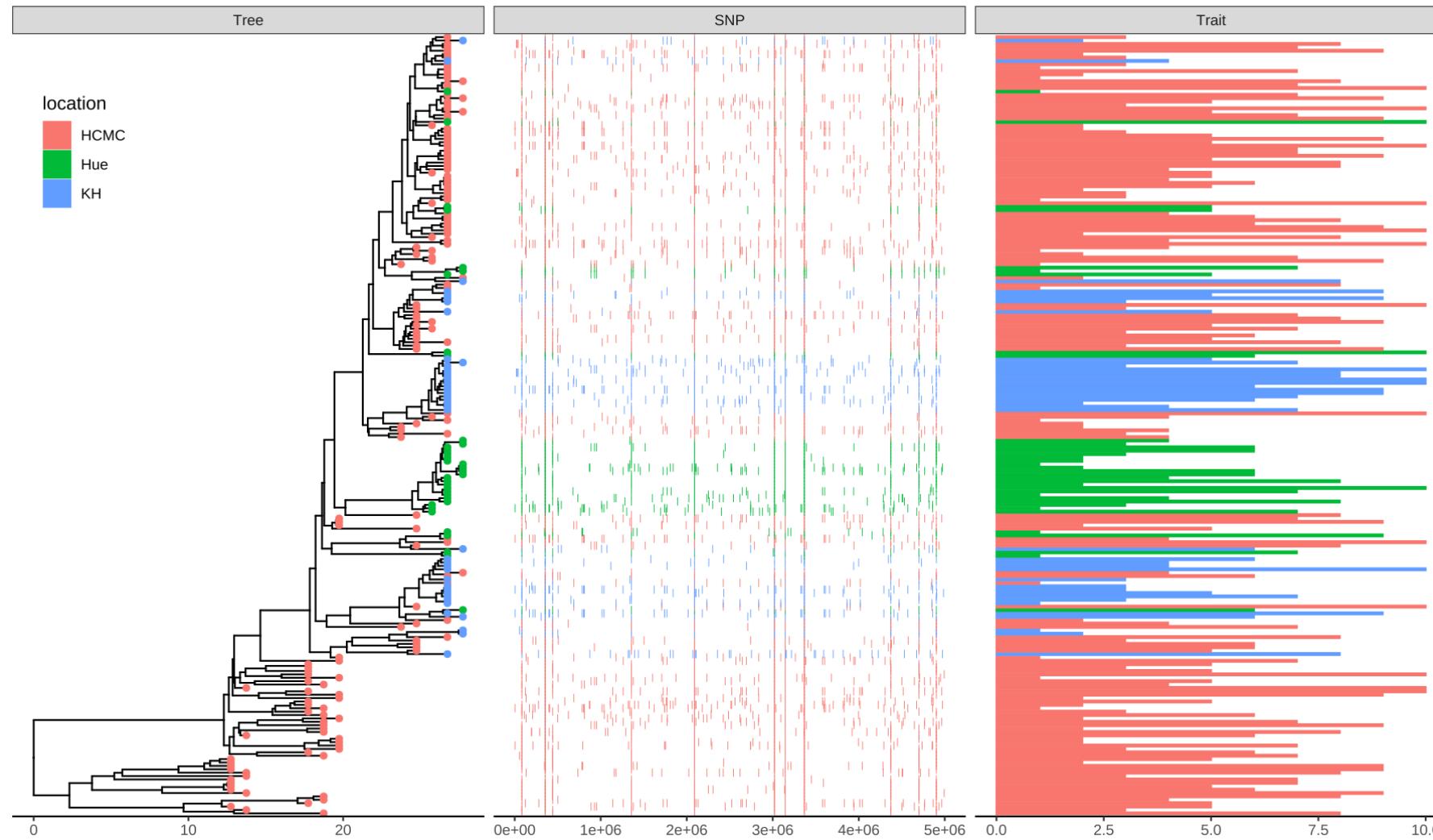


CM 515: Data Visualization Module I



Dan Sloan

Department of Biology

dan.sloan@colostate.edu

Image Source: Guangchuang Yu <https://yulab-smu.top/treedata-book/chapter7.html>

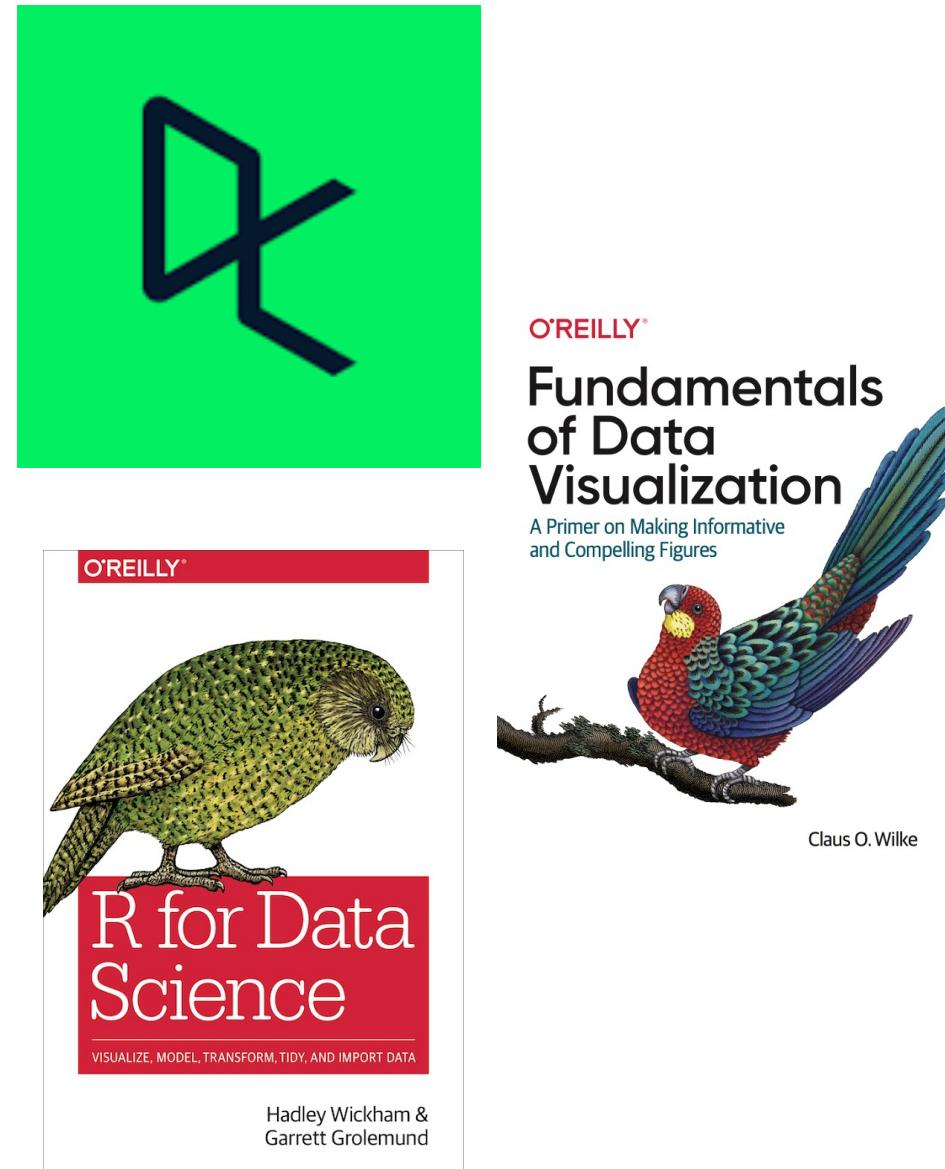
Data Visualization Resources

Datacamp

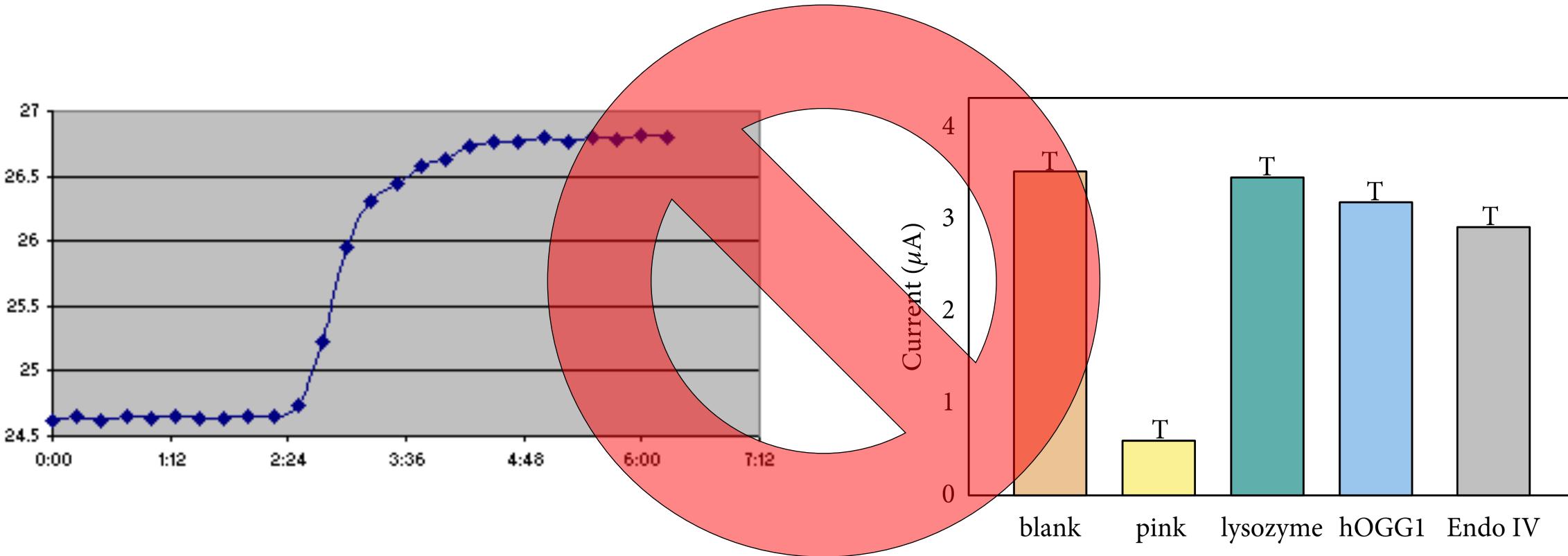
- [Introduction to Data Visualization with ggplot2](#)
- [Intermediate Data Visualization with ggplot2](#)
- [Visualization Best Practices in R](#)

Claus Wilke Data Visualization in R Course (U Texas)

R for Data Science



Scientific Communication and Professionalism

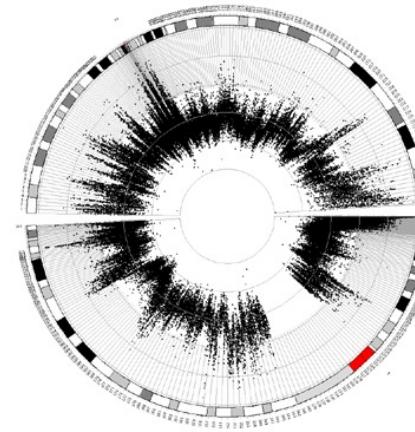


- Clear, accurate, and complete representation of your data
- Efficient, reproducible, and automated methods
- Clean, professional, and aesthetically pleasing appearance

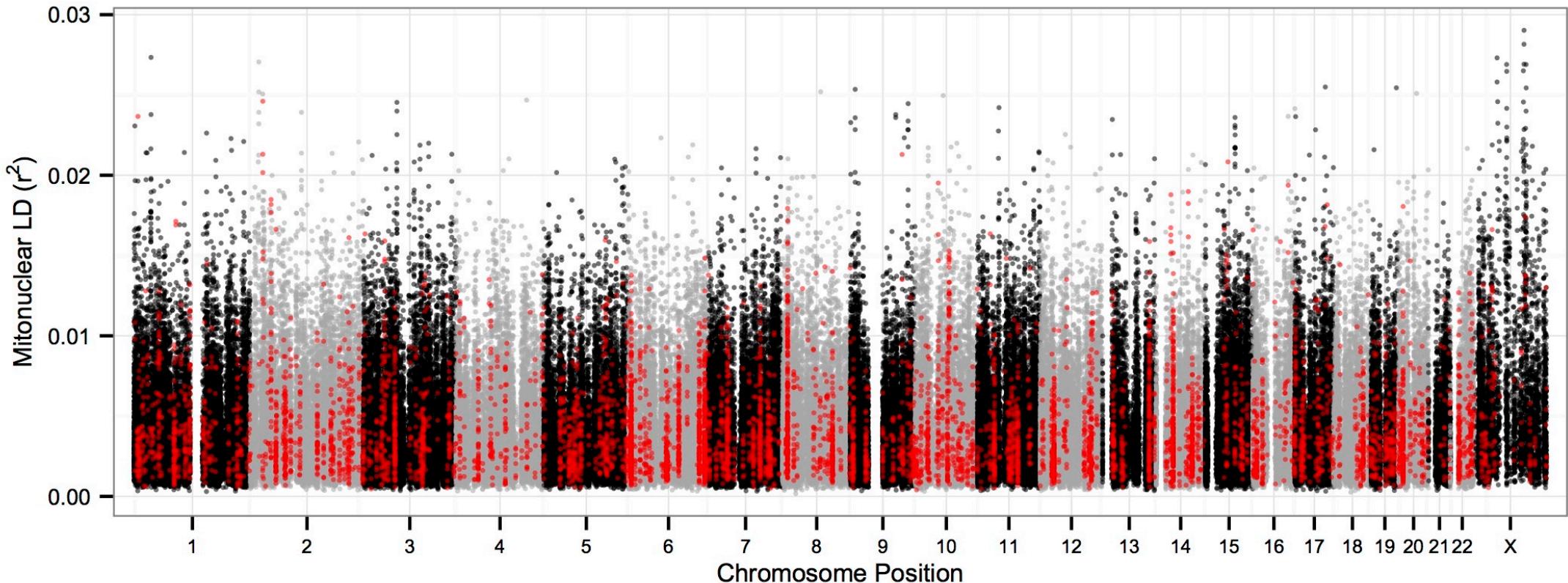
