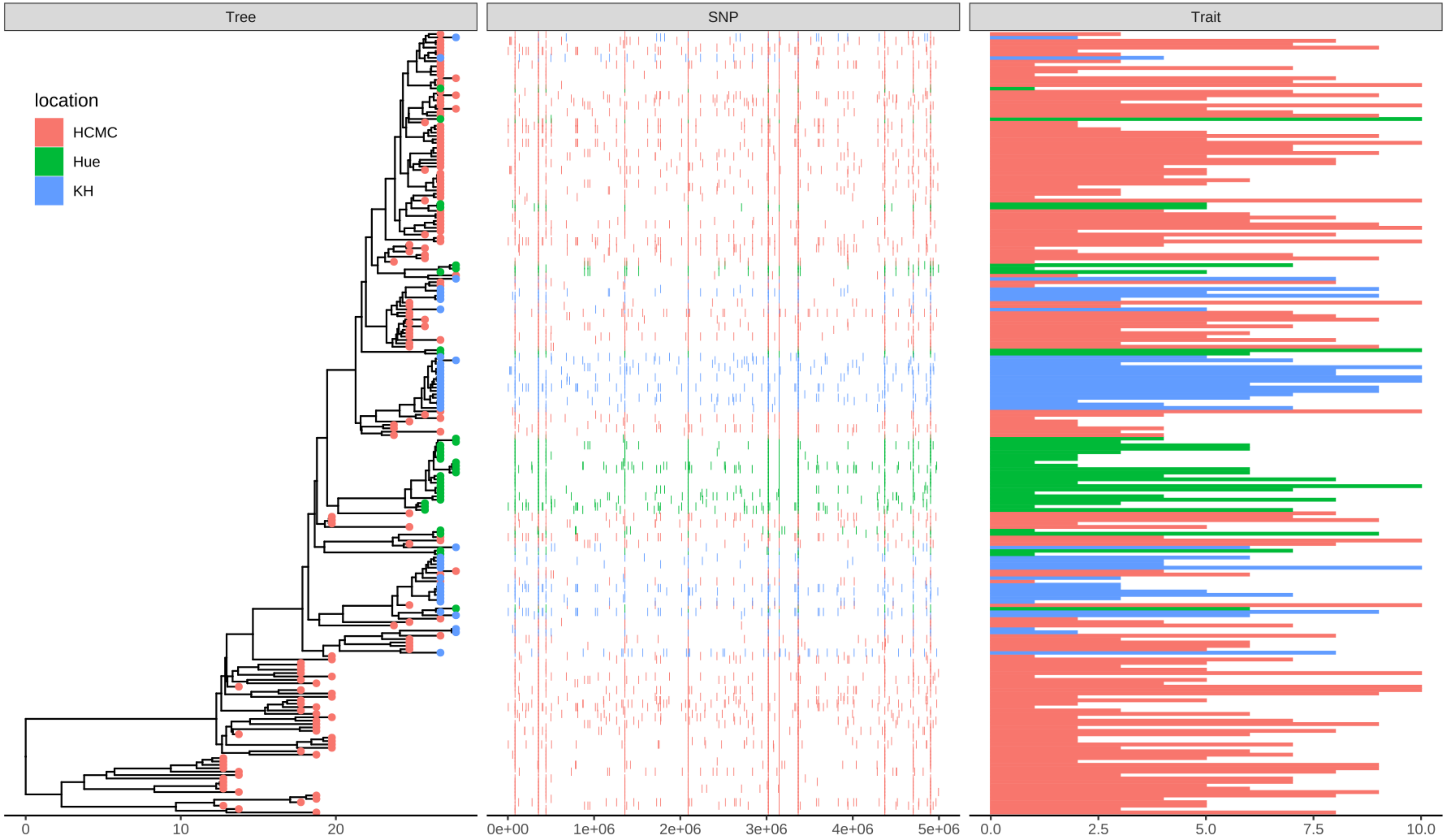
CM 515: Data Visualization Module I



Dan Sloan
Department of Biology
dan.sloan@colostate.edu

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

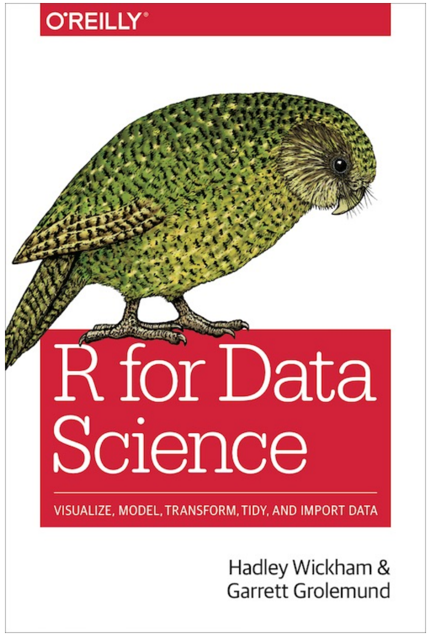
Data Visualization Resources

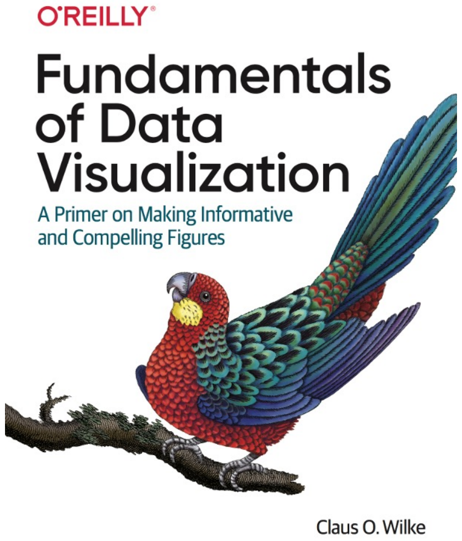
Datacamp

- Introduction to Data Visualization with ggplot2
- Intermediate Data Visualization with ggplot2
- Visualization Best Practices in R

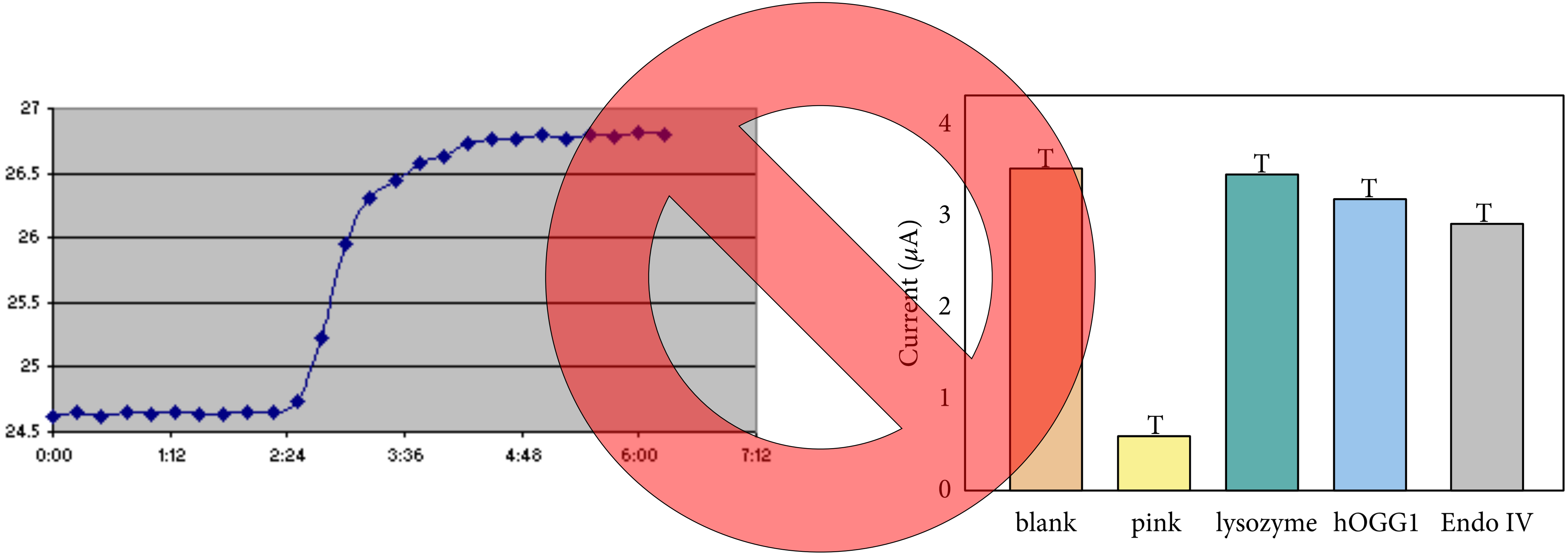
Claus Wilke Data Visualization in R Course (U Texas)

