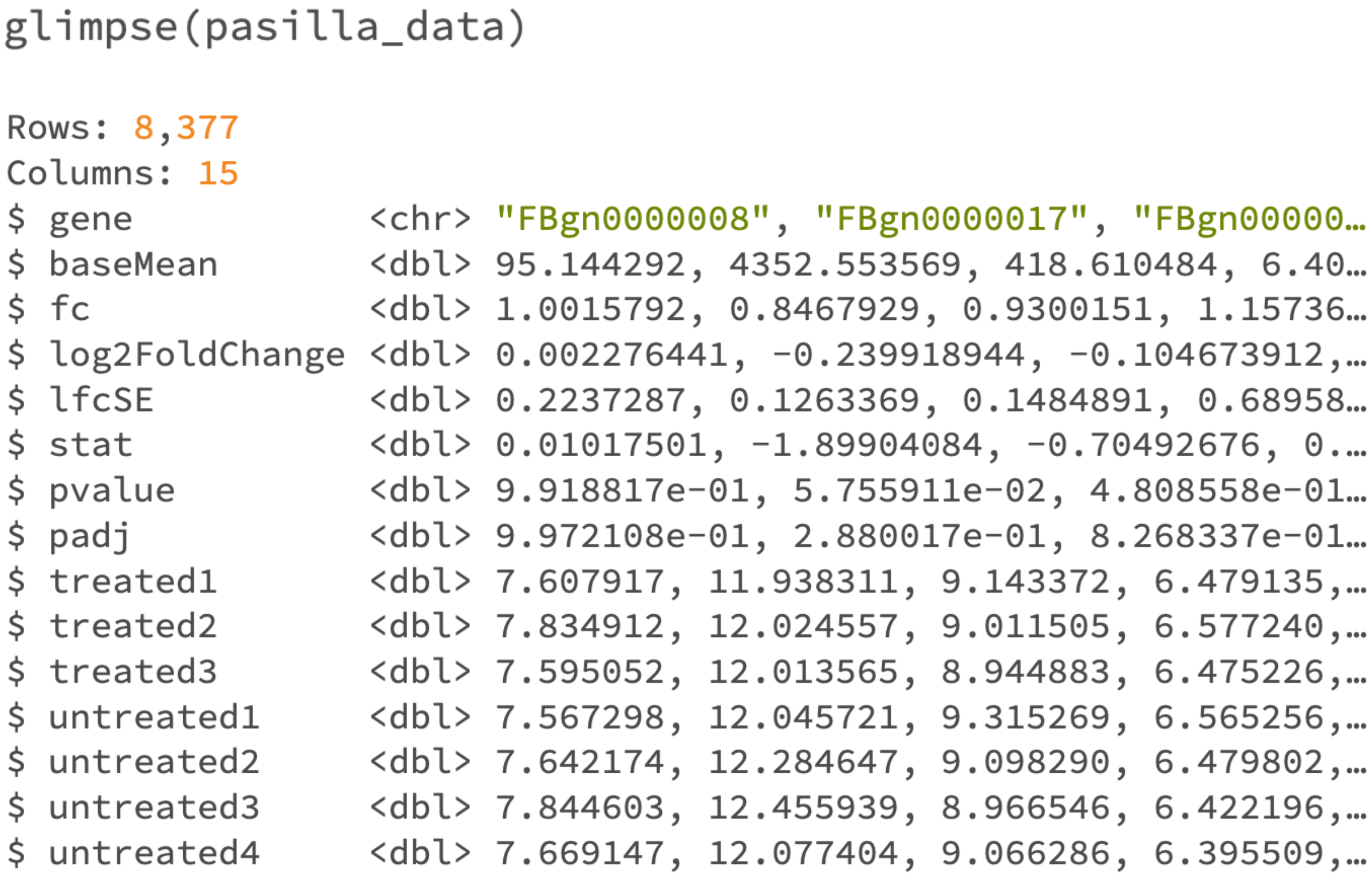
https://web.stanford.edu/class/bios221/book/Chap-Graphics.html

