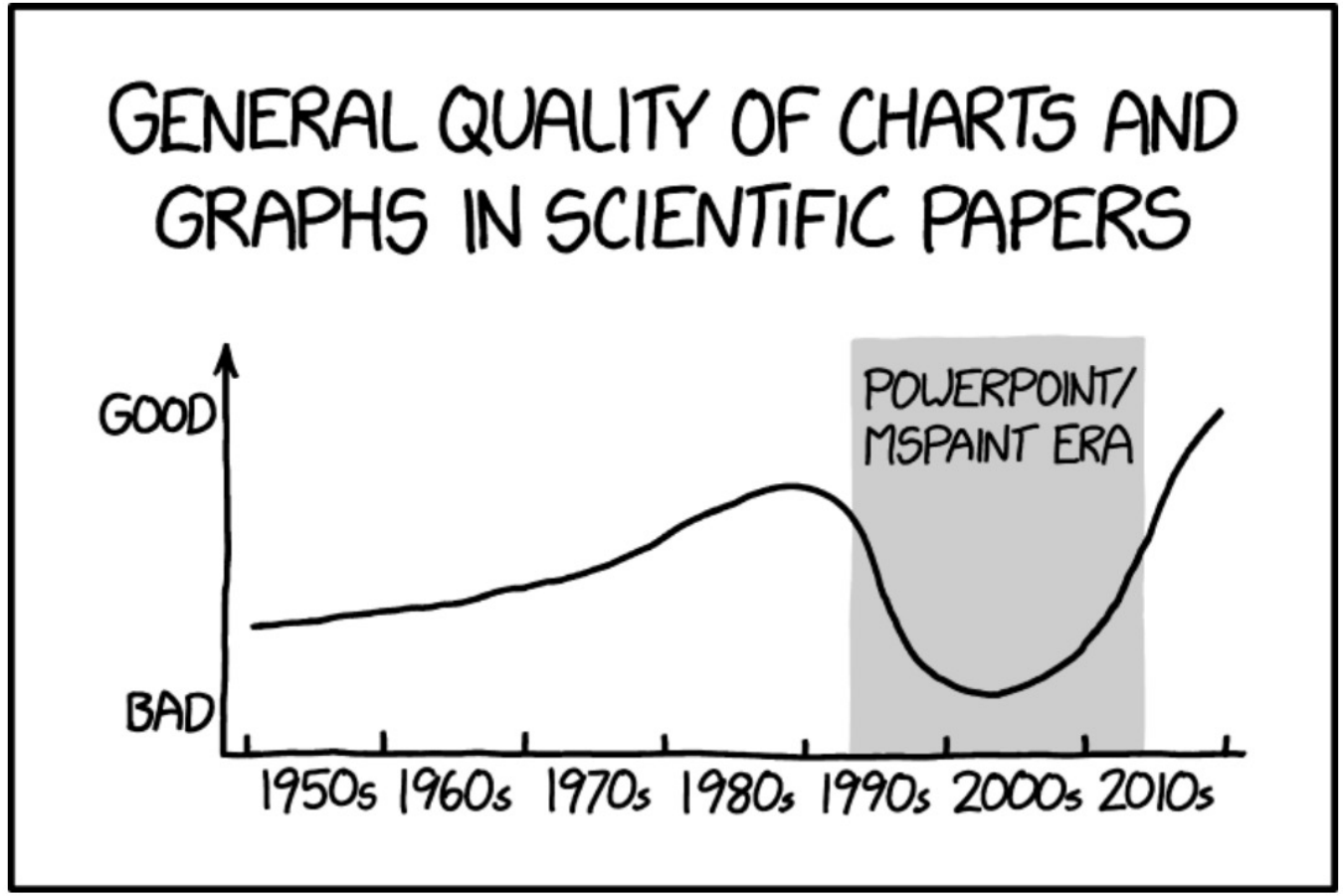
Visualizing Data

Biology 683

Week 2

Heath Blackmon

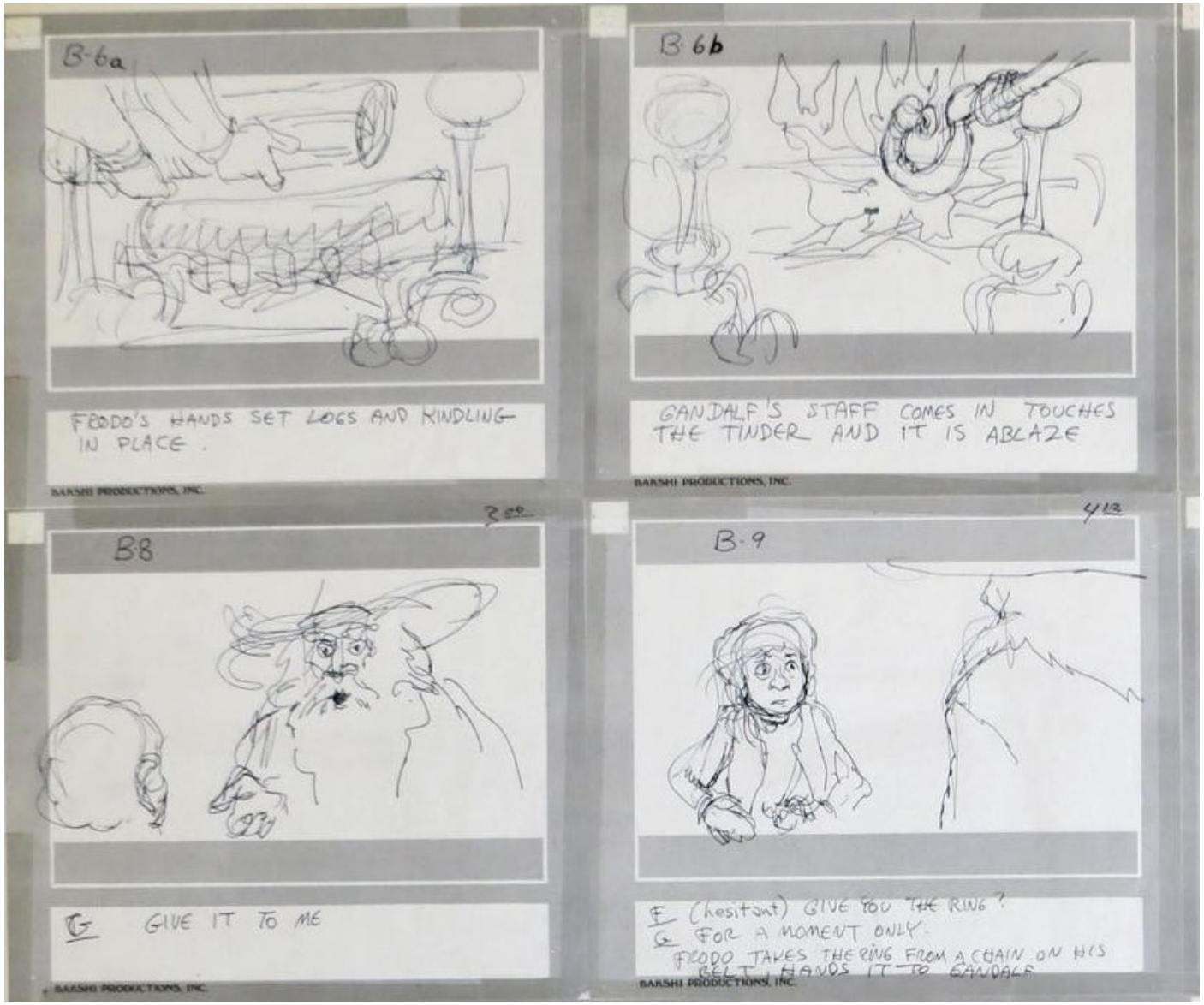


- What is the difference in a matrix and a dataframe?
- What is a p-value
- What are the types of data used in R

Plan for today

1. Importance of figures
2. General rules for making plots
3. Programs for plotting
4. Resources for plotting

Importance of figures



Serve a purpose

Meiotic drive shapes rates of karyotype evolution in mammals

Heath Blackmon,[1,2] Joshua Justison,[3] Itay Mayrose,[4] and Emma E. Goldberg[3]