The Right Tools for the Job

A few examples

- [R and ggplot](#)
- [Circos](#)
- [Processing](#)
- [Adobe Illustrator](#)
- [BioRender](#)



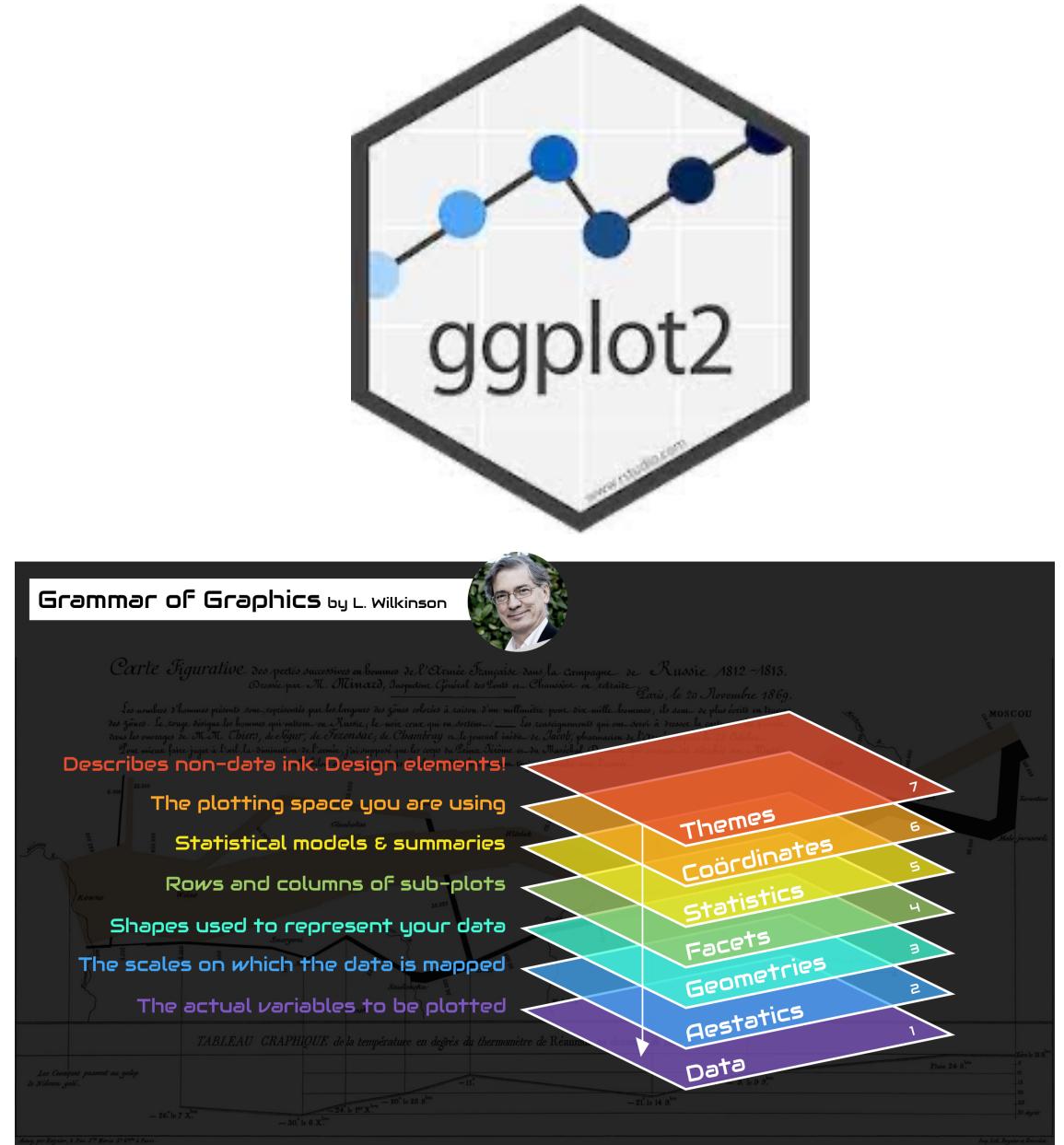
Making Figures with Code



```
ggplot(cnld) + geom_point(aes(x=CumPos, y=r2, size=0.75, colour=as.factor(ChromPrint), alpha = 1/8)) + scale_size_identity() + theme_bw(base_size=15) +  
scale_color_manual(values=c(rep(c('black', 'dark gray'),11), 'black', 'red')) +  
scale_x_continuous(expand = c(0.015, 0.015),labels=c(as.character(1:chrNum), "X"),  
breaks=bpMidVec) + theme(plot.margin = unit( c(0.03,0.03,0.03,0.03) , "in" ),  
legend.position='none', axis.text.x = element_text(size=6), axis.text.y = element_text(size=7),  
axis.title.x = element_text(size=8), axis.title.y = element_text(size=8)) + xlab('Chromosome  
Position') + ylab(expression(paste("Mitonuclear LD (", r^2, ")")))
```