Gene expression

```
pasilla_data <-
read_csv("../data/gene_expr_pasilla_result
s.csv")
```

Gene expression

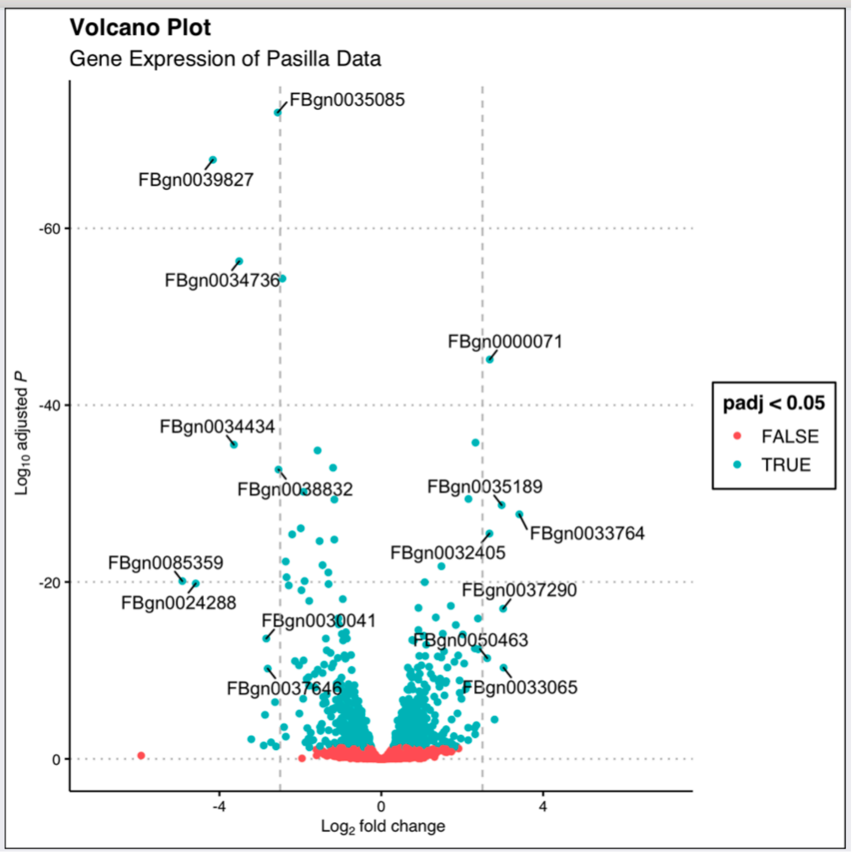
Pasilla data



```
glimpse(pasilla_data)
```

```
Rows: 8,377
Columns: 15
$ gene           <chr> "FBgn0000008", "FBgn0000017", "FBgn00000…
$ baseMean       <dbl> 95.144292, 4352.553569, 418.610484, 6.40…
$ fc             <dbl> 1.0015792, 0.8467929, 0.9300151, 1.15736…
$ log2FoldChange <dbl> 0.002276441, -0.239918944, -0.104673912,…
$ lfcSE          <dbl> 0.2237287, 0.1263369, 0.1484891, 0.68958…
$ stat           <dbl> 0.01017501, -1.89904084, -0.70492676, 0.…
$ pvalue         <dbl> 9.918817e-01, 5.755911e-02, 4.808558e-01…
$ padj           <dbl> 9.972108e-01, 2.880017e-01, 8.268337e-01…
$ treated1       <dbl> 7.607917, 11.938311, 9.143372, 6.479135,…
$ treated2       <dbl> 7.834912, 12.024557, 9.011505, 6.577240,…
$ treated3       <dbl> 7.595052, 12.013565, 8.944883, 6.475226,…
$ untreated1     <dbl> 7.567298, 12.045721, 9.315269, 6.565256,…
$ untreated2     <dbl> 7.642174, 12.284647, 9.098290, 6.479802,…
$ untreated3     <dbl> 7.844603, 12.455939, 8.966546, 6.422196,…
$ untreated4     <dbl> 7.669147, 12.077404, 9.066286, 6.395509,…
```