[1]Department of Biology, Texas A&M University, College Station Texas 77843

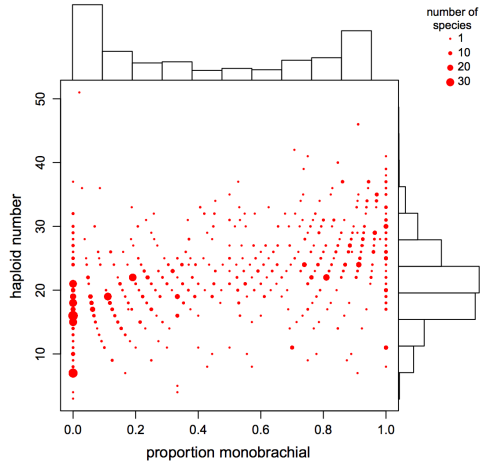
[2]E-mail: coleoguy@gmail.com

[3]Department of Ecology, Evolution, and Behavior, University of Minnesota, Saint Paul Minnesota 55108

[4]School of Plant Sciences and Food Security, Tel Aviv University Tel Aviv 69978, Israel

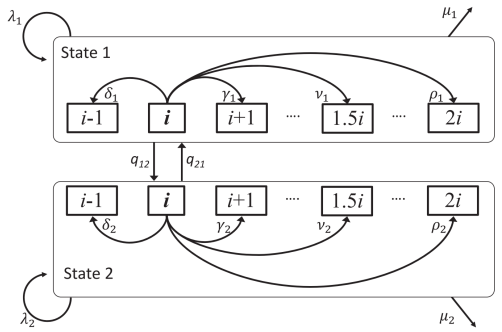
Serve a purpose

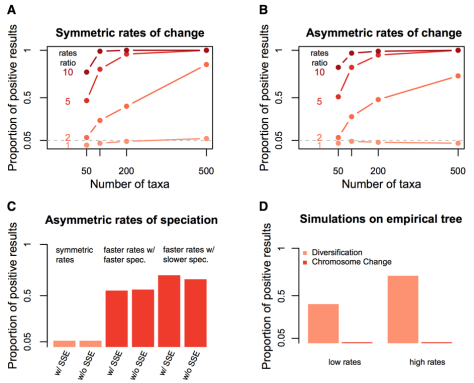
The Problem/Question



Mammals have a weird distribution of numbers and types of chromosomes

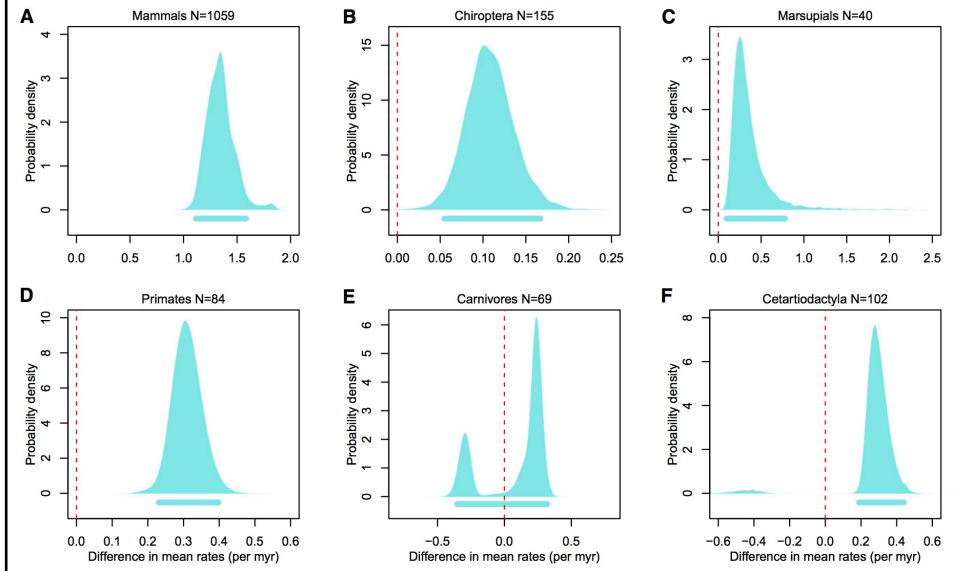
New tool and it works





I made this new model and it can test this hypothesis and it works pretty good

The Answer

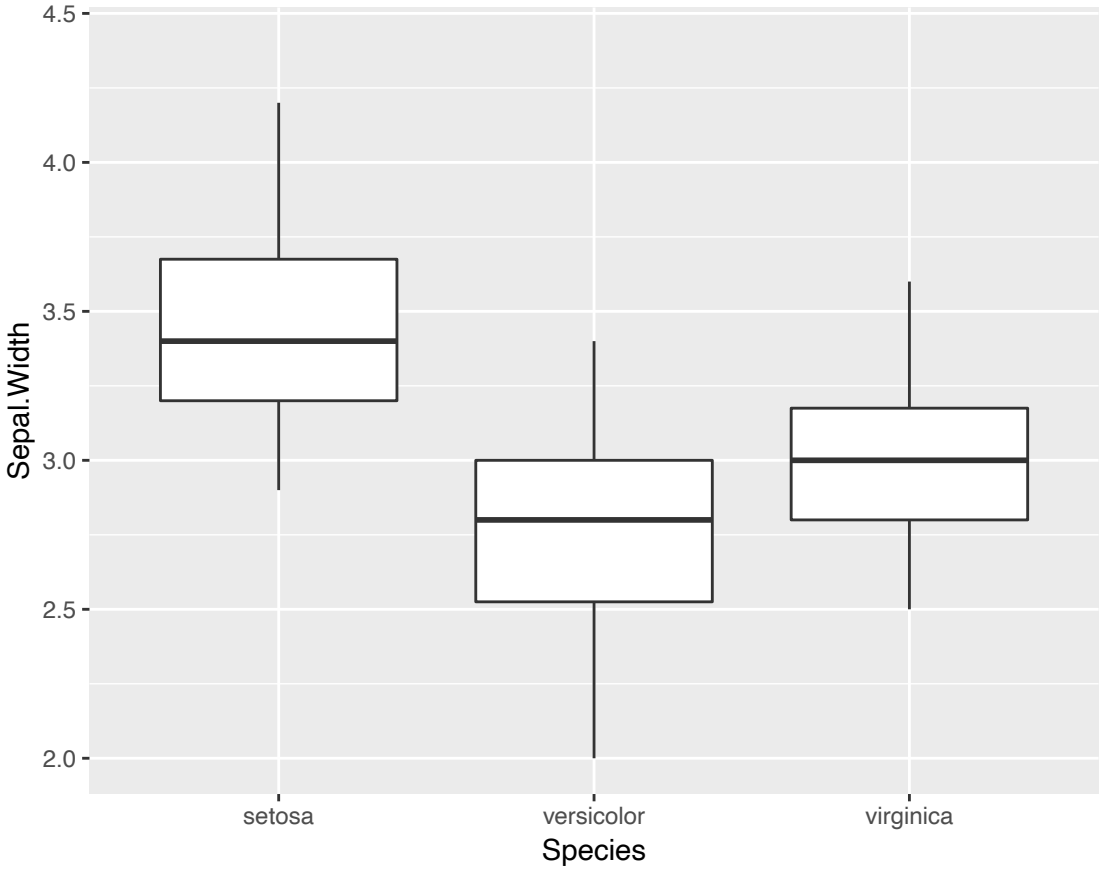


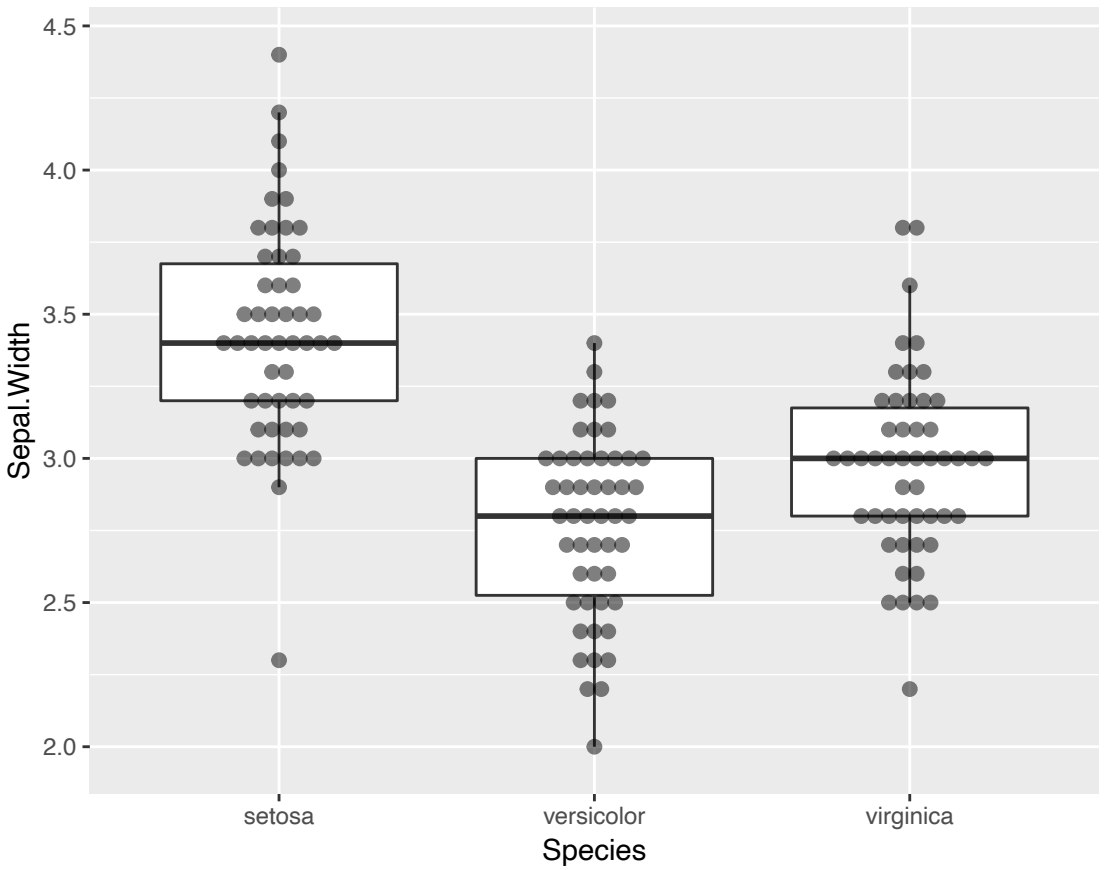
For most groups of mammals I show strong support for the existing hypothesis we should assume it is right but maybe not in 1-2 groups.