R for Data Science




O'REILLY
Fundamentals of Data Visualization
A Primer on Making Informative and Compelling Figures
Claus O. Wilke

Scientific Communication and Professionalism

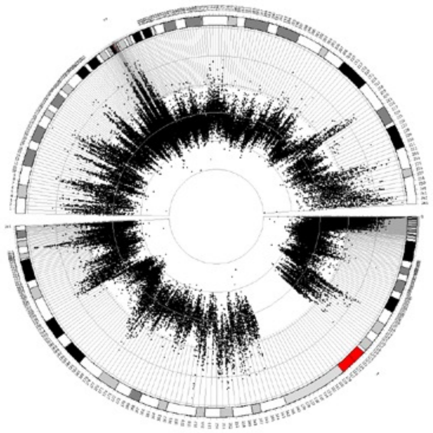


- Clear, accurate, and complete representation of your data
- Efficient, reproducible, and automated methods
- Clean, professional, and aesthetically pleasing appearance

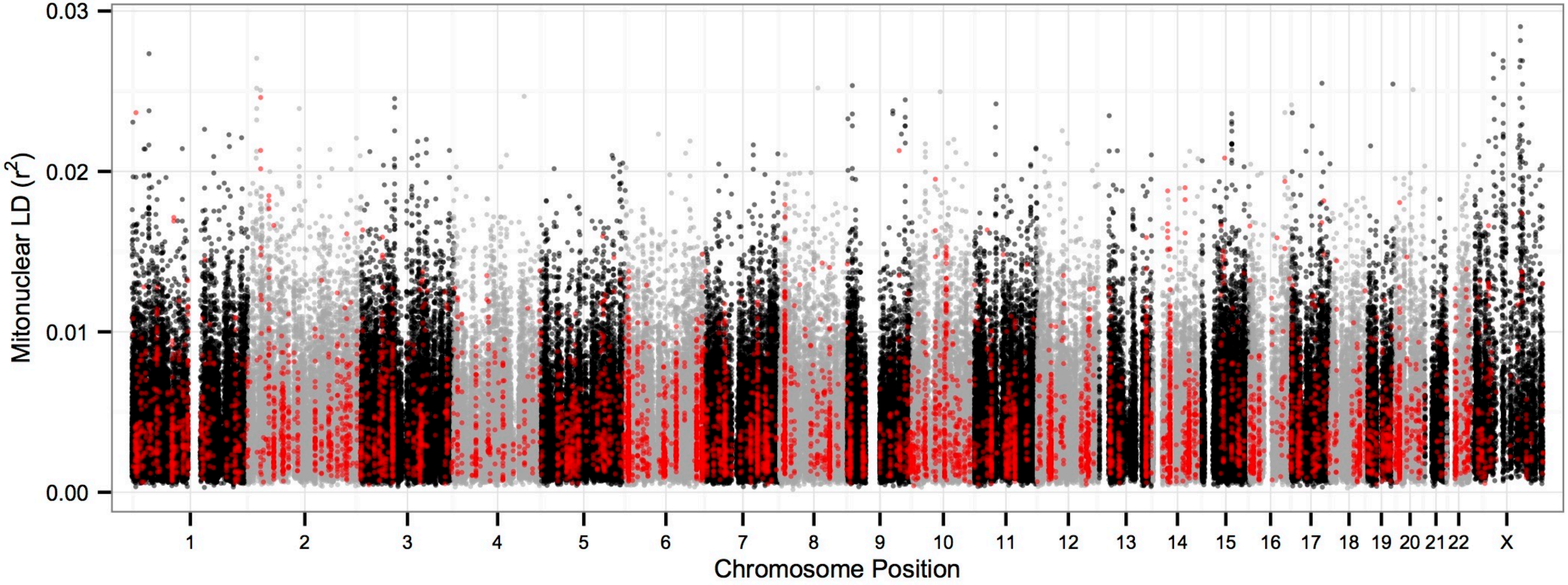
The Right Tools for the Job

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.

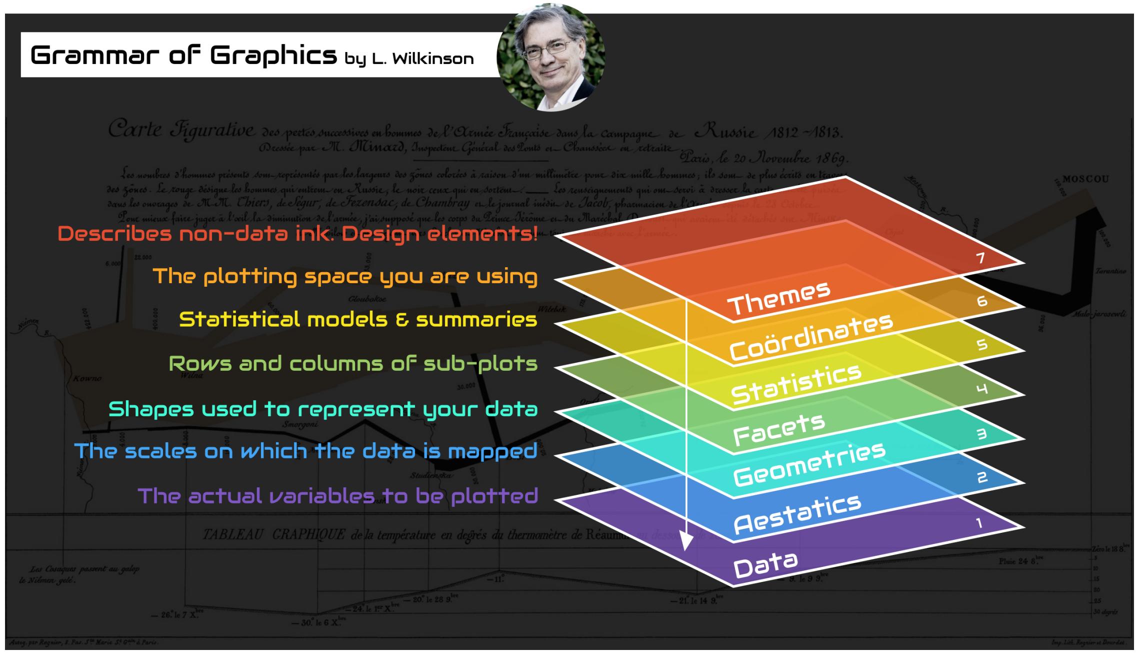
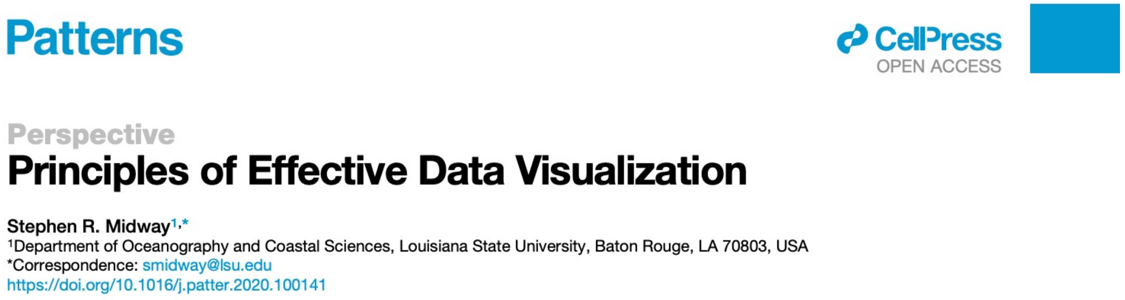

Grammar of Graphics by L. Wilkinson
Describes non-data ink. Design elements!
The plotting space you are using
Statistical models & summaries
Rows and columns of sub-plots
Shapes used to represent your data
The scales on which the data is mapped
The actual variables to be plotted
Themes 7
Coördinates 6
Statistics 5
Facets 4
Geometries 3
Aestatics 2
Data 1