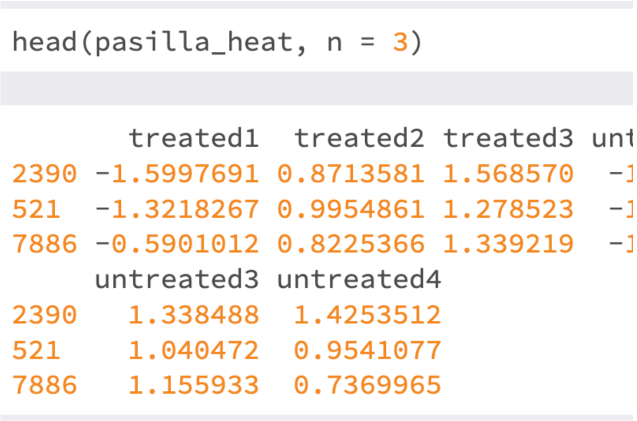
```
ggplot(data = pasilla_data,
       aes(x = log2FoldChange,
           y = log10(padj))) +
  geom_point() +
  scale_y_reverse() +
  aes(color = padj < 0.05) +
  ggrepel::geom_text_repel(
    data = pasilla_data_top,
    aes(label = gene), color = "black",
    box.padding = 0.5,
    min.segment.length = 0) +
  xlim(c(-7,7)) +
  geom_vline(xintercept = c(-2.5, 2.5),
             lty = "dashed", color="grey") +
  ggthemes::theme_clean() +
  labs(
    x = bquote(~Log[2]~ "fold change"),
    y = bquote(~Log[10]~adjusted~italic(P)),
    title = "Volcano Plot",
  subtitle="Gene Expression of Pasilla Data"
  )
```



Heatmap with `pheatmap::pheatmap()`

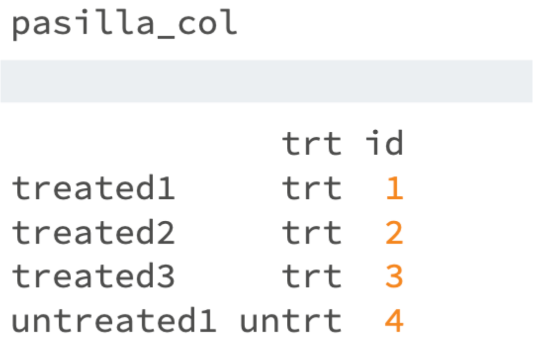
```
# select expression data
pasilla_heat <- pasilla_data %>%
  select(treated1:untreated4)
# subtract off gene-specific means
pasilla_heat <- pasilla_heat - rowMeans(pasilla_heat)
# calculate standard deviation of each centered gened
sd_gene <- apply(pasilla_heat,1,sd)
# select top 500 most variable
pasilla_heat <-
  pasilla_heat[order(sd_gene, decreasing = TRUE)[1:500],]

# create annotation data
pasilla_col <- data.frame(
  trt = factor(c(rep("trt",3), rep("untrt",4))),
  id = 1:7,
  row.names=colnames(pasilla_heat))
```



```
head(pasilla_heat, n = 3)
```

```
       treated1  treated2 treated3 unt
2390 -1.5997691 0.8713581 1.568570  -1
521  -1.3218267 0.9954861 1.278523  -1
7886 -0.5901012 0.8225366 1.339219  -1
     untreated3 untreated4
2390   1.338488  1.4253512
521    1.040472  0.9541077
7886   1.155933  0.7369965
```