Rules for plots

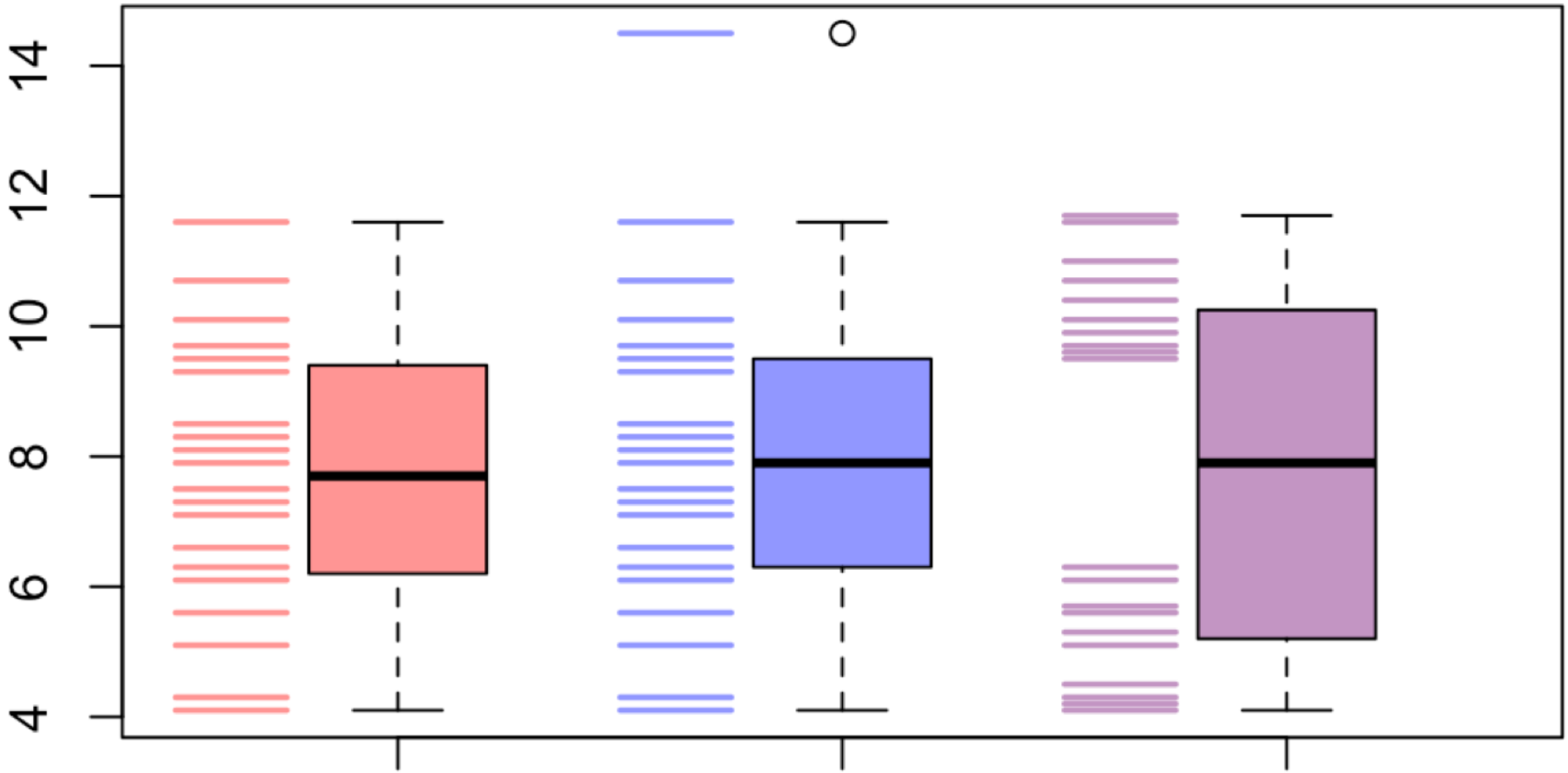
1. Show the data

Show the data





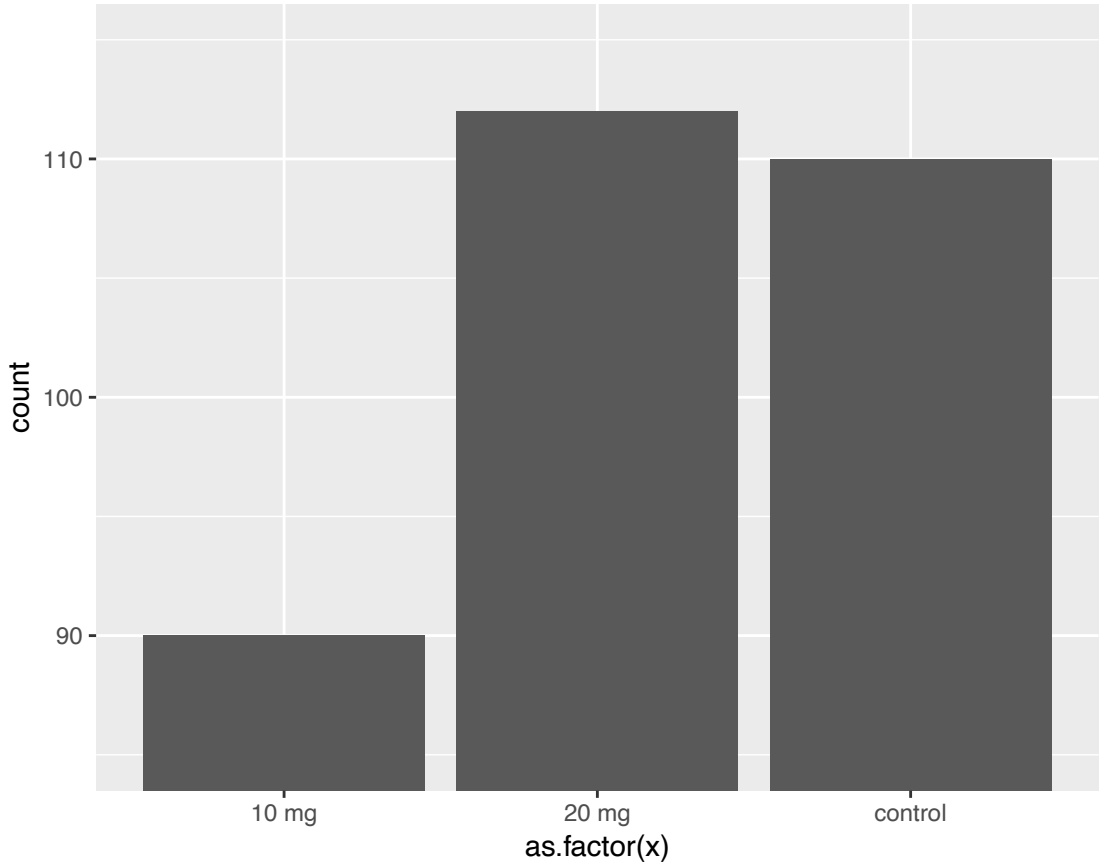
Showing the data can reveal patterns

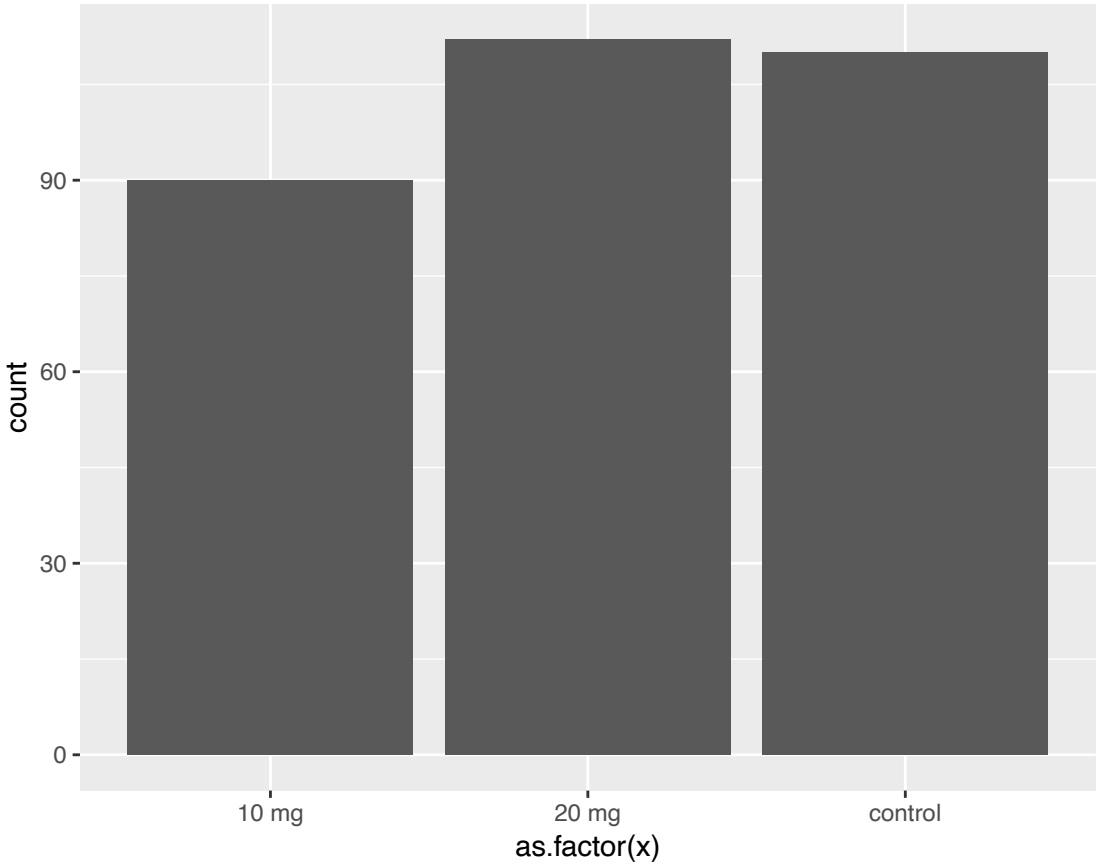


Rules for plots

1. Show the data