Image: Thomas de Beus

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

CellPress OPEN ACCESS

Perspective

Principles of Effective Data Visualization

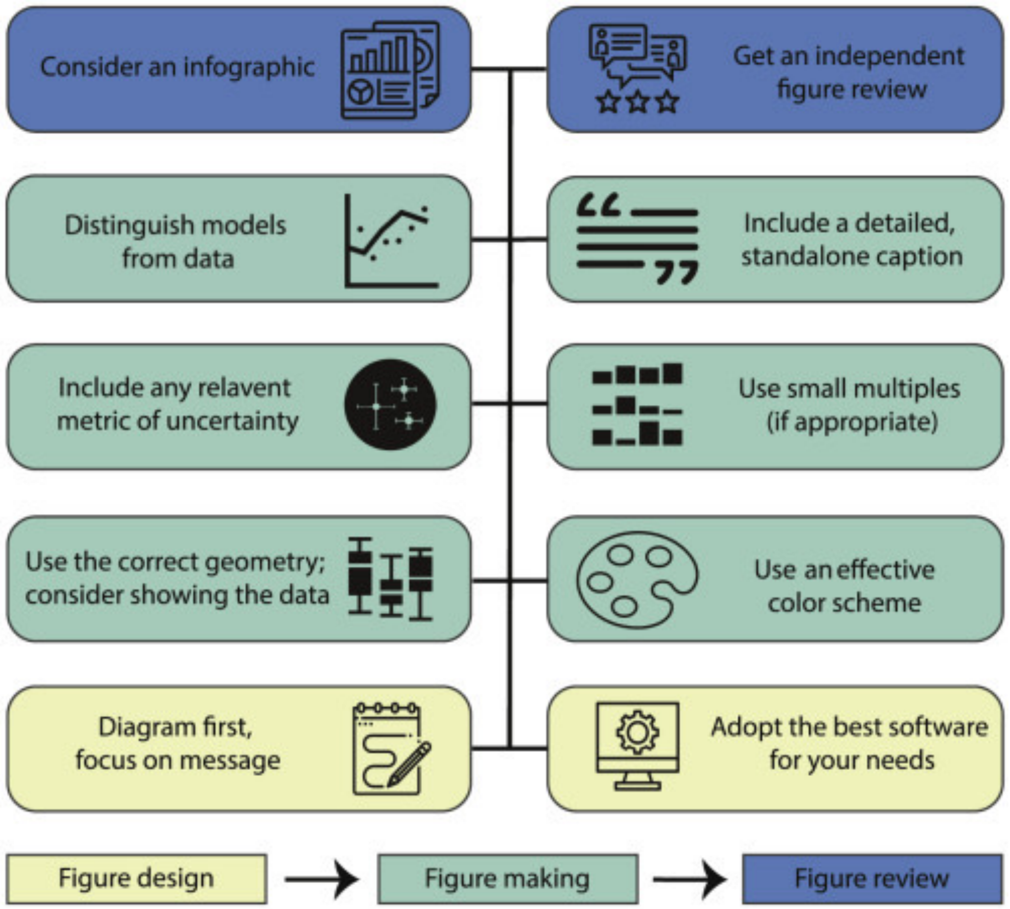
Stephen R. Midway[1,*]

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

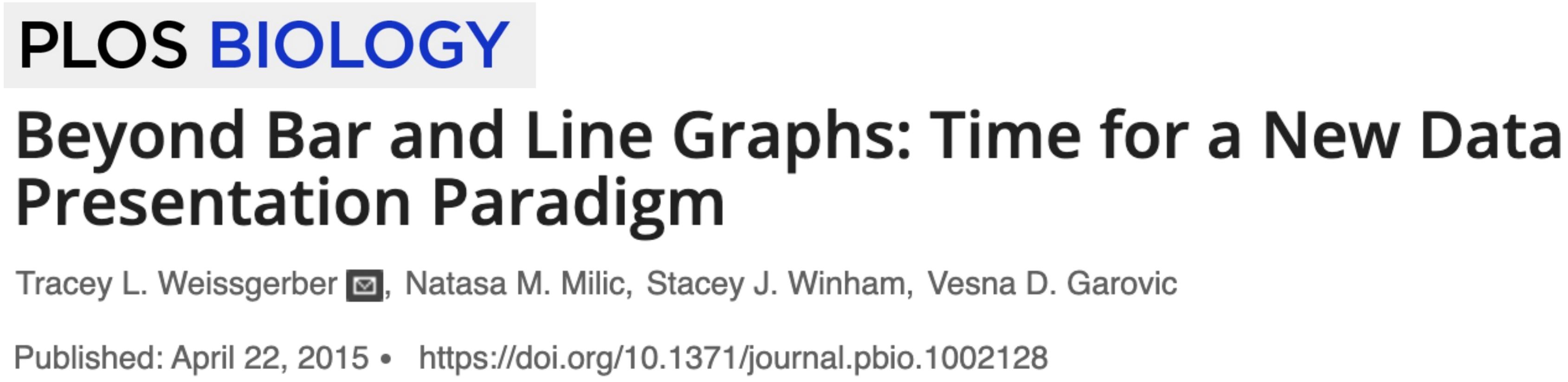
*Correspondence: smidway@lsu.edu

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

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


Consider an infographic
Get an independent figure review
Distinguish models from data
Include a detailed, standalone caption
Include any relavent metric of uncertainty
Use small multiples (if appropriate)
Use the correct geometry; consider showing the data
Use an effective color scheme
Diagram first, focus on message
Adopt the best software for your needs
Figure design
Figure making
Figure review