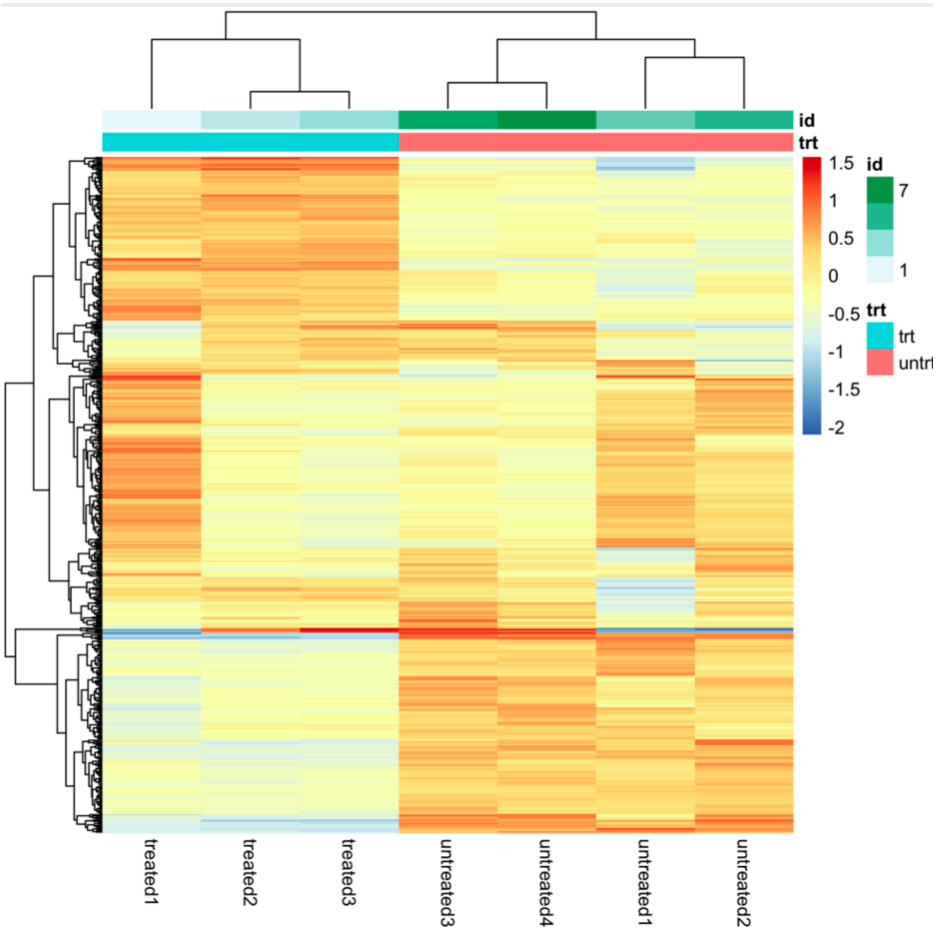


```
pasilla_col
```

```
              trt id
treated1      trt  1
treated2      trt  2
treated3      trt  3
untreated1 untrt   4
```