The Grammar of Graphics

- **aes:** Aesthetic mapping of data to plot elements
 - position (X or Y coordinates), shapes, sizes, color, line weight/type, transparency, etc.
- **geoms:** Layers visually representing your mapped data
 - points, lines, bars, density curves, etc.
- **facets:** Dividing into subplots
 - partitions dataset based on one or more variables
- **themes:** Non-data plot elements
 - axis labels, grid lines, titles, etc.



A Few Principles of Data Visualization

- What are you trying to communicate??
- The demise of the bar plot
- Choosing a scale: log vs. linear
- Use and choice of color palettes

Patterns

CelPress
OPEN ACCESS

Perspective

Principles of Effective Data Visualization

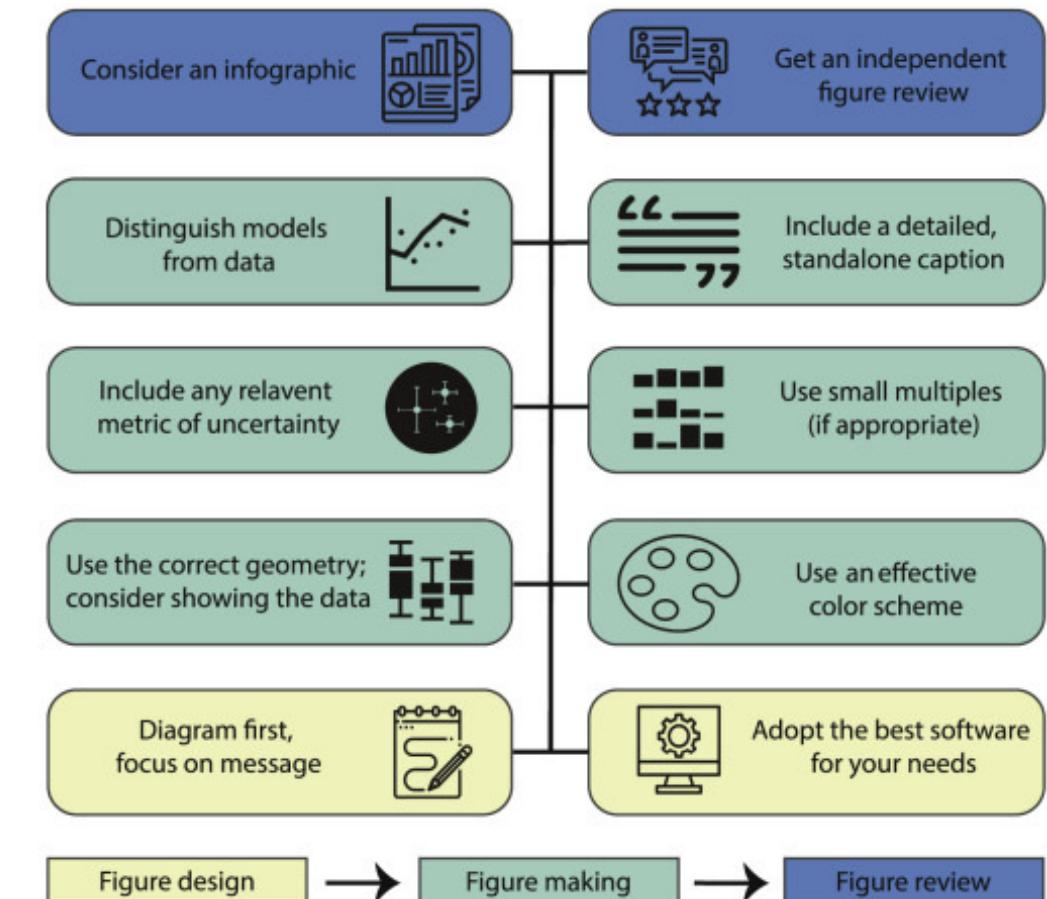
Stephen R. Midway^{1,*}

¹Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA 70803, USA