2. Avoid distorting data

Avoid distorting data

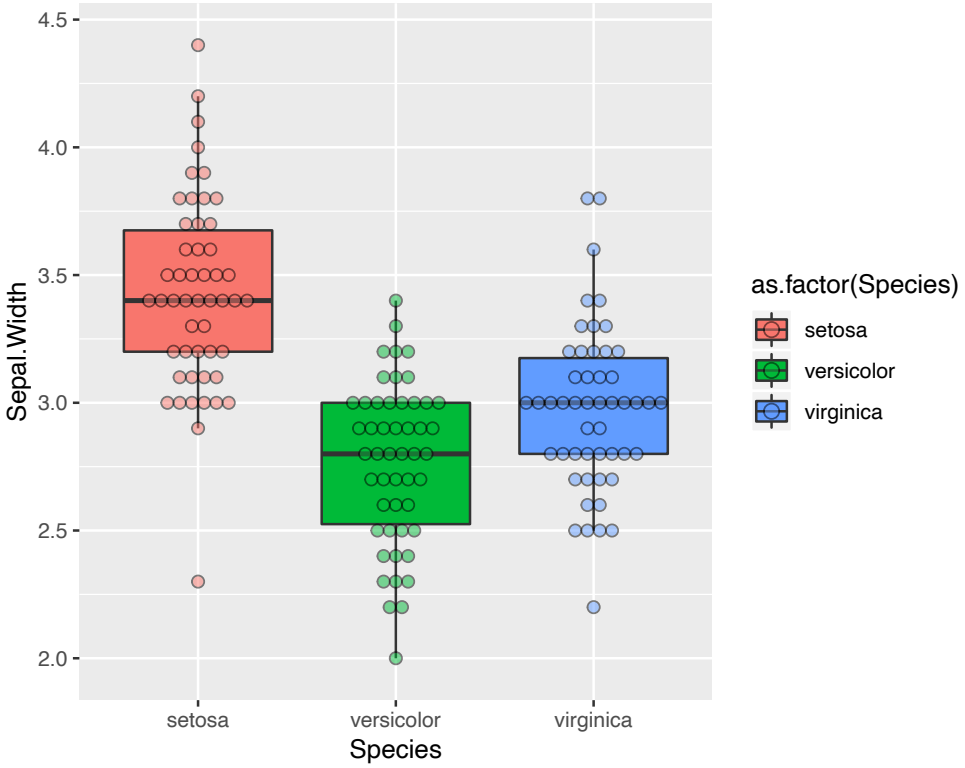


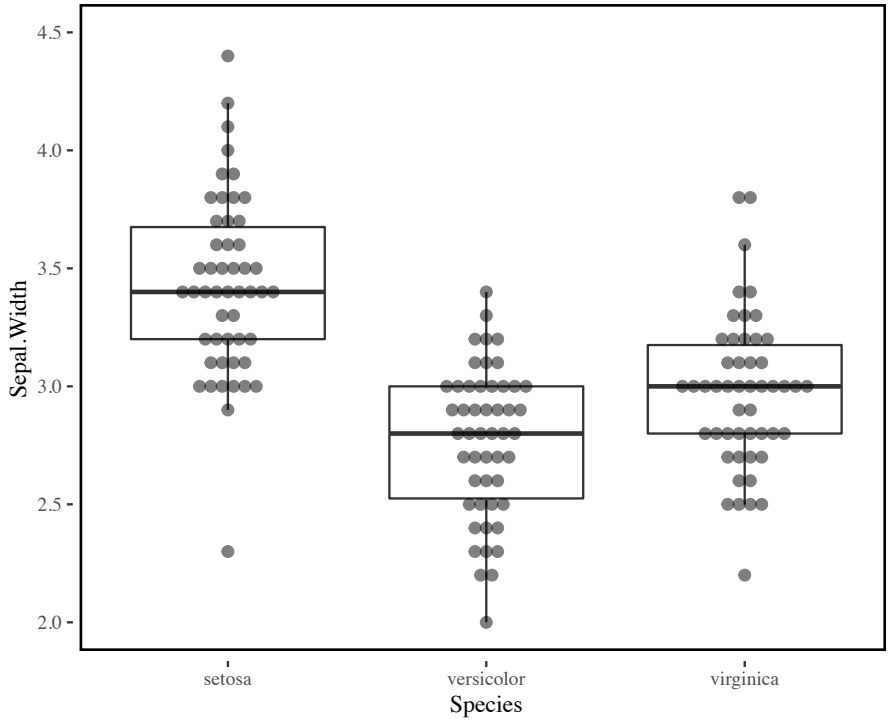


Rules for plots

1. Show the data
2. Avoid distorting data
3. Avoid chart junk

Avoid chart junk