Heatmap with `pheatmap::pheatmap()`

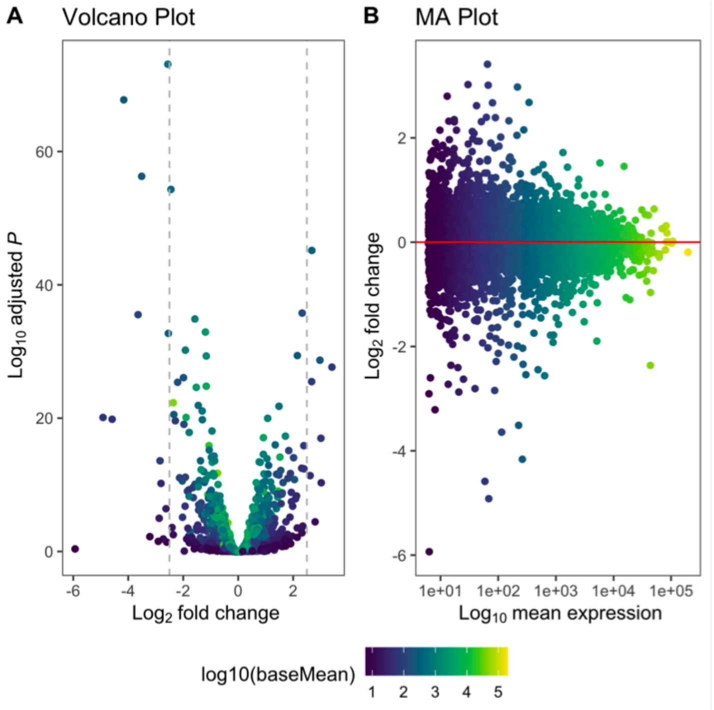
```
pheatmap::pheatmap(
  mat = pasilla_heat,
  show_rownames = FALSE,
  annotation_col = pasilla_col
)
```



Side by side plot with `ggpubr`

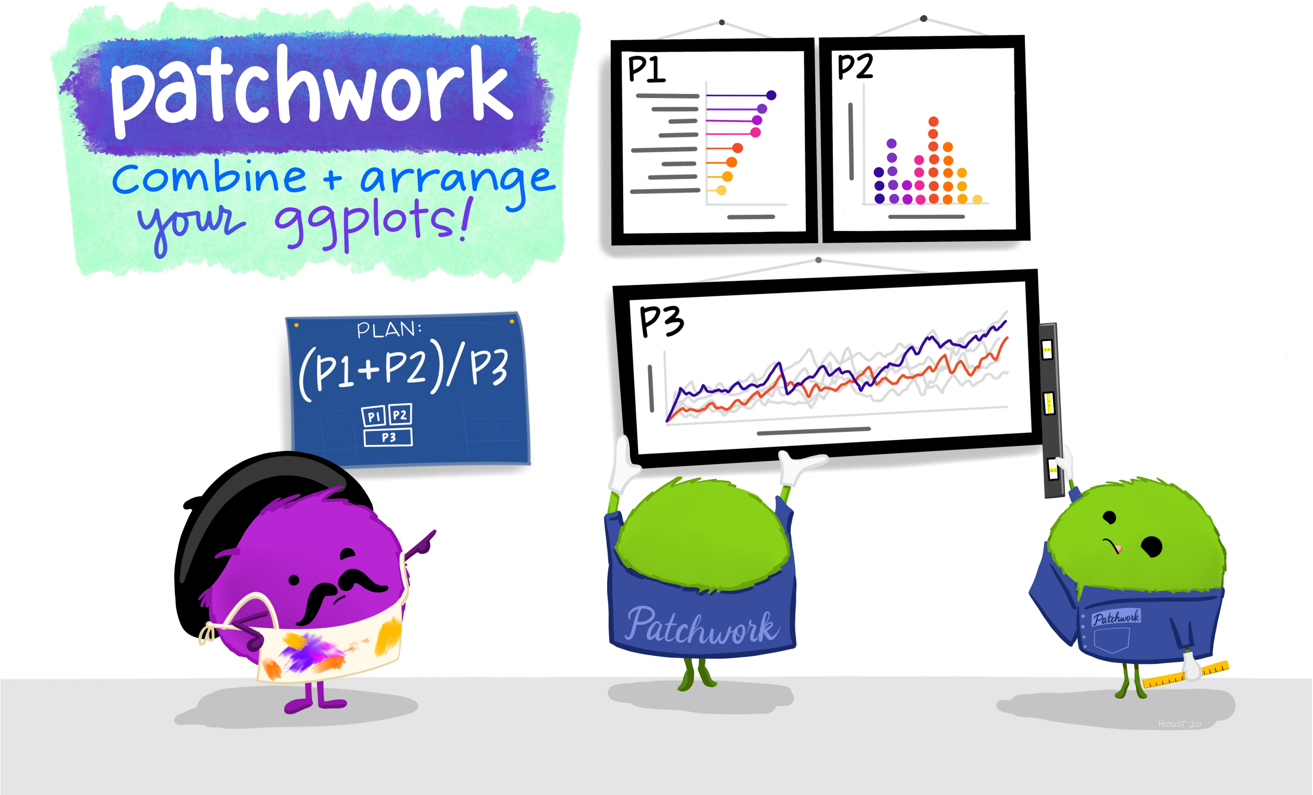
```
p1 <- ggplot(data = pasilla_data,
       aes(x = log2FoldChange,
           y = -log10(padj),
           color = log10(baseMean))) +
  geom_point() +
  geom_vline(xintercept = c(-2.5, 2.5),
             lty = 2, color="grey") +
  theme_few() + scale_color_viridis_c() +
  labs(x = bquote(~Log[2]~ "fold change"),
       y = bquote(~Log[10]~adjusted~italic(P)),
       title = "Volcano Plot")
p2 <- ggplot(data = pasilla_data,
       aes(x = baseMean,
           y = log2FoldChange,
           color = log10(baseMean))) +
  geom_point() +
  scale_x_log10() +
  geom_hline(yintercept = 0, color = "red") +
  theme_few() + scale_color_viridis_c() +
  labs(y = bquote(~Log[2]~ "fold change"),
       x = bquote(~Log[10]~ "mean expression"),
       title = "MA Plot")
```

```
ggpubr::ggarrange(p1, p2, labels = "AUTO",
  common.legend = TRUE, legend = "bottom")
```