The Demise of the Bar Plot

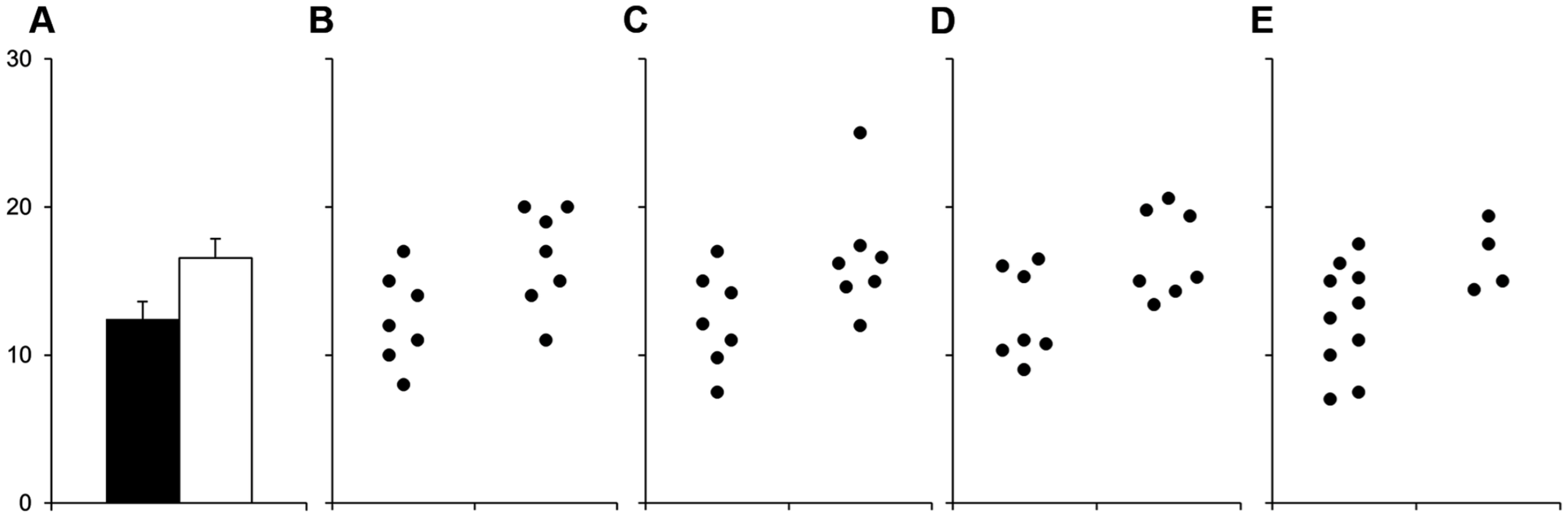


PLOS BIOLOGY

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

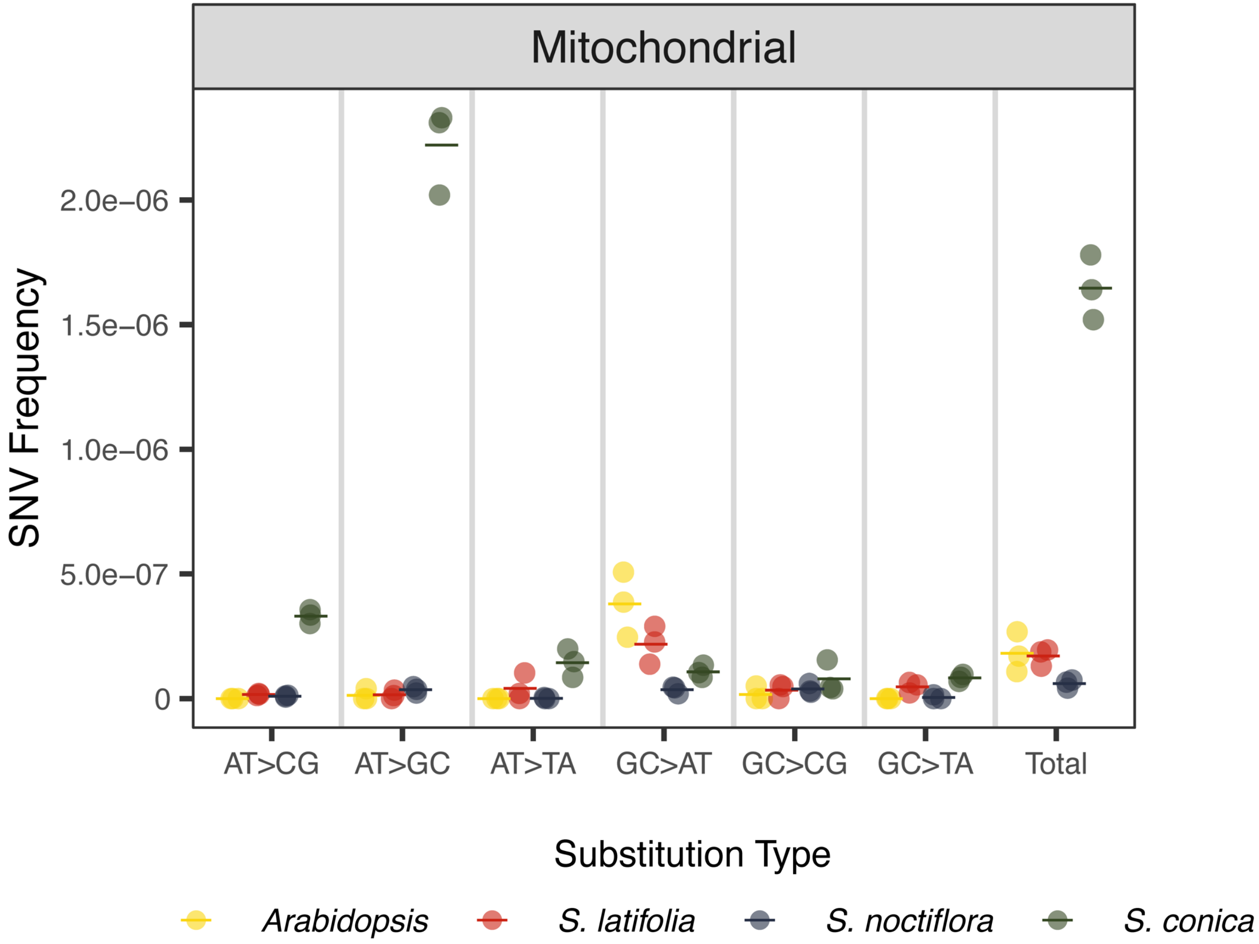
Tracey L. Weissgerber, Natasa M. Milic, Stacey J. Winham, Vesna D. Garovic

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



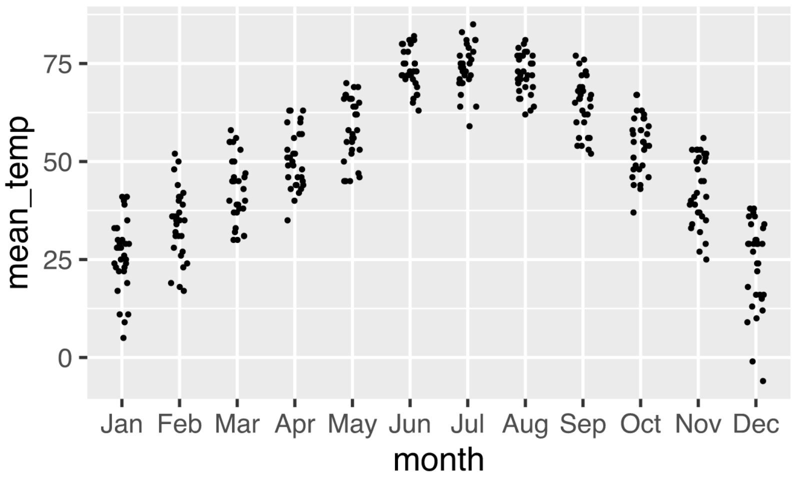
When Possible… Show All the Data!

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

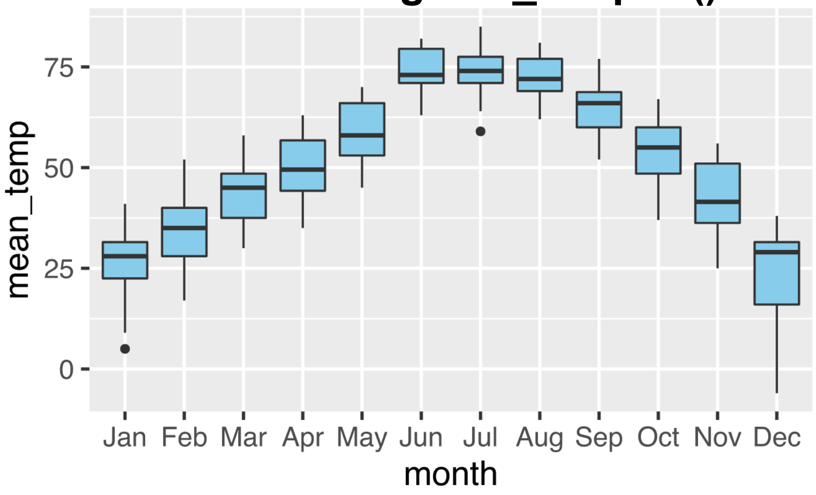


Better Ways of Comparing and Summarizing Distributions

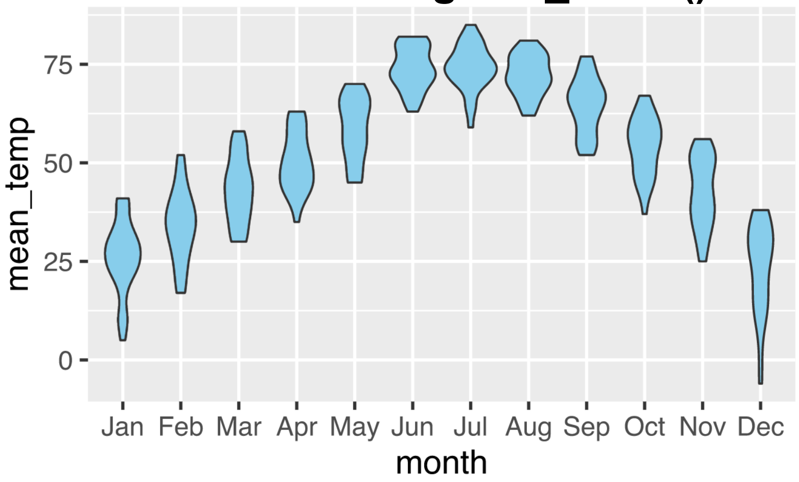
All Data Points: geom_point() [with jitter]



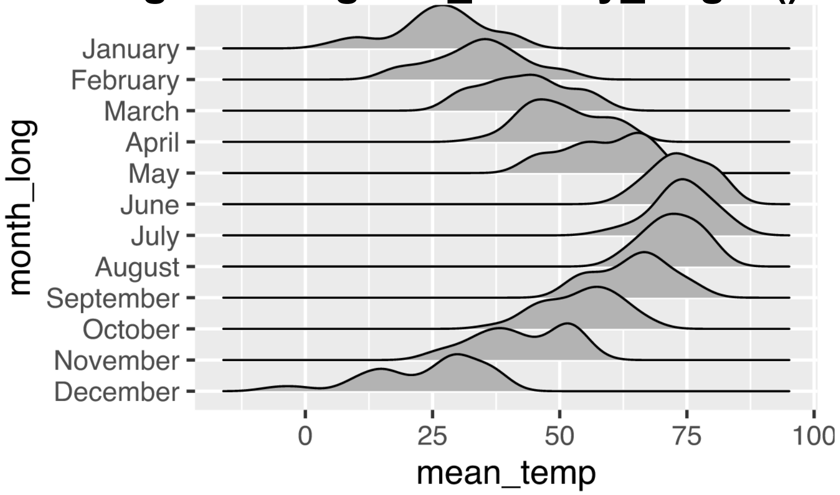
Box Plots: geom_boxplot()



Violin Plots: geom_violin()