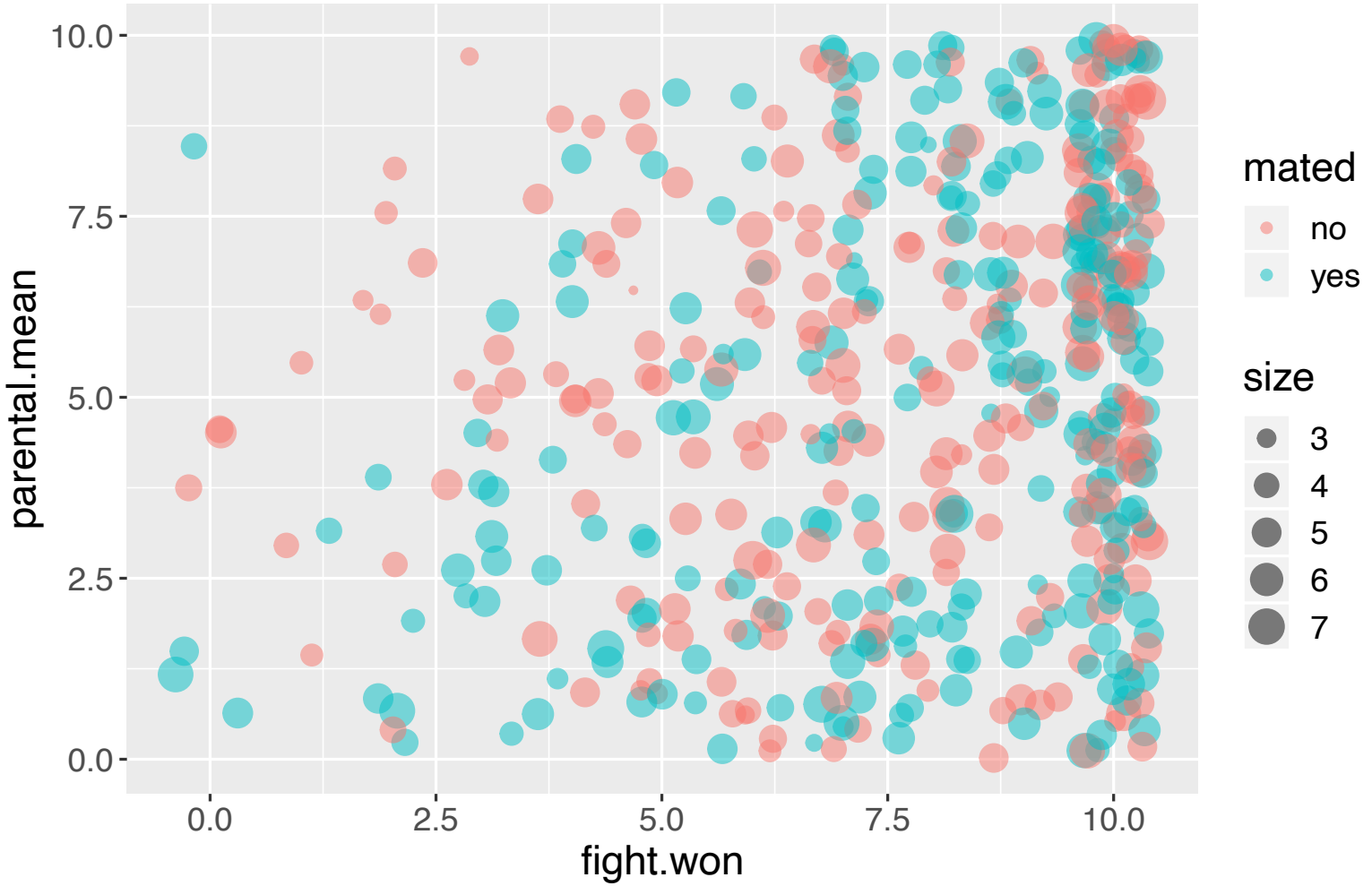




Rules for plots

1. Show the data
2. Avoid distorting data
3. Avoid chart junk
4. Maximize data:ink ratio maximize information

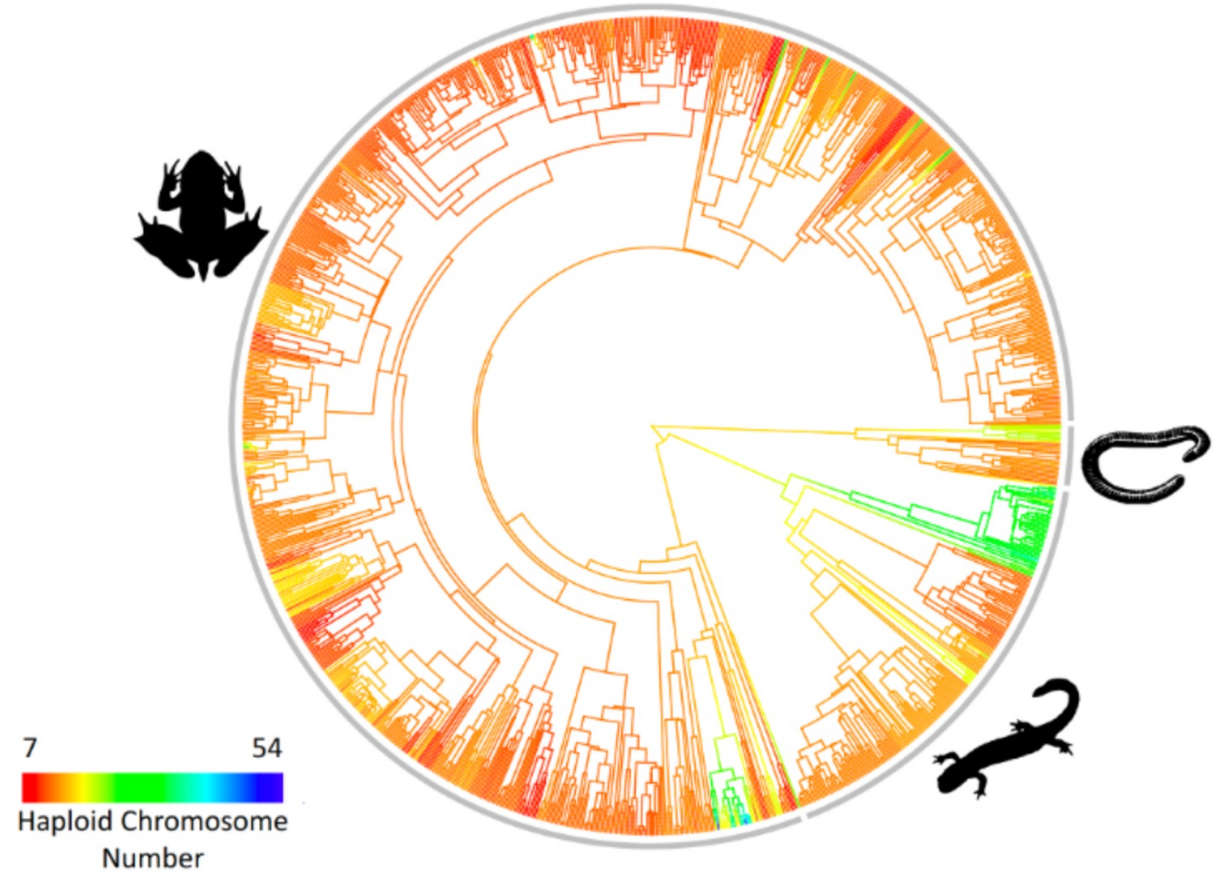
Maximize data:ink ratio maximize information