Other options:

cowplot (https://wilkelab.org/cowplot/articles/index.html) and patchwork (https://github.com/thomasp85/patchwork)



Resources

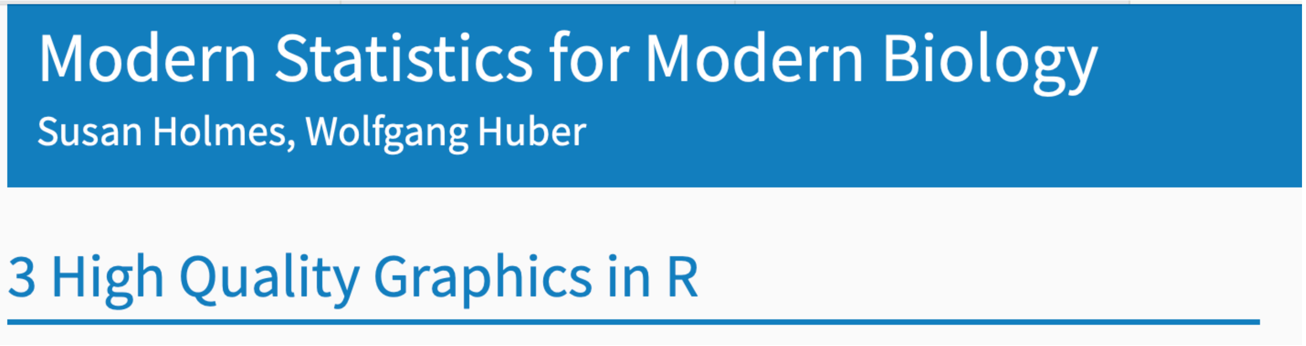
ggbio: visualization toolkits for genomic data

Tengfei Yin[1]

[1] tengfei.yin@sbgenomics.com

April 27, 2020

http://127.0.0.1:12631/library/ggbio/doc/ggbio.pdf



Modern Statistics for Modern Biology

Susan Holmes, Wolfgang Huber

3 High Quality Graphics in R

https://web.stanford.edu/class/bios221/book/Chap-Graphics.html

Inspiration and slides for this talk
Thanks!

Open content & slides:

- Jessica Minnier · Meike Niederhausen. bit.ly/berd_ggplot

Artwork by Allison Horst