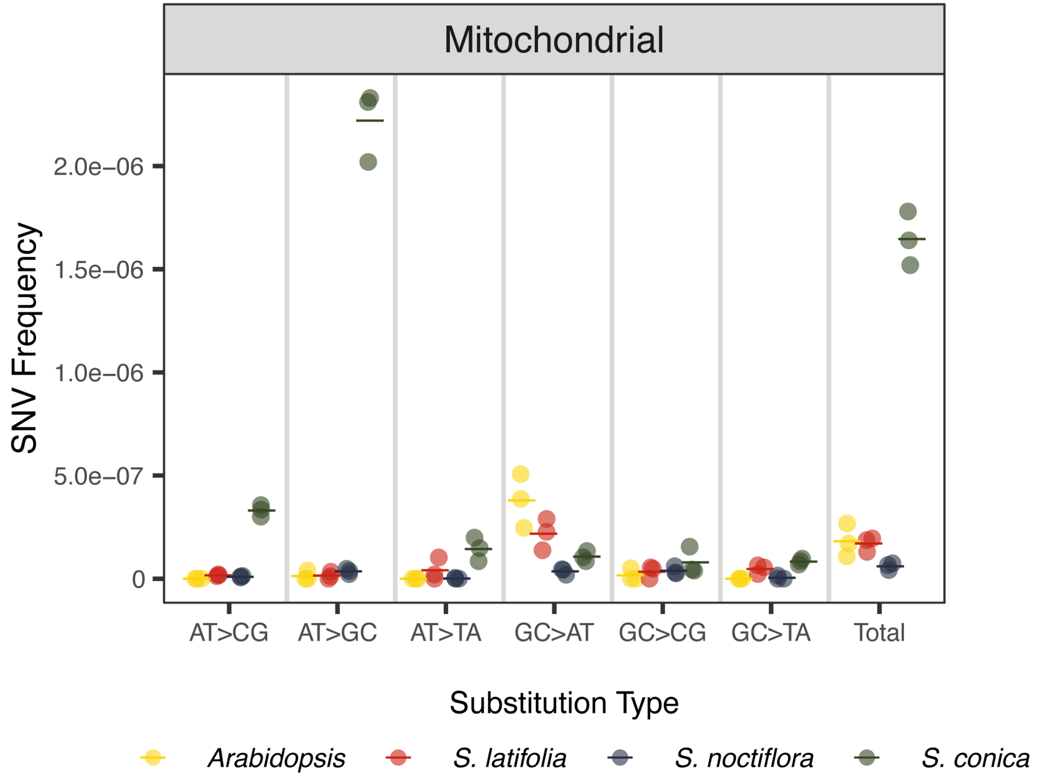
Ridge Plots: geom_density_ridges()



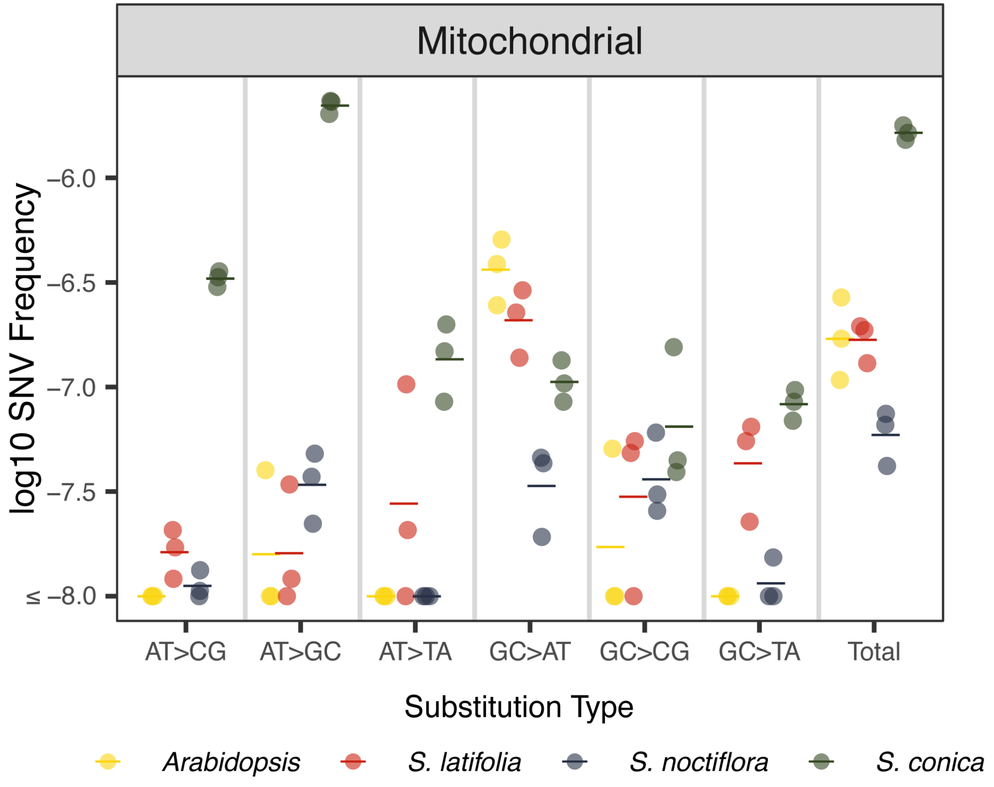
Examples from Claus Wilke: https://wilkelab.org/SDS375/slides/visualizing-distributions-2.html#1

Linear vs. Log Scales

Linear Scale



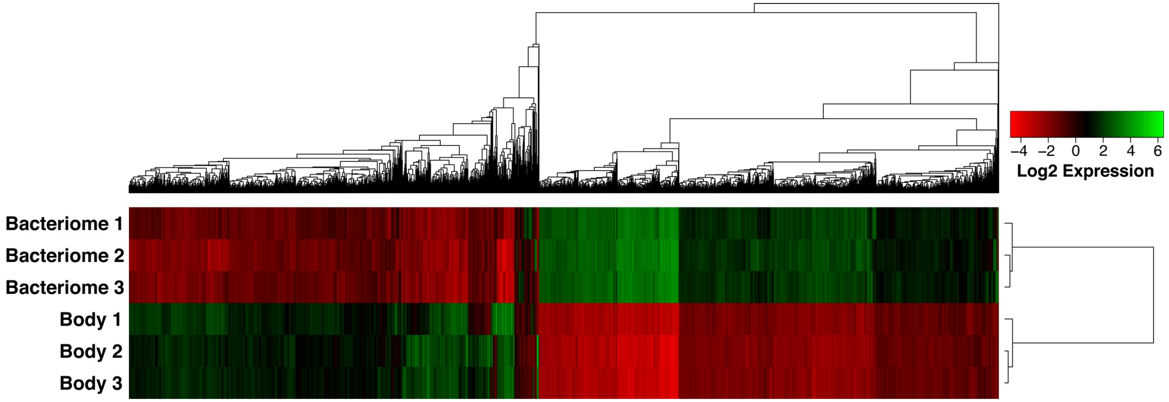
Log Scale



- 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 bind individuals….



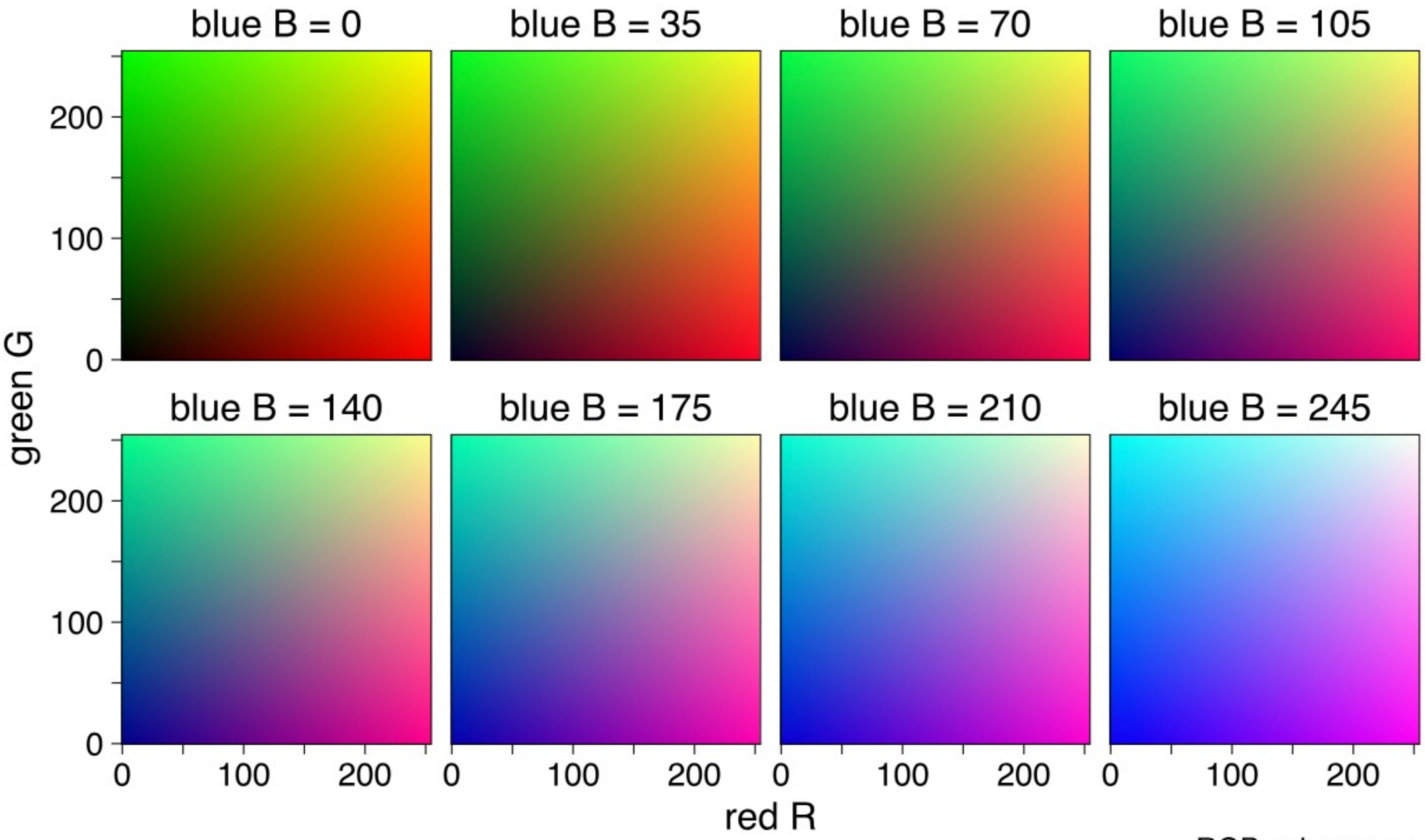
- Use palettes consisting of colors that are more distinguishable for individuals with common forms of color blindness.