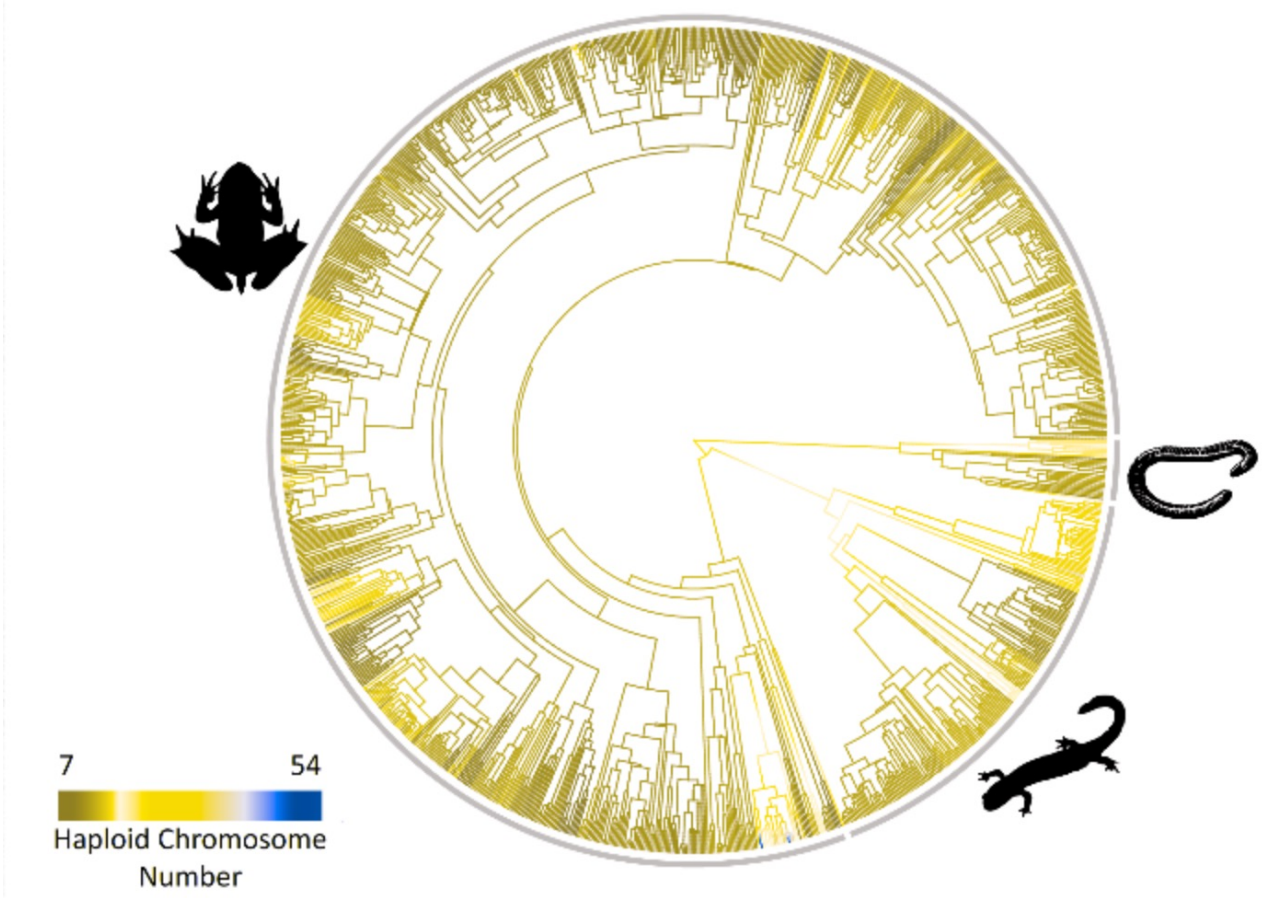
Rules for plots

1. Show the data
2. Avoid distorting data
3. Avoid chart junk
4. Maximize data:ink ratio maximize information
5. Make it accessible to all (5% color blind + BW prints).

Make it accessible to all



Make it accessible to all



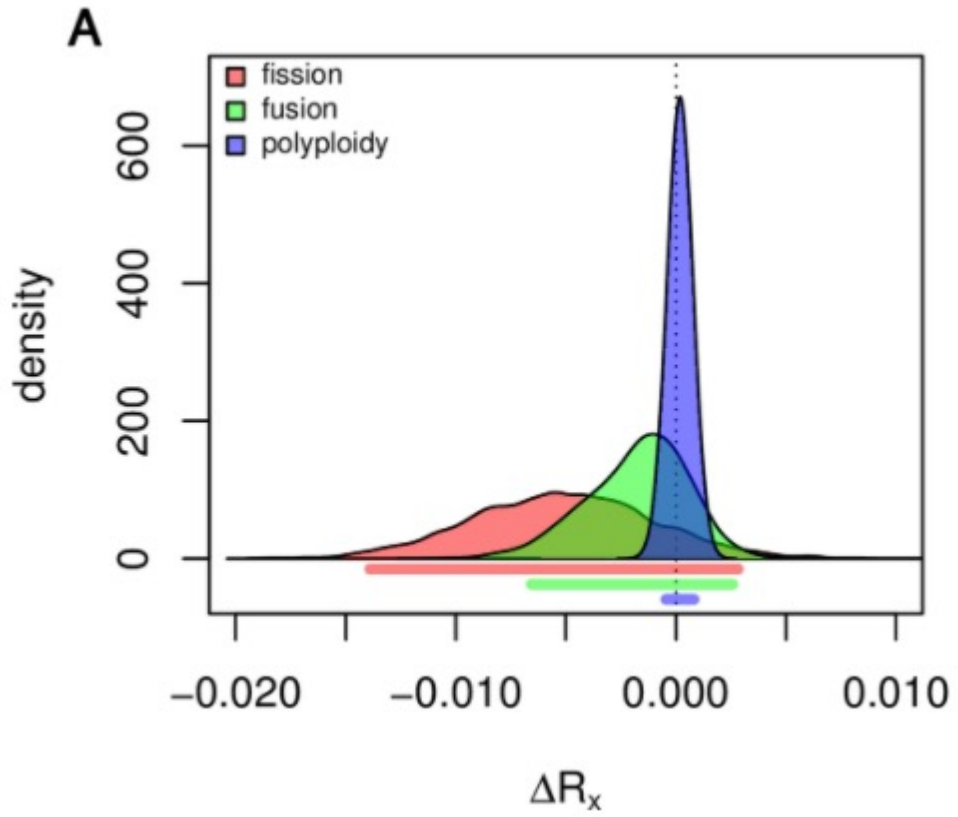
Use the viridis color palette

[Color blind simulator]

Rules for plots

1. Show the data
2. Avoid distorting data
3. Avoid chart junk
4. Maximize data:ink ratio maximize information
5. Make it accessible to all (5% color blind + BW prints).
6. Axes and legends that are informative and useful

