INTRO TO VISUALIZATION
IN RNA SEQ EXPERIMENTS

TREx - Faraz Ahmed
A WORLD OF POSSIBILITIES

- Violin
- Density
- Histogram
- Boxplot
- Ridgeline
- Scatter
- Heatmap
- Correlogram
- Bubble
- Connected scatter
- Density 2d
- Barplot
- Spider / Radar
- Wordcloud
- Parallel
- Lollipop
- Circular Barplot
- Treemap
- Venn diagram
- Doughnut
- Pie chart
- Dendrogram
- Circular packing
- Sunburst
A WORLD OF POSSIBILITIES

Venn diagram
Doughnut
Pie chart
Dendrogram
Circular packing
Sunburst
Line plot
Area
Stacked area
Streamchart
Map
Choropleth
Hexbin map
Cartogram
Connection
Bubble map
Chord diagram
Network
Sankey
Arc diagram
Edge bundling
WHAT IS MY GOAL?

WHICH *PLATFORM* DO I USE TO GENERATE MY PLOTS 🤔

- DATA EXPLORATION?
- DATA SUMMARY?
WHICH **PLATFORM** DO I USE TO GENERATE MY PLOTS 🤔

Command Line

OR

Graphical User Interface (GUI)
WHICH **PLATFORM** DO I USE TO GENERATE MY PLOTS 🤔

- **Command Line:**
  - R/RStudio (ggplot2, reshape, plotly, viridis)
  - Python (matplotlib, plotly)
  - jQuery

- **GUI’s:**
  - JMP Pro
  - GraphPad Prism
  - R-shiny
WHAT IS MY GOAL?

- DATA EXPLORATION?
- DATA SUMMARY?
DATA EXPLORATION

• GENERALLY MEANS, PLOTTING ALL DATA POINTS;

• FINDING PATTERNS;
Data can either be **discrete** or **continuous**

**Discrete data:**

Can only take particular values

Each value is distinct (up to $\infty$) - NO Grey Area

Can be numeric -- like numbers of DE genes

but it can also be categorical -- like case or control, or male or female, or WT or KO.
DATA can either be **DISCRETE** or **CONTINUOUS**

**Continuous data:**

*Not restricted to defined separate values*

*Can take any value over a continuous range*

*EX: normalized expression of a sequenced gene*
RStudio, comes pre-loaded with example data-sets.
## DATA EXPLORATION

### Diamonds Data Set

```r
> str(diamonds)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame': 53940 obs. of 10 variables:
$ carat  : num 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ... 
$ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<...: 5 4 2 4 2 3 3 3 1 3 ... 
$ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<...: 2 2 2 6 7 7 6 5 2 5 ... 
$ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<...: 2 3 5 4 2 6 7 3 4 5 ... 
$ depth  : num 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ... 
$ table  : num 55 61 65 58 58 57 57 55 61 61 ... 
$ price  : int 326 326 327 334 335 336 336 337 337 338 ... 
$ x      : num 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ... 
$ y      : num 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ... 
$ z      : num 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ... 
```
## DATA EXPLORATION

**Diamonds Data Set**

```r
> head(diamonds)
# A tibble: 6 x 10
    carat  cut  color  clarity depth  table price  x   y   z
   <dbl> <ord> <ord> <ord>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
1  0.23  Ideal  E     SI2   61.5   55    326    3.95   3.98  2.43
2  0.21  Premium E     SI1   59.8   61    326    3.89   3.84  2.31
3  0.23  Good   E     VS1   56.9   65    327    4.05   4.07  2.31
4  0.290 Premium I     VS2   62.4   58    334    4.2    4.23  2.63
5  0.31  Good   J     SI2   63.3   58    335    4.34   4.35  2.75
6  0.24  Very Good J     VVS2   62.8   57    336    3.94   3.96  2.48
```
DATA EXPLORATION

Density Plot

Allows to study the distribution of a *NUMERIC* continuous variable.
library(ggplot2)
ggplot(data=diamonds,
      aes(x=price, group=cut, fill=cut)) +
geom_density(adjust = 1.5, alpha = 0.4)
DATA EXPLORATION

```
ggplot(data=diamonds, 
aes(x=price, group=cut, fill=cut)) + 
geom_density(adjust=1.5) + 
facet_wrap(~cut)
```
IN CONTEXT OF RNA-seq

Density Plot

Density of counts distribution

- ctrl3hr
- ctrl18hr
- trt18hr
- trt3hr

log₂ (raw count + 1)

Density
DATA EXPLORATION

Like Density Plot, allows us to study the distribution of a NUMERIC continuous variable.