- 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



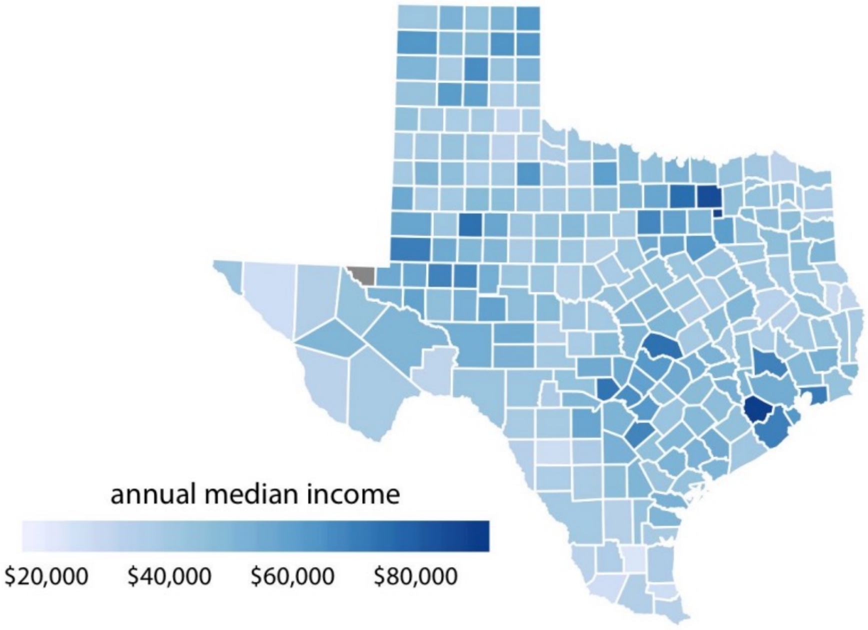
RGB color space

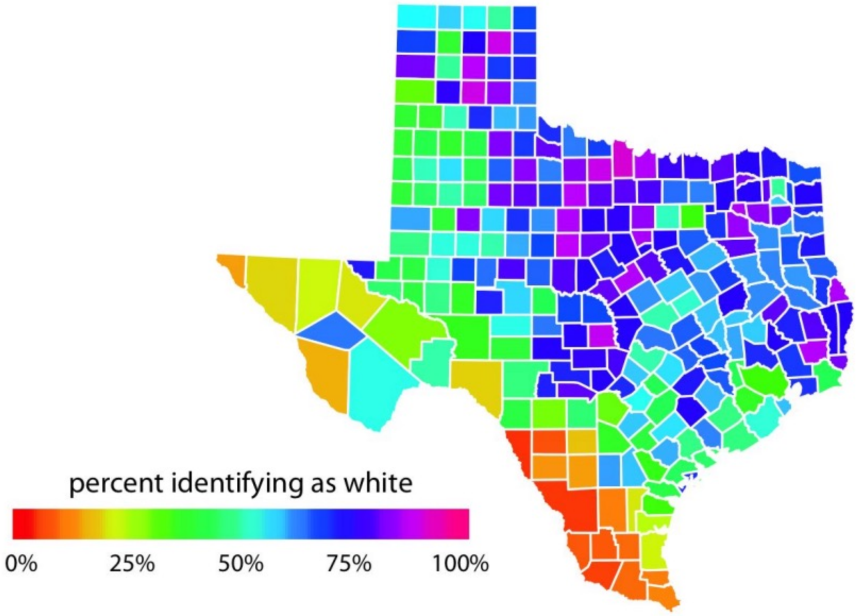
https://clauswilke.com/art/post/colors-color-spaces

Go Easy on Our Eyes

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

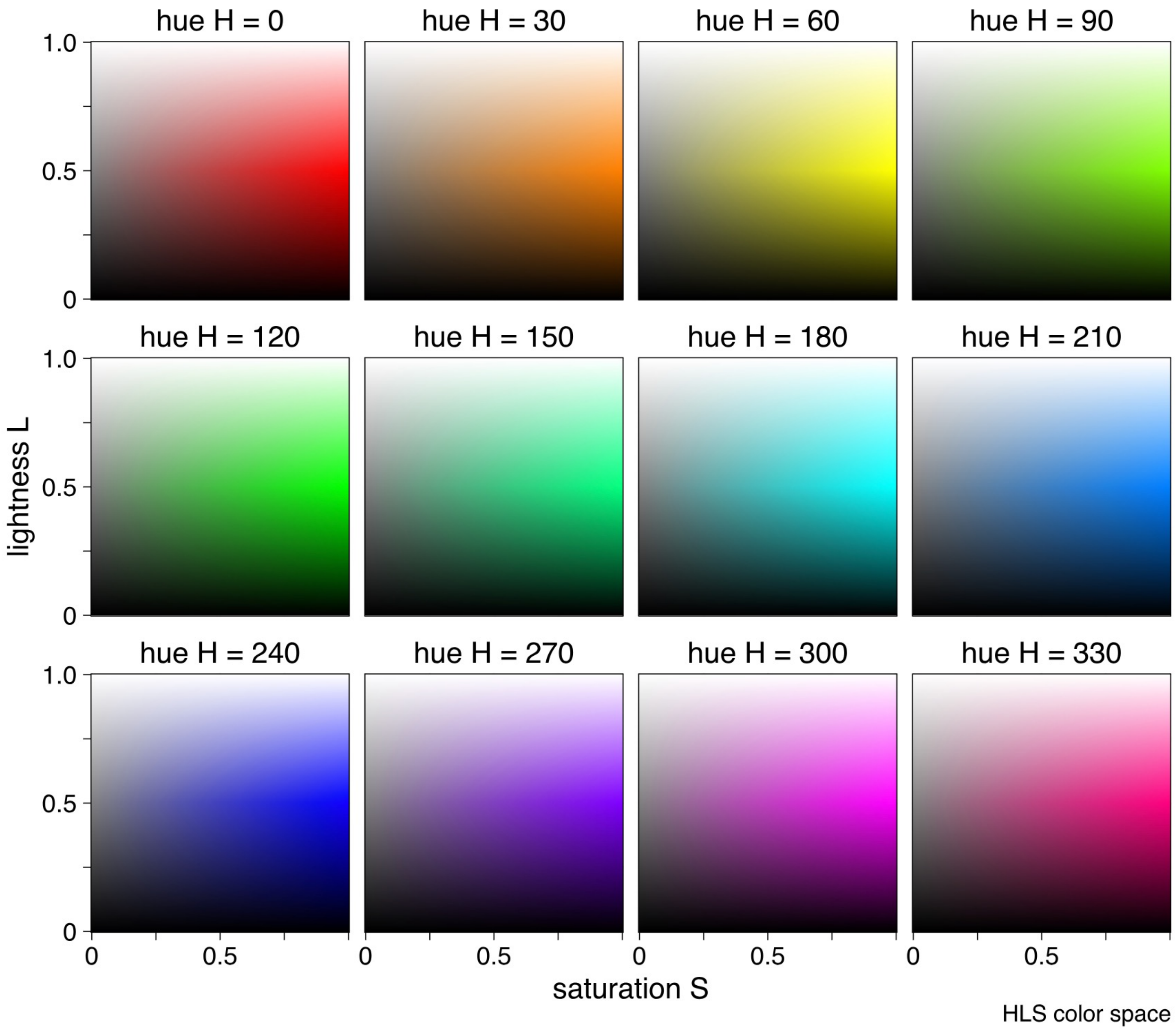
🟥	red
🟧	orange
🟨	yellow
🟩	green
🟦	blue
🟪	purple