Axes and legends

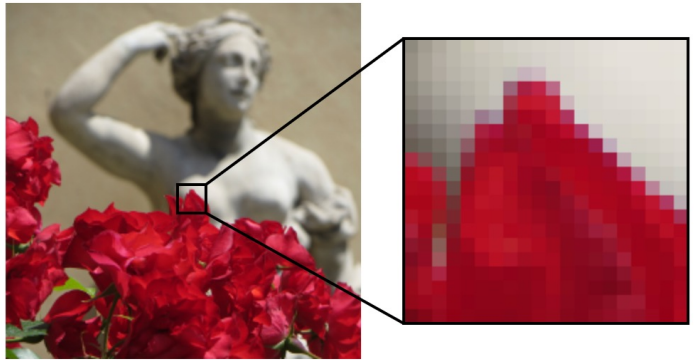


Rules for plots

1. Show the data
2. Avoid distorting data
3. Avoid chart junk
4. Maximize data:ink ratio maximize information
5. Make it accessible to all (5% color blind + BW prints).
6. Axes and legends that are informative and useful
7. Use vector art when possible

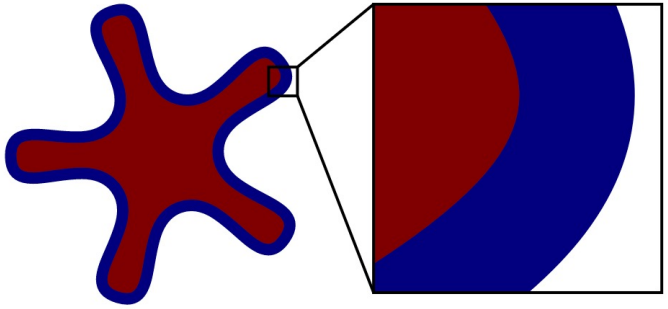
Types of images

Raster images



jpg, tiff, png, bmp, raw

Vector images



eps, pdf, svg

Example of journal requirements

Photos (edit in Adobe photoshop; gimp)

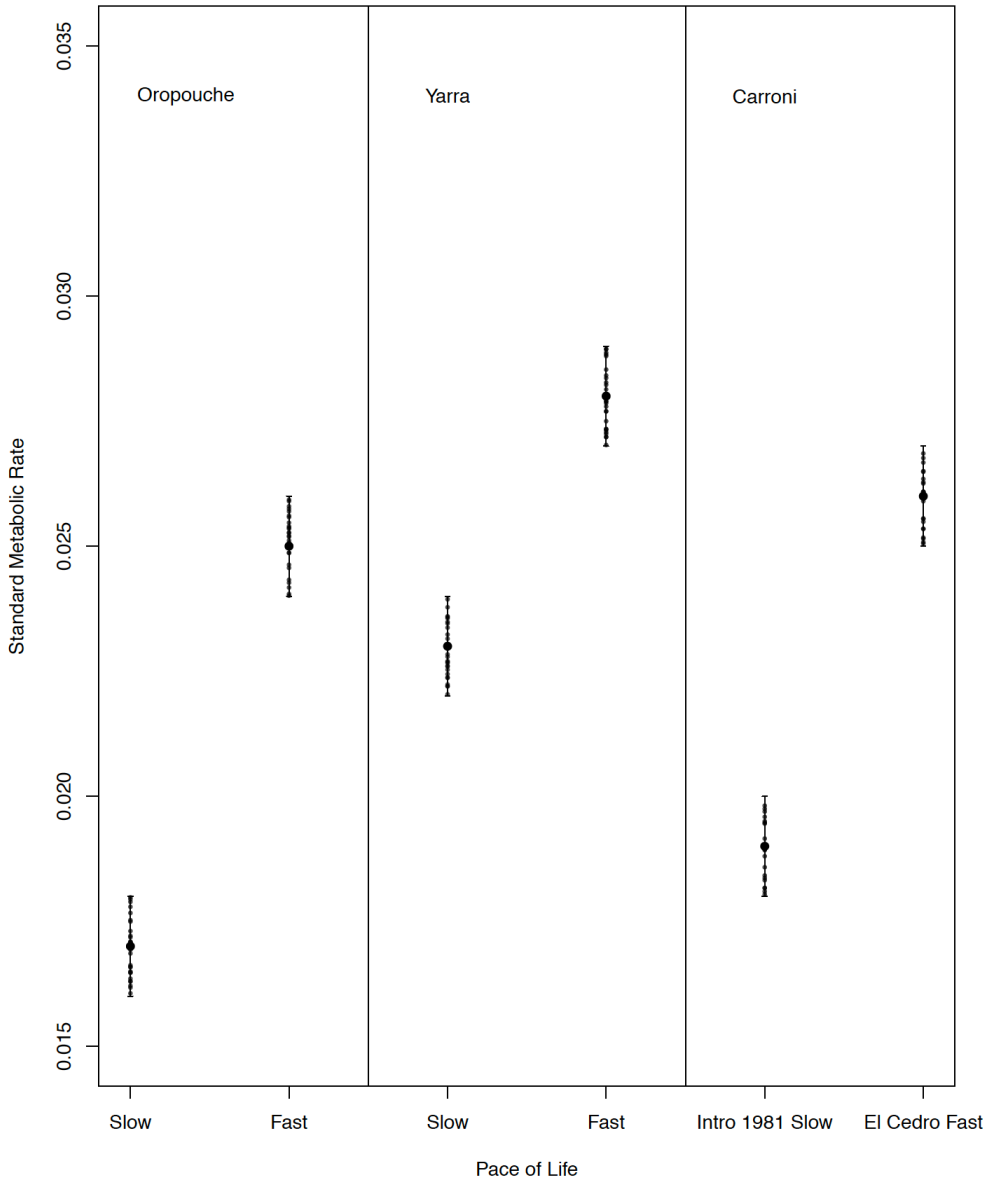
Plots (R; edit in inkscape; Adobe illustrator)

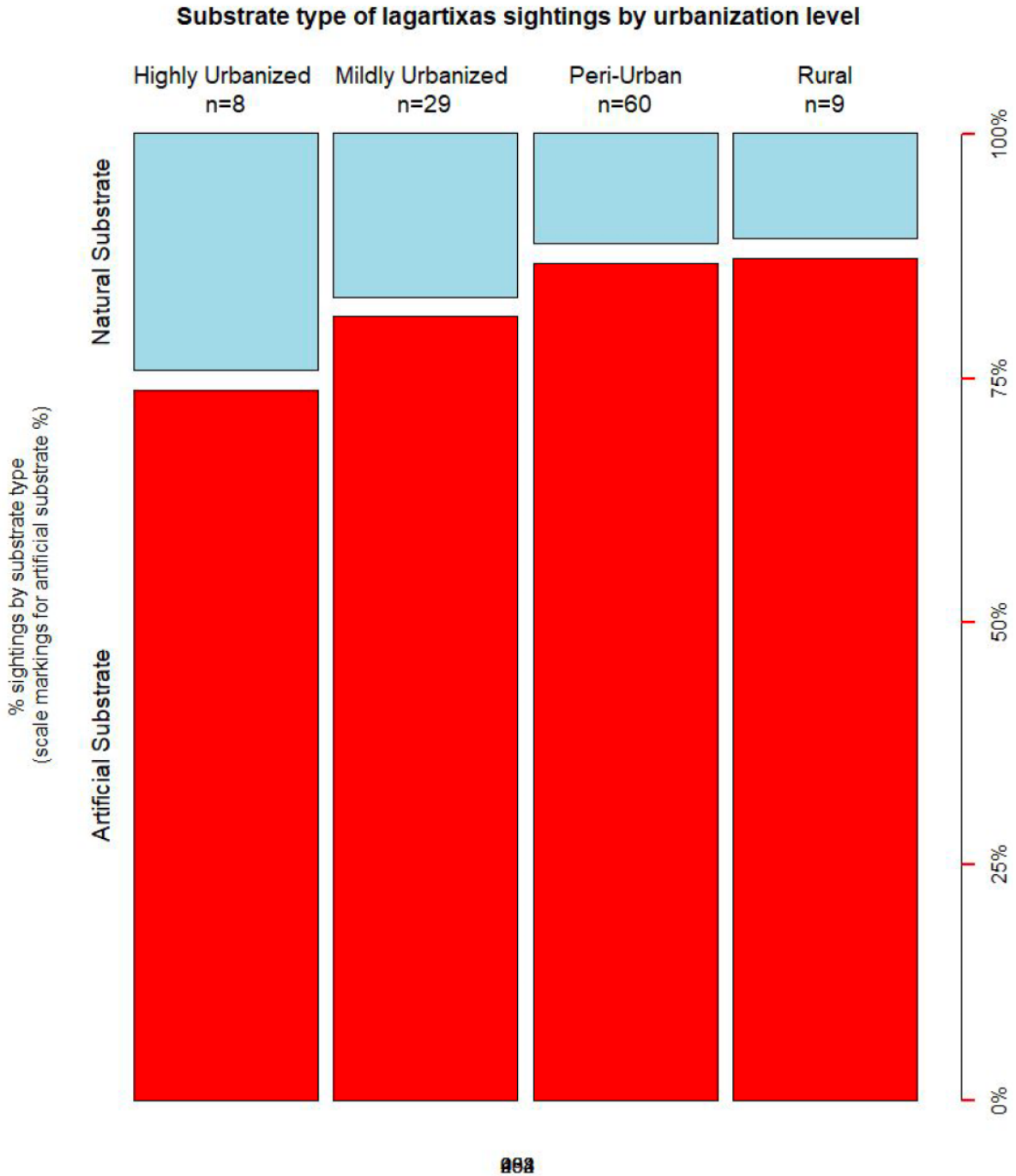
Complex plots (PPT; Adobe illustrator)