*Correspondence: smidway@lsu.edu

<https://doi.org/10.1016/j.patter.2020.100141>

<https://www.sciencedirect.com/science/article/pii/S2666389920301896>



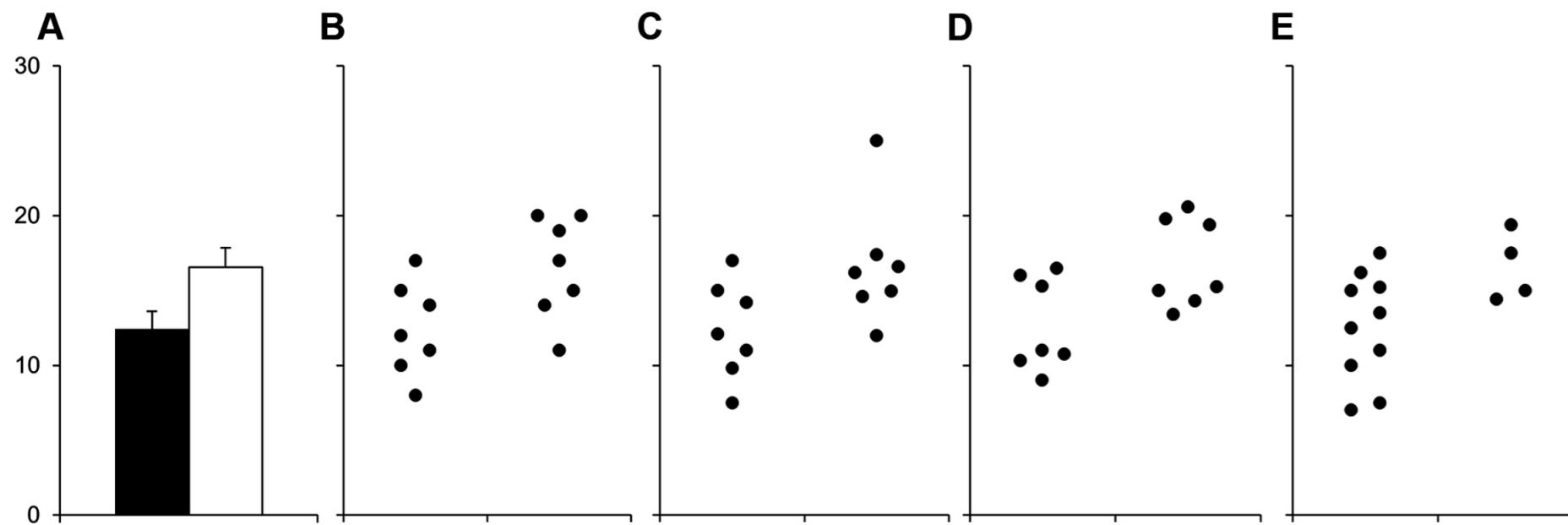
The Demise of the Bar Plot

PLOS BIOLOGY

Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm

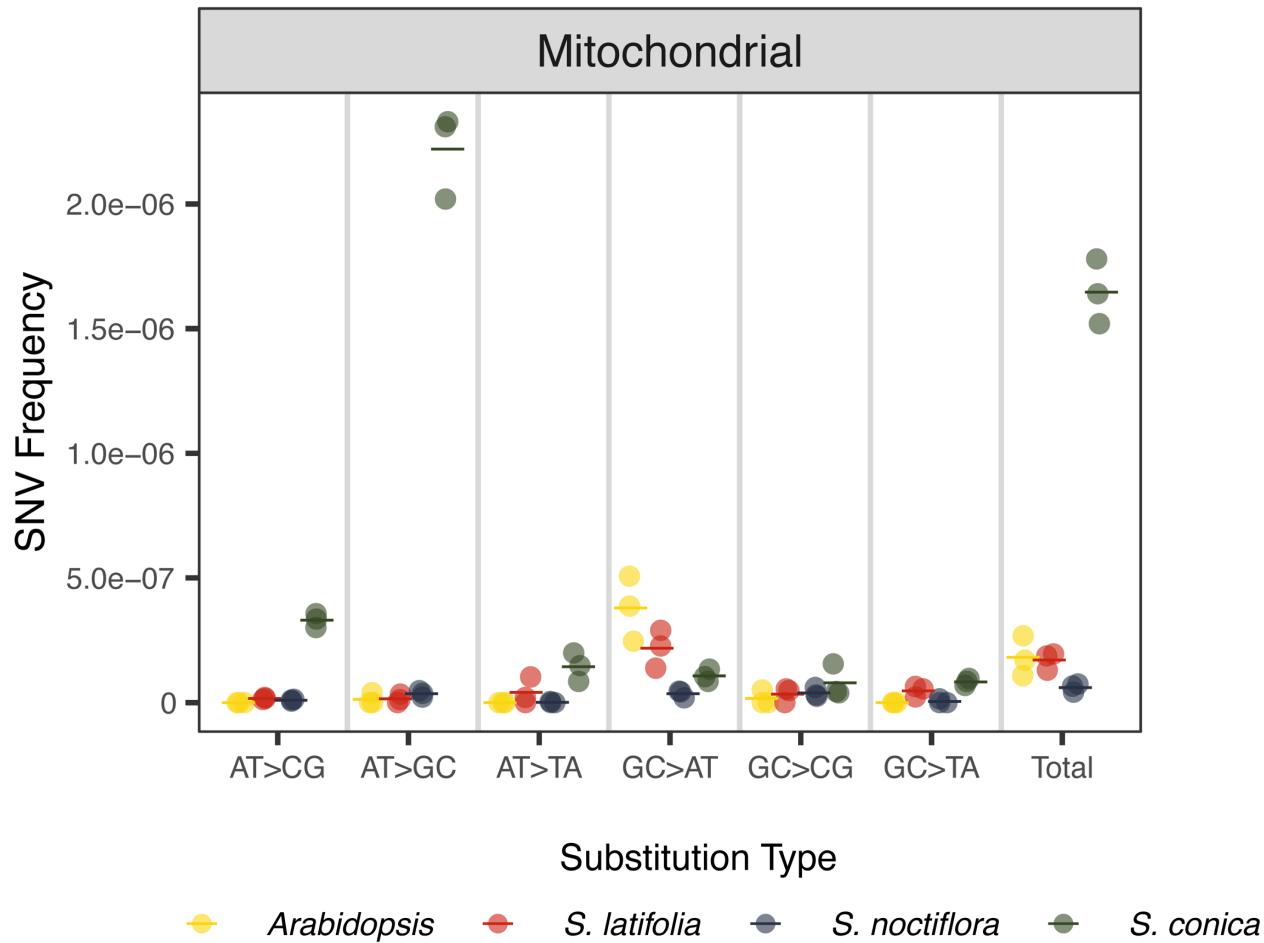
Tracey L. Weissgerber , Natasa M. Milic, Stacey J. Winham, Vesna D. Garovic

Published: April 22, 2015 • <https://doi.org/10.1371/journal.pbio.1002128>



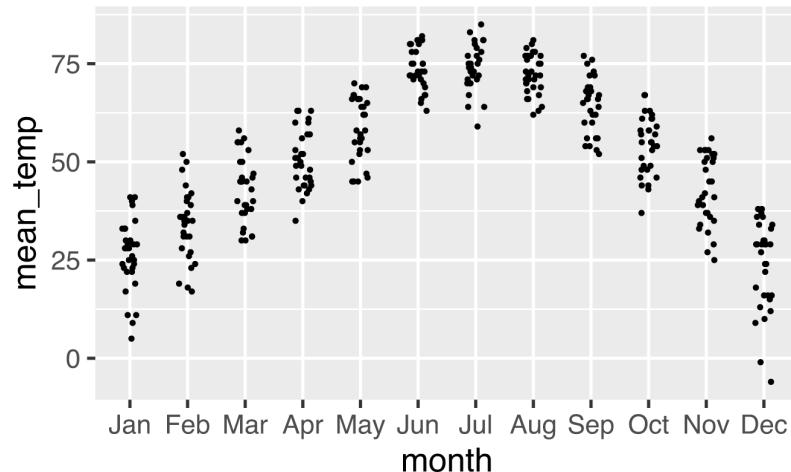
When Possible... Show All the Data!

Use point size, jitter, and/or transparency to mitigate the effects of overlapping points.

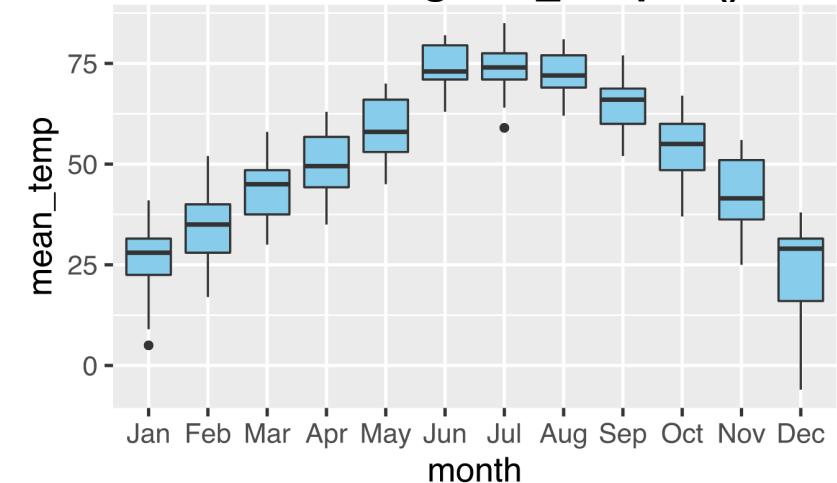


Better Ways of Comparing and Summarizing Distributions

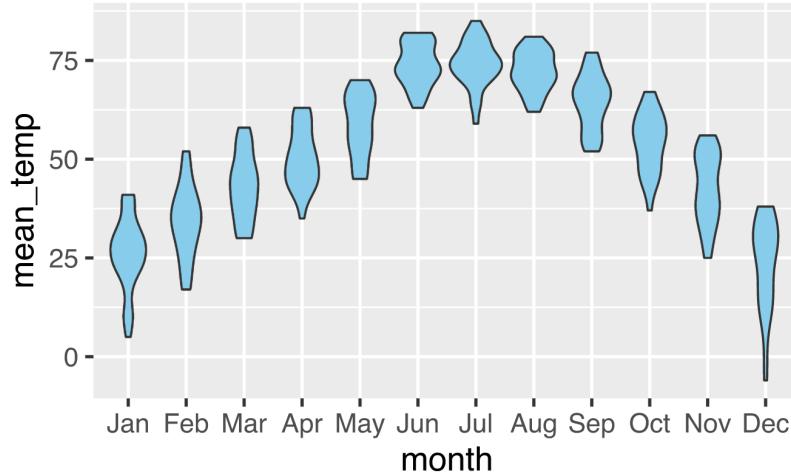
All Data Points: `geom_point()` [with jitter]