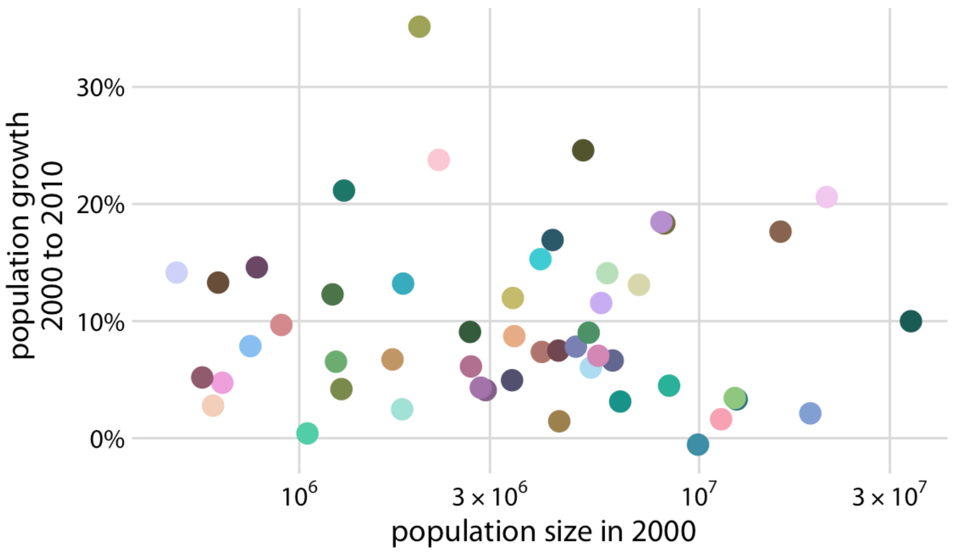
Jess Cohen-Tanugi

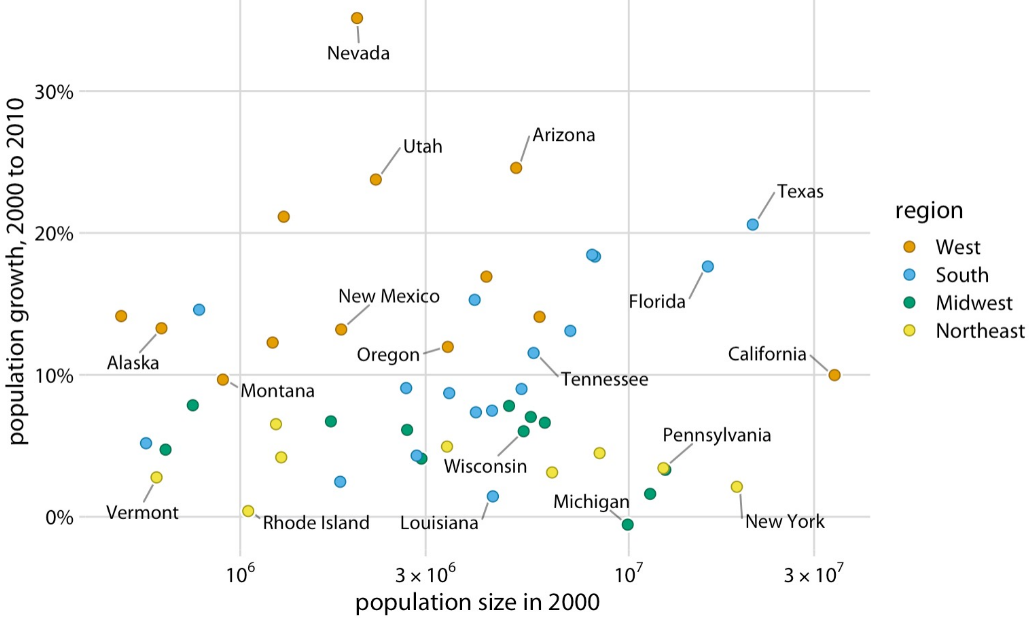
Thinking of Colors in Terms of Hue, Lightness and Saturation