The variable is cut into several bins, and the number of observation per bin is represented by the height of the bar.
library(ggplot2)
ggplot(data=diamonds, 
aes(x=price, group=cut, fill=cut)) + 
geom_histogram(bins = 20,binwidth = 300) + 
facet_wrap(~cut)
DATA EXPLORATION

- Density
- Histogram
- Barplot
- Violin
- Ridgeline
DATA EXPLORATION

Ridgeline (Joyplot)

Allows to study the distribution of a NUMERIC variable for several groups.
Ridgeline (Joyplot)

```r
geom_density_ridges()
```
Ridgeline (Joyplot)

In Context of scRNA-seq:

**LYZ**
- Platelet
- DC
- NK
- FCGR3A+ Mono
- CD8 T
- B
- CD14+ Mono
- Memory CD4 T
- Naive CD4 T

**CCL5**
- Platelet
- DC
- NK
- FCGR3A+ Mono
- CD8 T
- B
- CD14+ Mono
- Memory CD4 T
- Naive CD4 T

Expression Level
Ridgeline (Joyplot)

Variations:
Boxplots

- Minimum Value in the Data
- 25th Percentile ($Q_1$)
- Median ($Q_2$)
- 75th Percentile ($Q_3$)
- Interquartile Range (IQR)
- Maximum Value in the Data

Potential Outliers:
- Minimum (Minimum Value in the Data, $Q_1 - 1.5 \times IQR$)
- Maximum (Maximum Value in the Data, $Q_3 + 1.5 \times IQR$)
A somewhat misleading boxplot
A boxplot with jitter
# Plot
data %>%
  ggplot( aes(x=name, y=value, fill=name)) +
  geom_boxplot() +
  scale_fill_viridis(discrete = TRUE) +
  geom_jitter(color="grey", size=0.7, alpha=0.5) +
  theme_ipsum() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A boxplot with jitter") +
  xlab("")
# sample size

```r
sample_size = data %>% group_by(name) %>% summarize(num=n())
```

# Plot

```r
data %>%
  left_join(sample_size) %>%
  mutate(myaxis = paste0(name, "\n", "n=", num)) %>%
  ggplot(
    aes(x=myaxis, y=value, fill=name)) +
  geom_violin(width=1.4) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_fill_viridis(discrete = TRUE) +
  theme_ipsum() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A boxplot with jitter") +
  xlab(""")
```
```r
# sample size
sample_size = data %>% group_by(name) %>% summarize(num=n())

# Plot
data %>%
  left_join(sample_size) %>%
  mutate(myaxis = paste0(name, "\n", "n=", num)) %>%
  ggplot( aes(x=myaxis, y=value, fill=name)) +
  geom_violin(width=1.4) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_fill_viridis(discrete = TRUE) +
  theme_ipsum() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A boxplot with jitter") +
  xlab("")
```
IN CONTEXT OF scRNA-seq

**Violin Plot**

- **DLX6-AS1**
  - Expression level (log TPM)
  - Cell Type: 1 to 11

- **GAD2**
  - Expression level (log TPM)
  - Cell Type: 1 to 11

- **NEUROD2**
  - Expression level (log TPM)
  - Cell Type: 1 to 11

- **TFAP2C**
  - Expression level (log TPM)
  - Cell Type: 1 to 11
WHAT IS MY GOAL?