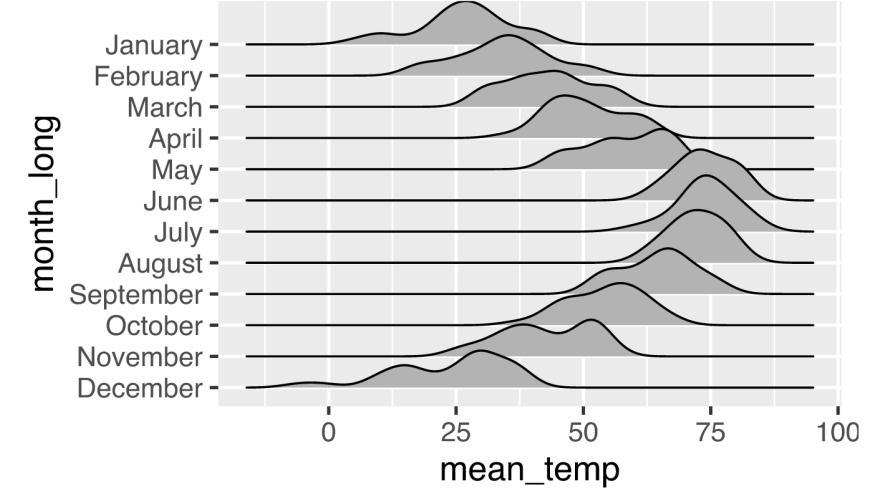
Box Plots: `geom_boxplot()`



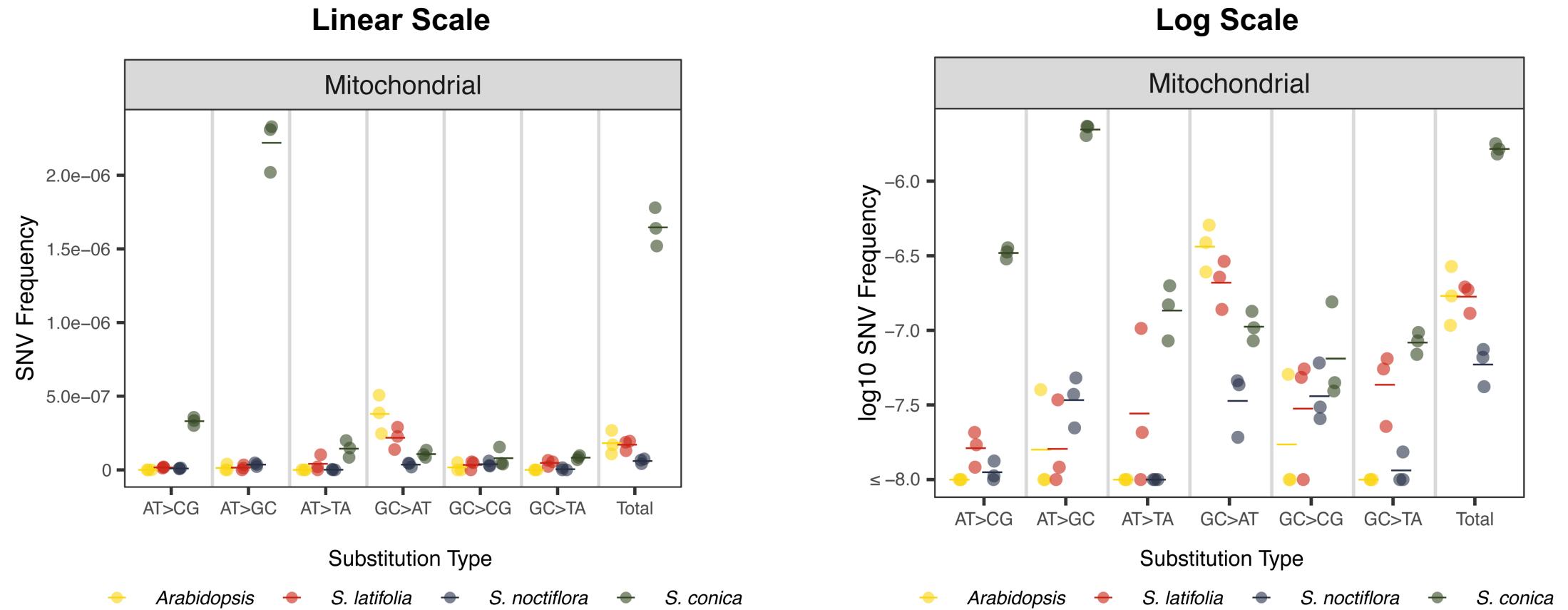
Violin Plots: `geom_violin()`



Ridge Plots: `geom_density_ridges()`



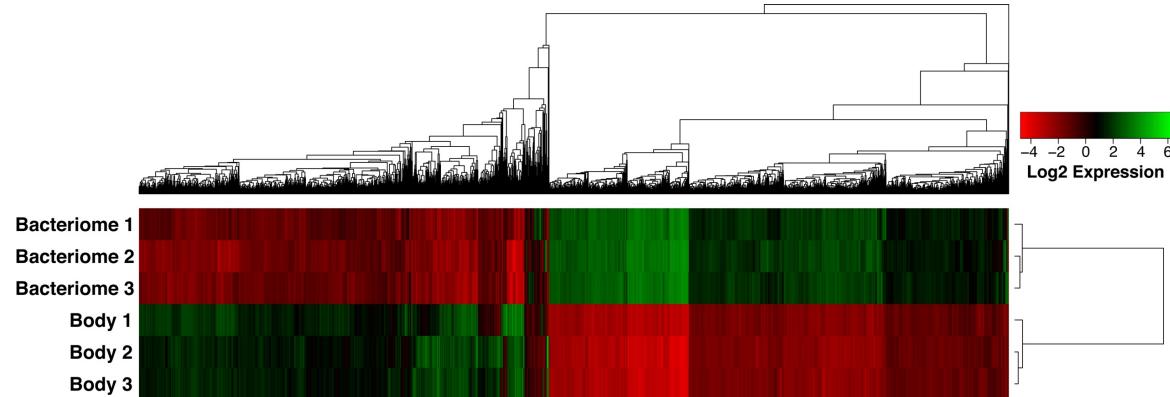
Linear vs. Log Scales



- Use **linear** scales to emphasize **absolute** differences.
- Use **log** scales to emphasize **proportional** differences.

Accessibility

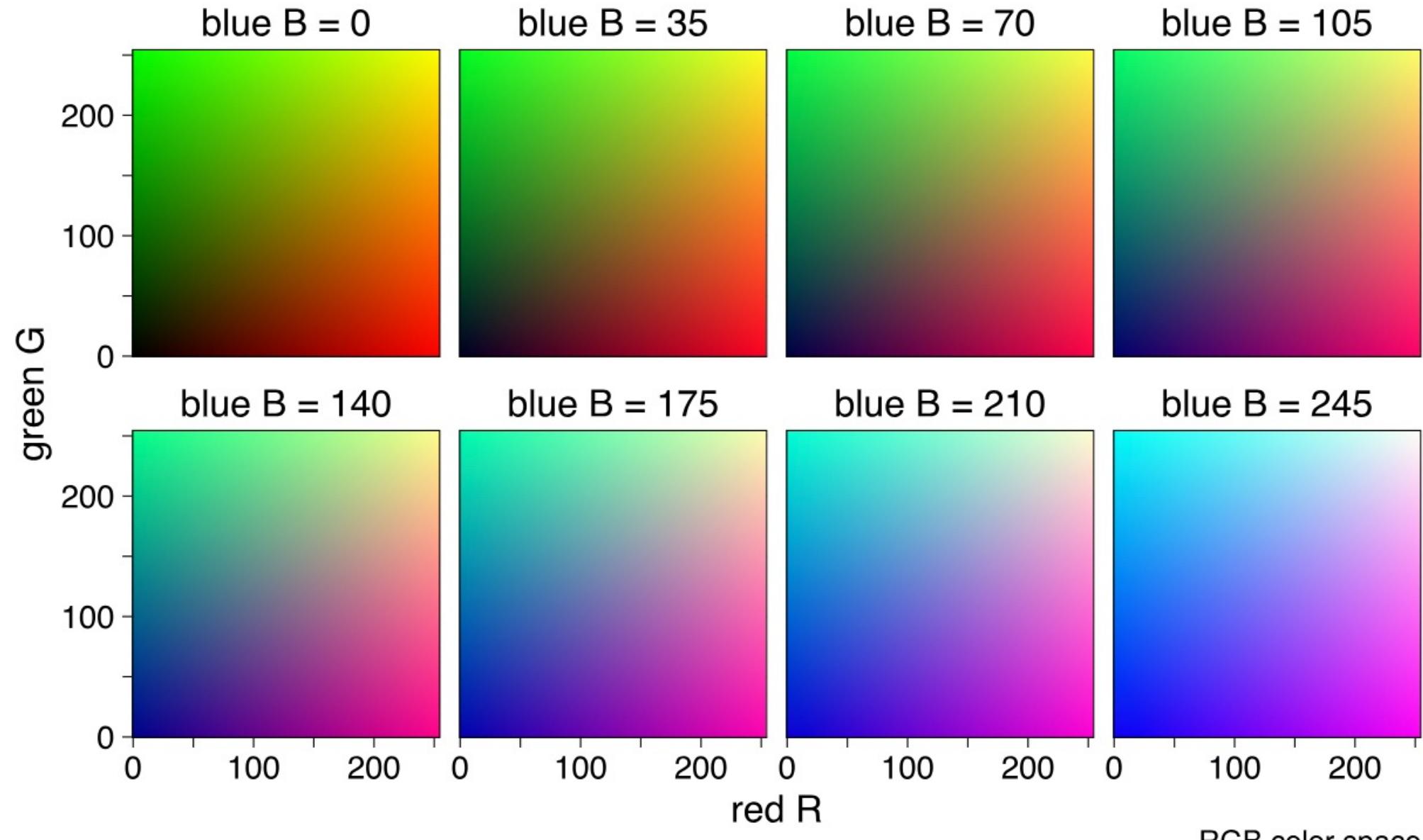
Color is a powerful tool for visualizations, but it will not be perceived in the same way by everyone in your audience. Tips for making your visualizations accessible to color blind individuals....