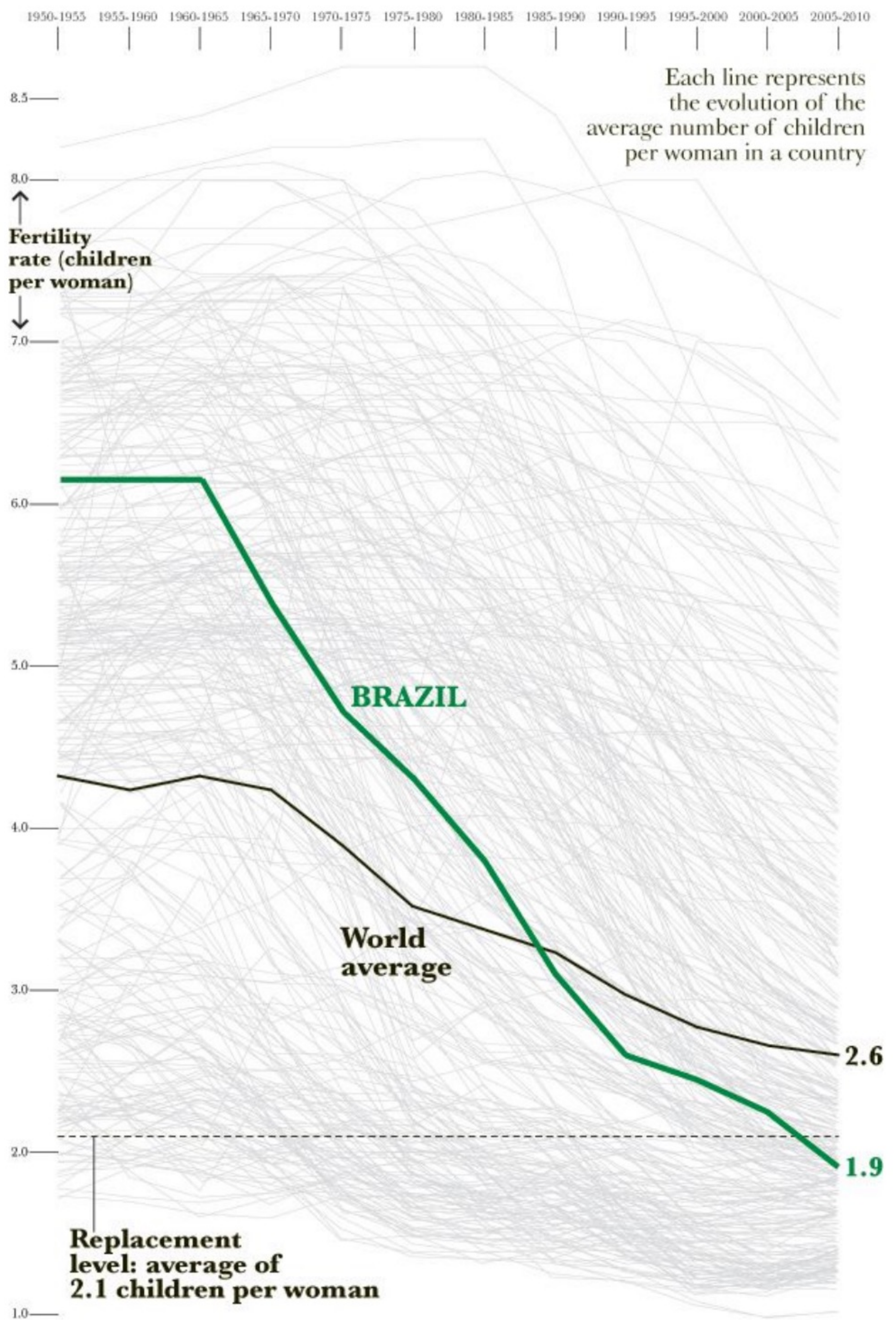


https://clauswilke.com/art/post/colors-color-spaces

Color to Emphasize, Not to Overwhelm







Jess Cohen-Tanugi

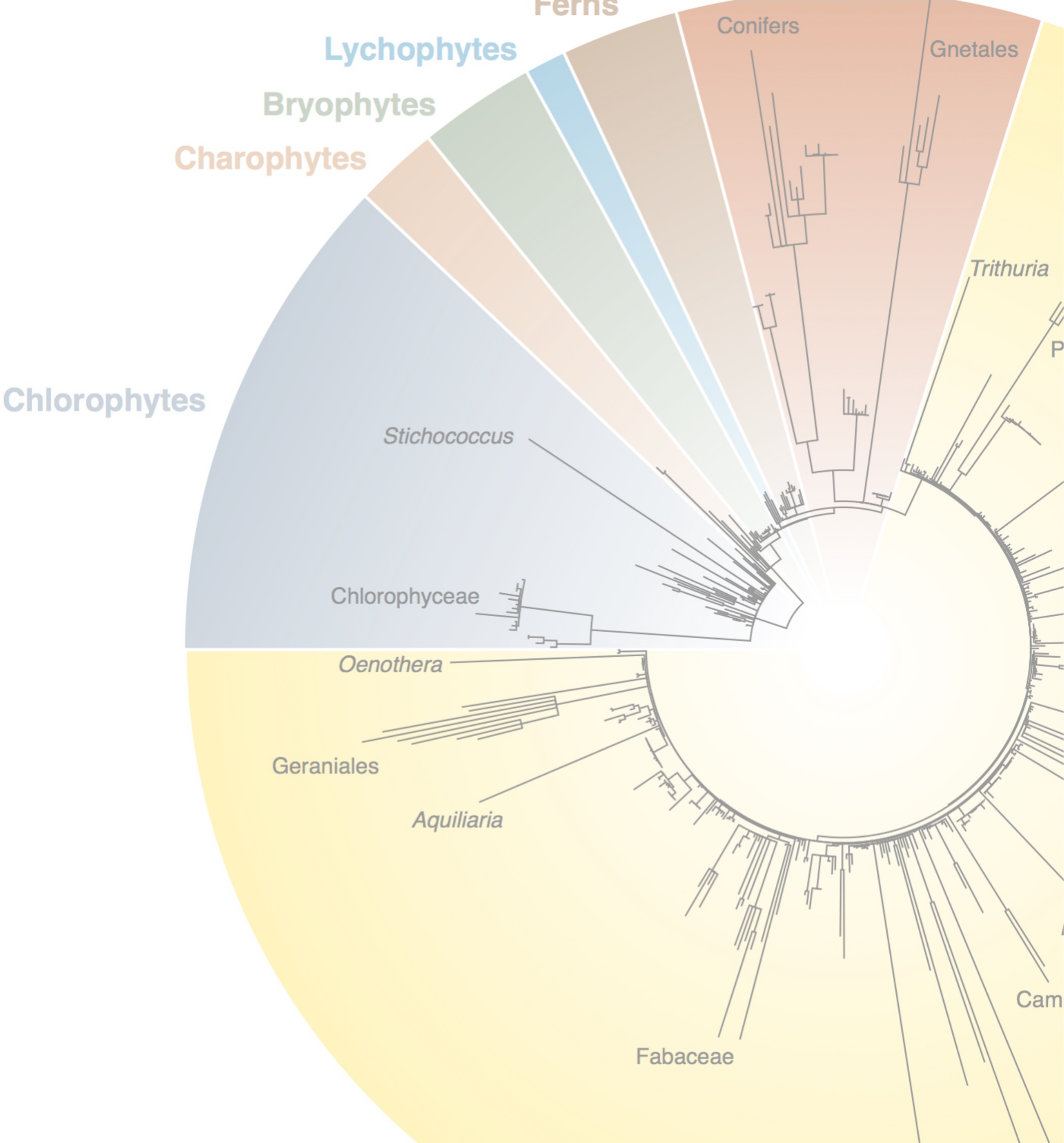
https://clauswilke.com/dataviz/color-pitfalls.html

Exercise and Assignment

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

Tree-Like Data Structures in Biology

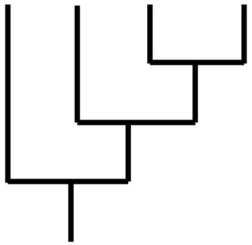
- The tree of life and species relationships
- Gene family evolution
- Hierarchical clustering of gene expression patterns, ecological/microbiome communities, etc.



Newick Tree Format

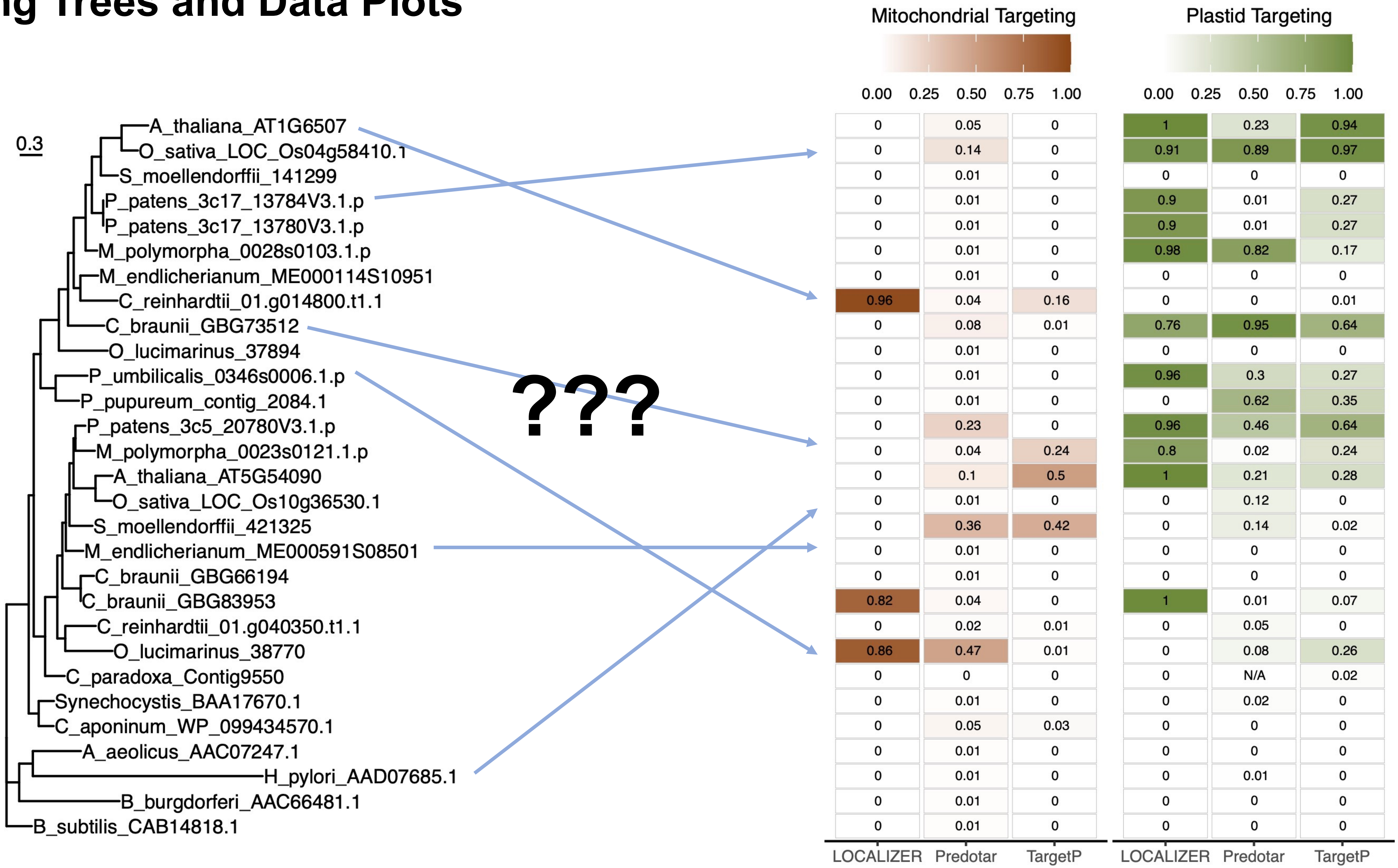
- Standard parentheses/comma-based format for summarizing tree branching patterns

(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



Mitochondrial Targeting			Plastid Targeting		
LOCALIZER	Predotar	TargetP	LOCALIZER	Predotar	TargetP
0	0.05	0	1	0.23	0.94
0	0.14	0	0.91	0.89	0.97
0	0.01	0	0	0	0
0	0.01	0	0.9	0.01	0.27
0	0.01	0	0.9	0.01	0.27
0	0.01	0	0.98	0.82	0.17
0	0.01	0	0	0	0
0.96	0.04	0.16	0	0	0.01
0	0.08	0.01	0.76	0.95	0.64
0	0.01	0	0	0	0
0	0.01	0	0.96	0.3	0.27
0	0.01	0	0	0.62	0.35
0	0.23	0	0.96	0.46	0.64
0	0.04	0.24	0.8	0.02	0.24
0	0.1	0.5	1	0.21	0.28
0	0.01	0	0	0.12	0
0	0.36	0.42	0	0.14	0.02
0	0.01	0	0	0	0
0	0.01	0	0	0	0
0.82	0.04	0	1	0.01	0.07
0	0.02	0.01	0	0.05	0
0.86	0.47	0.01	0	0.08	0.26
0	0	0	0	N/A	0.02
0	0.01	0	0	0.02	0
0	0.05	0.03	0	0	0
0	0.01	0	0	0	0
0	0.01	0	0	0.01	0
0	0.01	0	0	0	0
0	0.01	0	0	0	0

Exercise and Assignment

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