• DATA SUMMARY?
A heatmap is a graphical representation of data, where the individual values contained in a matrix are represented as colors.
### Heatmap

#### Mtcars Data Set

```r
> str(mtcars)
'data.frame': 32 obs. of 11 variables:
$ mpg  : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
$ cyl  : num  6 6 4 6 8 6 8 4 4 6 ...  
$ disp : num 160 160 108 258 360 ...  
$ hp   : num 110 110 93 110 175 105 245 62 95 123 ...  
$ drat : num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
$ wt   : num 2.62 2.88 2.32 3.21 3.44 ...  
$ qsec : num 16.5 17 18.6 19.4 17 ...  
$ vs   : num  0  0  1  1  0  1  0  1  1  1 ...  
$ am   : num  1  1  0  0  0  0  0  0  0  0 ...  
$ gear : num  4  4  4  3  3  3  4  4  4 ...  
$ carb : num  4  4  1  1  2  1  4  2  2  4 ...  
```
## Mtcars Data Set

```r
> head(mtcars)

<table>
<thead>
<tr>
<th></th>
<th>mpg</th>
<th>cyl</th>
<th>disp</th>
<th>hp</th>
<th>drat</th>
<th>wt</th>
<th>qsec</th>
<th>vs</th>
<th>am</th>
<th>gear</th>
<th>carb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mazda RX4</td>
<td>21.0</td>
<td>6</td>
<td>160</td>
<td>110</td>
<td>3.90</td>
<td>2.620</td>
<td>16.46</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mazda RX4 Wag</td>
<td>21.0</td>
<td>6</td>
<td>160</td>
<td>110</td>
<td>3.90</td>
<td>2.875</td>
<td>17.02</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Datsun 710</td>
<td>22.8</td>
<td>4</td>
<td>108</td>
<td>93</td>
<td>3.85</td>
<td>2.320</td>
<td>18.61</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Hornet 4 Drive</td>
<td>21.4</td>
<td>6</td>
<td>258</td>
<td>110</td>
<td>3.08</td>
<td>3.215</td>
<td>19.44</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Hornet Sportabout</td>
<td>18.7</td>
<td>8</td>
<td>360</td>
<td>175</td>
<td>3.15</td>
<td>3.440</td>
<td>17.02</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Valiant</td>
<td>18.1</td>
<td>6</td>
<td>225</td>
<td>105</td>
<td>2.76</td>
<td>3.460</td>
<td>20.22</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
```
library(pheatmap)
pheatmap(mat = mtcars, scale = "column")
In Context of RNA-seq:
DATA EXPLORATION

Dendrogram

A network structure consisting of nodes and edges
library(DESeq2)
hc2 <- hclust(dist(t(assay(vsd))),
               method="ward.D")
plot(hc2, hang=-1, ylab="Height",
     xlab="Method: Euclidean distance - Ward criterion",
     main="Cluster Dendrogram")
DATA EXPLORATION

Scatter Plot

Displays a relationship between two NUMERICAL variables
Iris Data Set

```r
> str(iris)
'data.frame': 150 obs. of 5 variables:
$ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
$ Sepal.Width : num 3.5 3.3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
$ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
$ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
$ Species   : Factor w/ 3 levels "setosa","versicolor","virginica": 1 1 1 1 1 1 1 1 1 1 ...  
```
Iris Data Set

> head(iris)

<table>
<thead>
<tr>
<th>Sepal.Length</th>
<th>Sepal.Width</th>
<th>Petal.Length</th>
<th>Petal.Width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>4.6</td>
<td>3.1</td>
<td>1.5</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>5.0</td>
<td>3.6</td>
<td>1.4</td>
<td>0.2</td>
<td>setosa</td>
</tr>
<tr>
<td>5.4</td>
<td>3.9</td>
<td>1.7</td>
<td>0.4</td>
<td>setosa</td>
</tr>
</tbody>
</table>
library(ggplot2)
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, color=Species)) + geom_point(size=2)
IN CONTEXT OF RNA-seq

Principal Component Analysis - Axes 1 and 2

PC2 (40.83%)

PC1 (58.24%)

g3r1

g3r2

-60 -40 -20 0 20 40

-40 0 20 40

-20
IN CONTEXT OF scRNA-seq
A *bubble plot* is a *scatterplot* where a third dimension is added: the value of an additional numeric variable is represented through the size of the dots.
IN CONTEXT OF scRNA-seq
DATA EXPLORATION

A **Venn diagram** shows all possible logical relationships between a finite collection of different sets.