- Use [palettes consisting of colors that are more distinguishable](#) for individuals with common forms of color blindness.
- Use color and shape of points redundantly to distinguish among groups in plot.

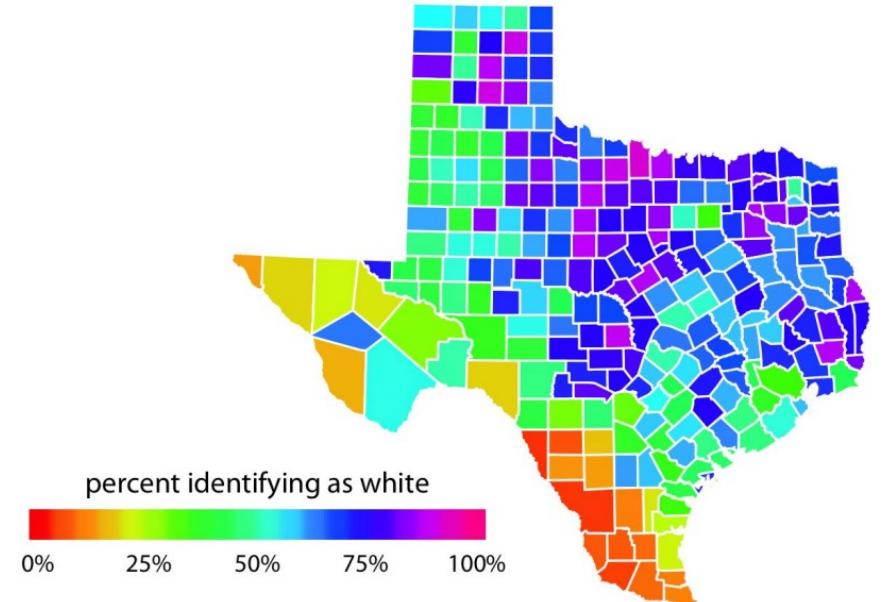
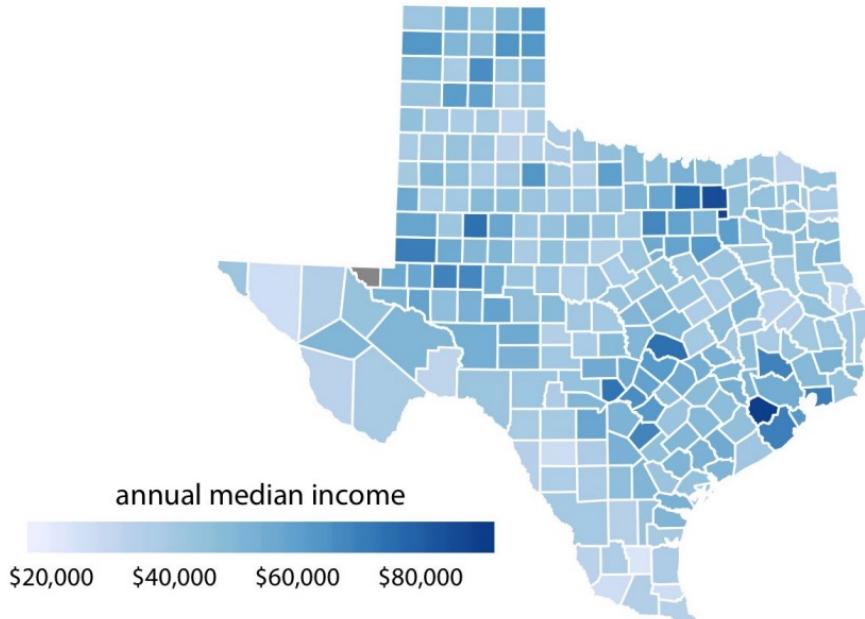
```
> ggplot(MouseData, aes(x=age, y=weight, color=genotype, shape=genotype))  
+ geom_point()
```
- Make use of figure labeling and legend descriptions to make the plot accessible even if colors are difficult to distinguish.

Choosing Colors

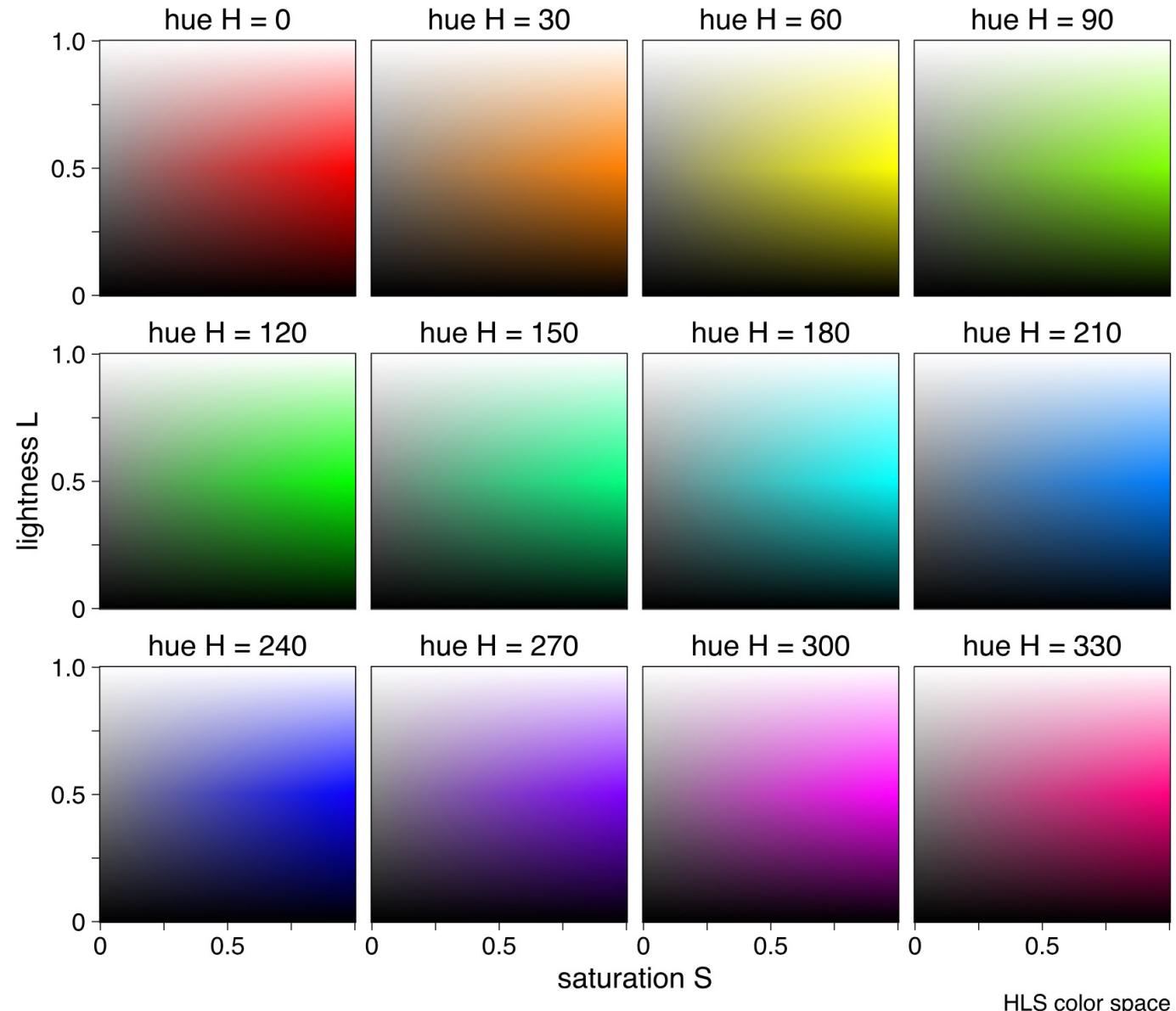


Go Easy on Our Eyes

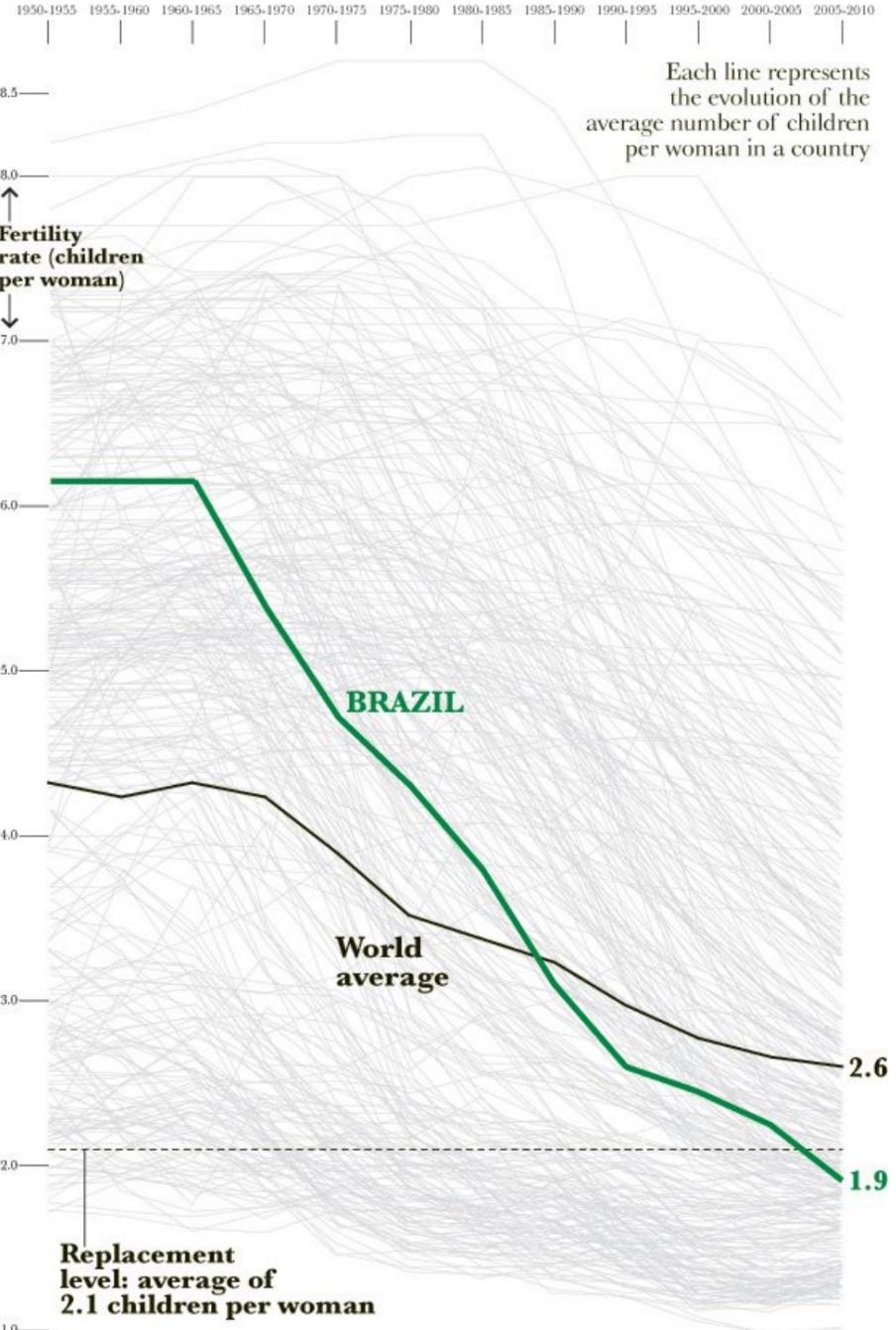
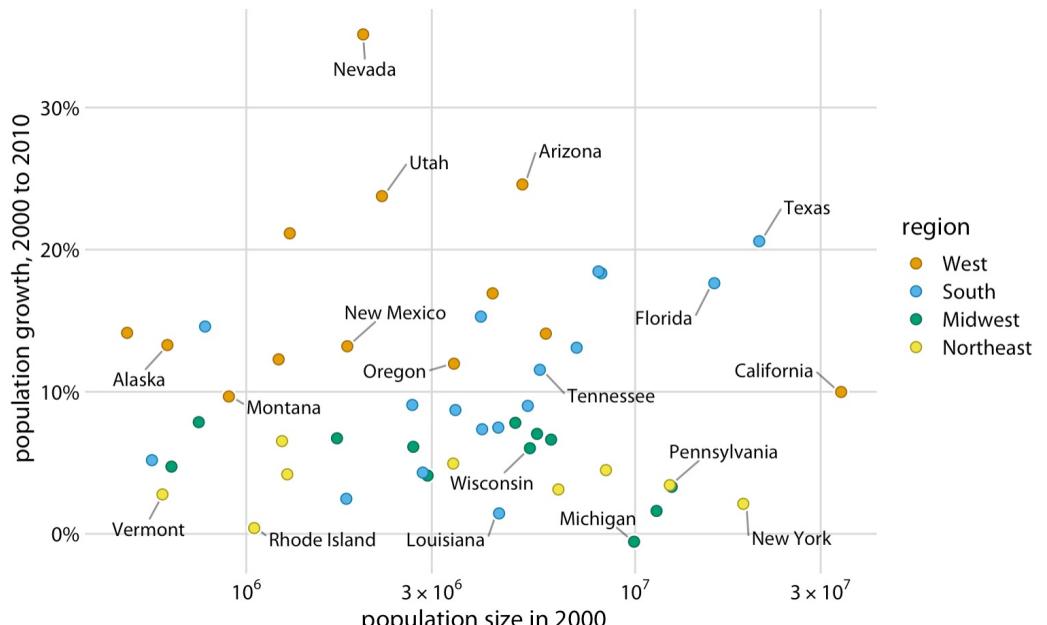
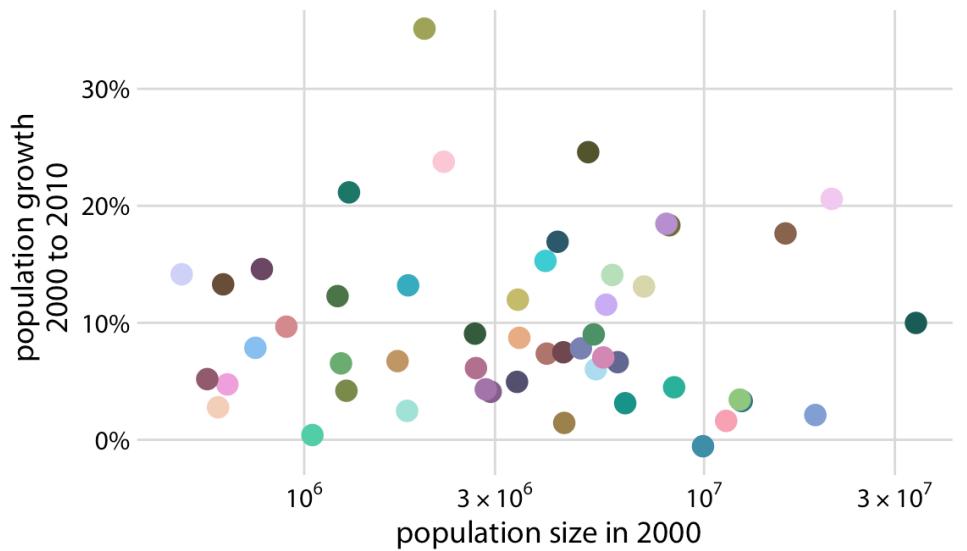
Generic named colors in R tend to have extremely high saturation.



Thinking of Colors in Terms of Hue, Lightness and Saturation



Color to Emphasize, Not to Overwhelm



Jess Cohen-Tanugi

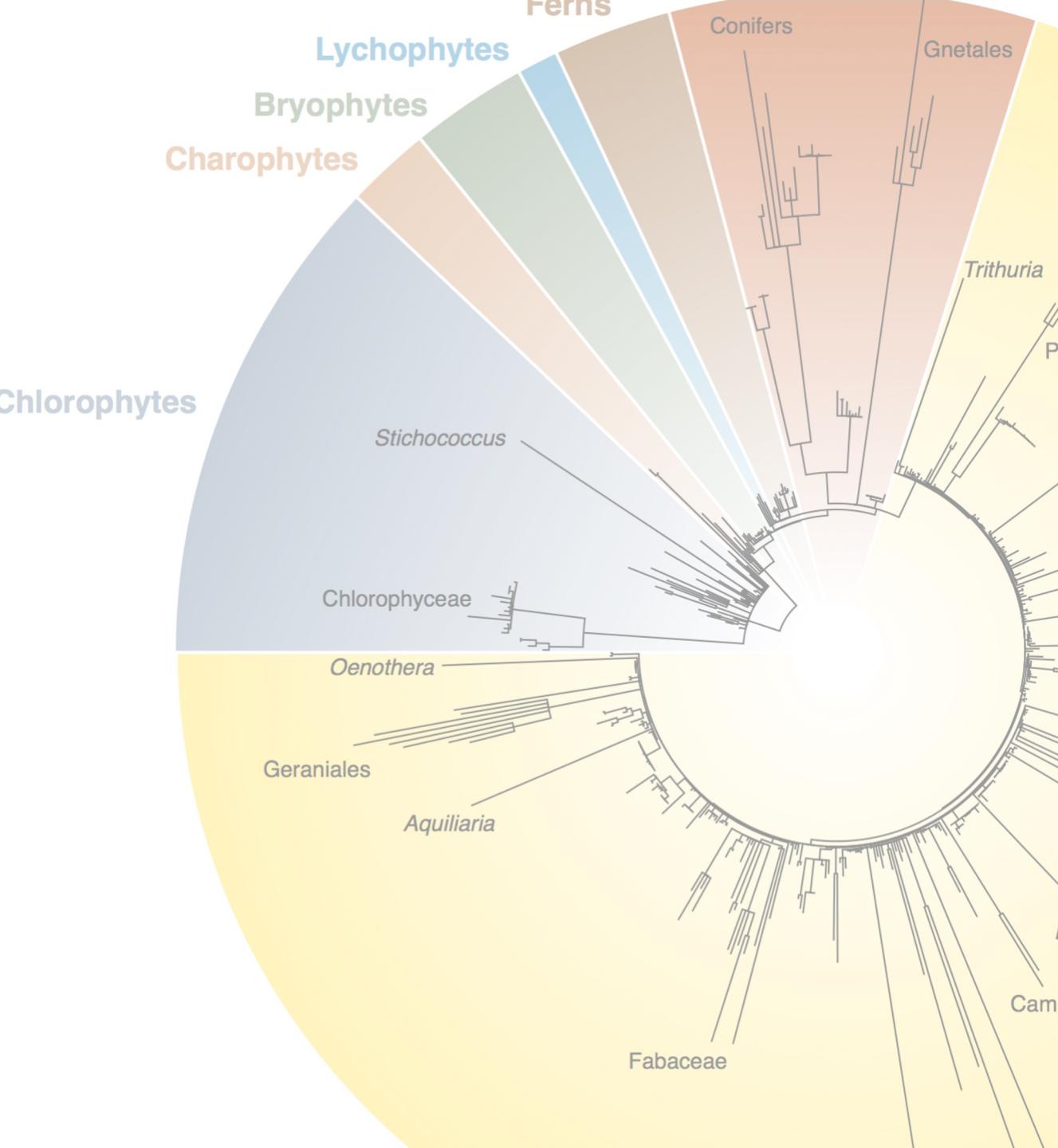
<https://clauswilke.com/dataviz/color-pitfalls.html>

Exercise and Assignment

<https://dbsloan.github.io/CM515/SP24/ggplot/>

Tree-Like Data Structures in Biology

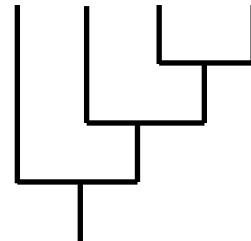
- The tree of life and species relationships
- Gene family evolution
- Hierarchical clustering of gene expression patterns, ecological/microbiome communities, etc.



Newick Tree Format

- Standard parentheses/comma-based format for summarizing tree branching patterns

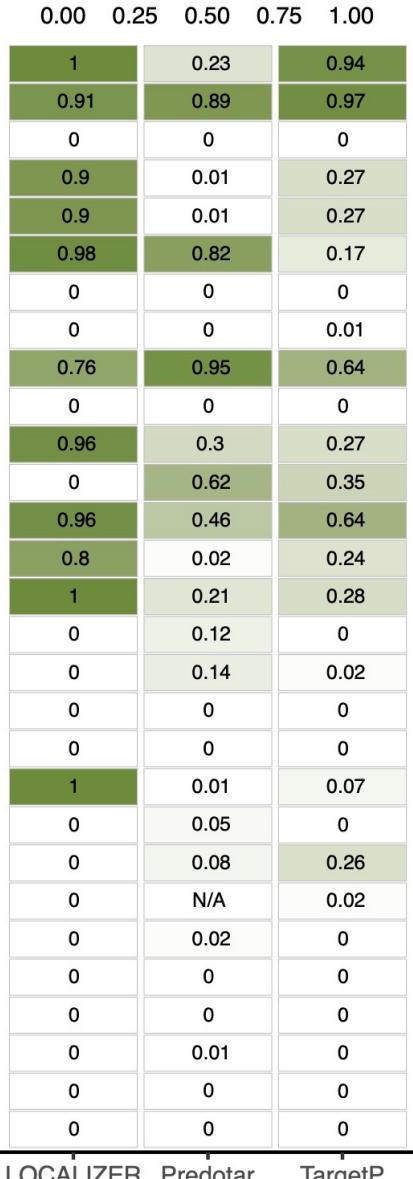
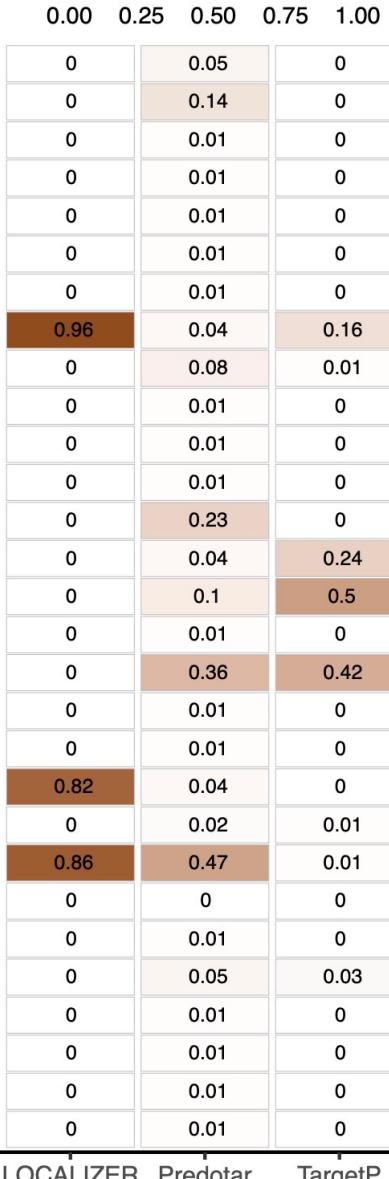
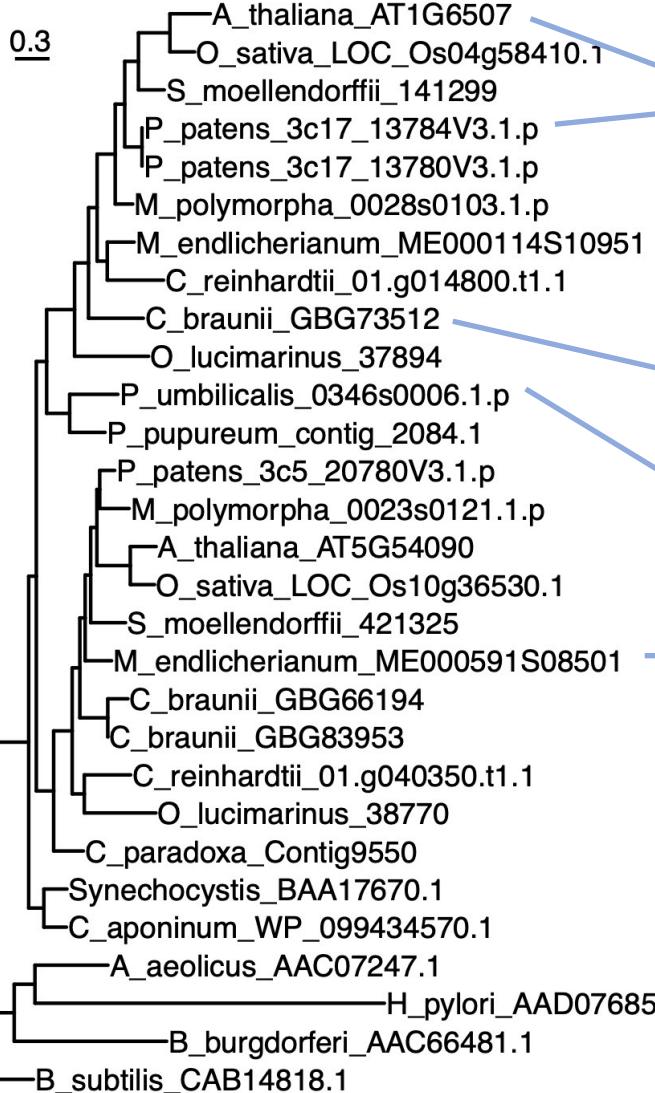
(A, (B, (C, D)));



- Numerical values and annotations can be added to record features such as branch lengths, statistical (e.g., bootstrap) support, node/branch labels, and other features.



Linking Trees and Data Plots



Exercise and Assignment

<https://dbsloan.github.io/CM515/SP24/ggtree/>