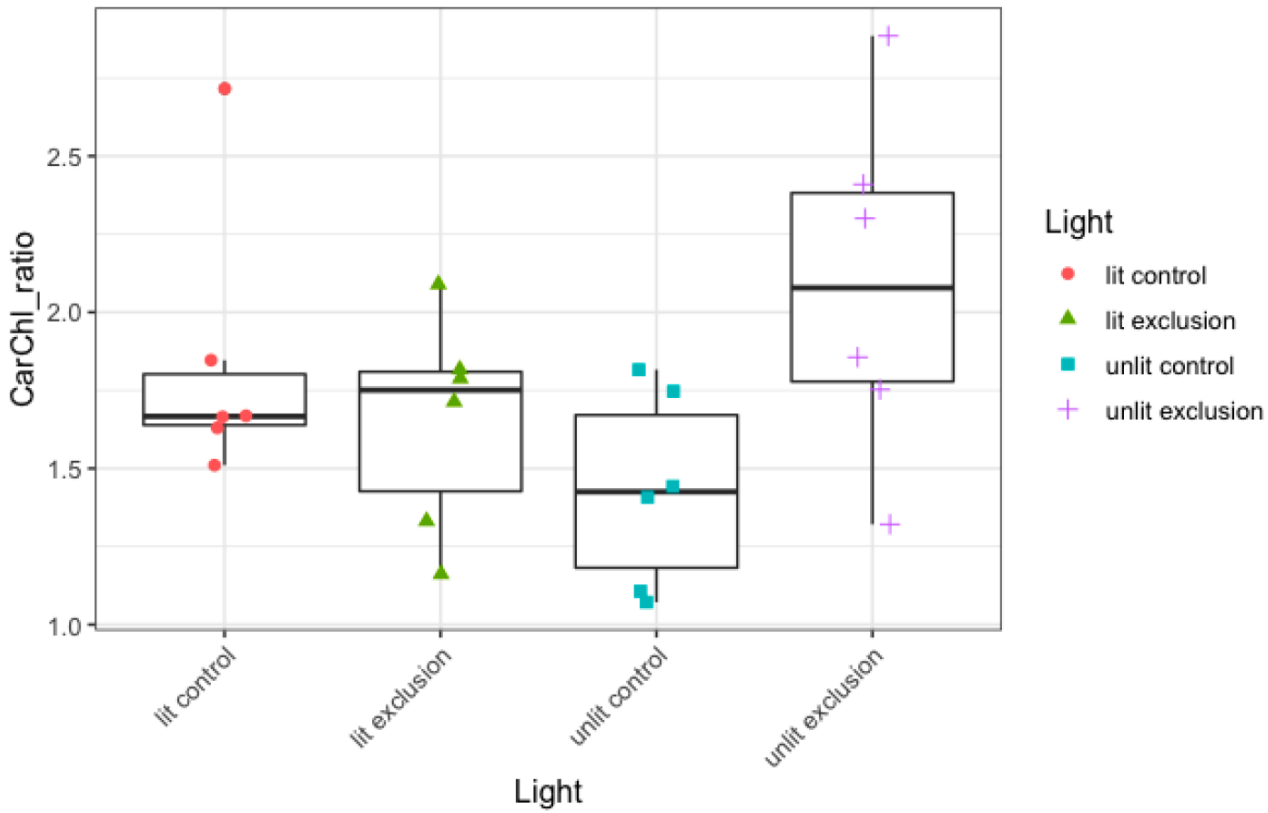
Rules for plots

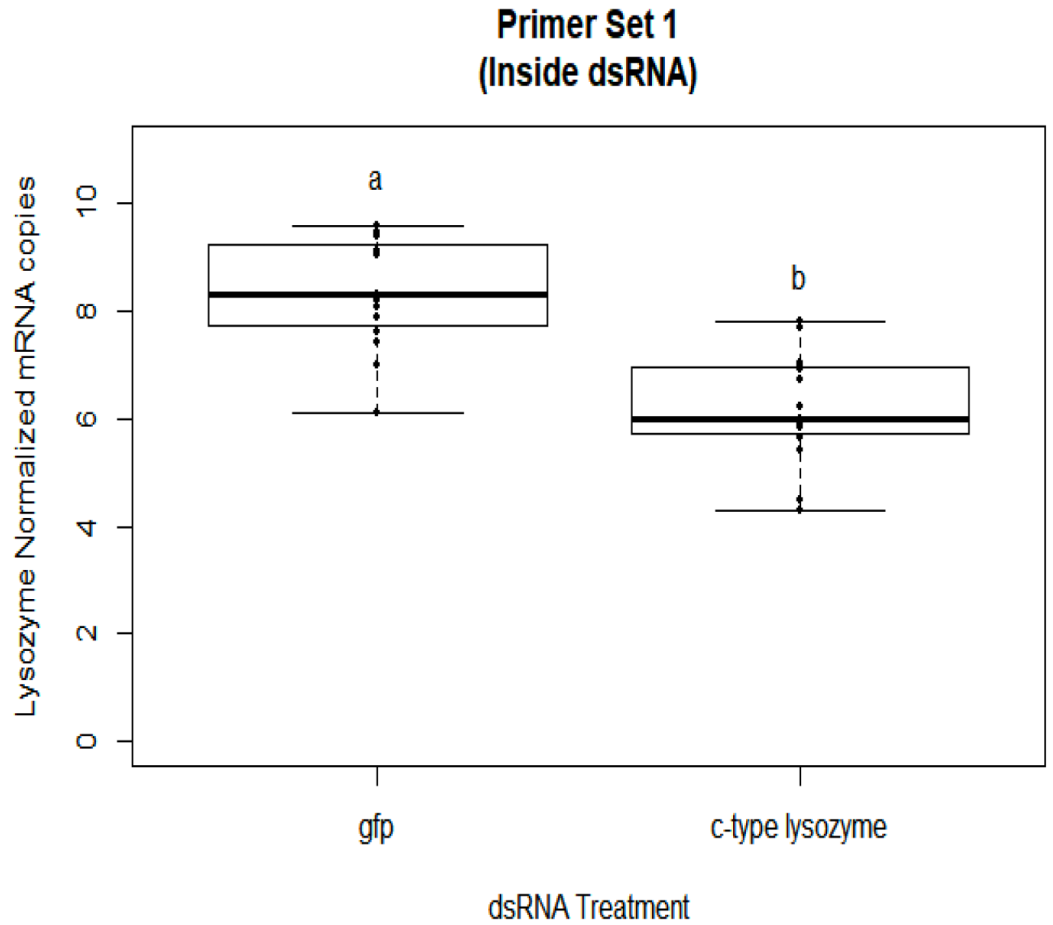
1. Show the data
2. Avoid distorting data
3. Avoid chart junk
4. Maximize data:ink ratio maximize information
5. Make it accessible to all (5% color blind + BW prints).
6. Axes and legends that are informative and useful
7. Use vector art when possible
8. Humans aren’t good at pie charts don’t use them

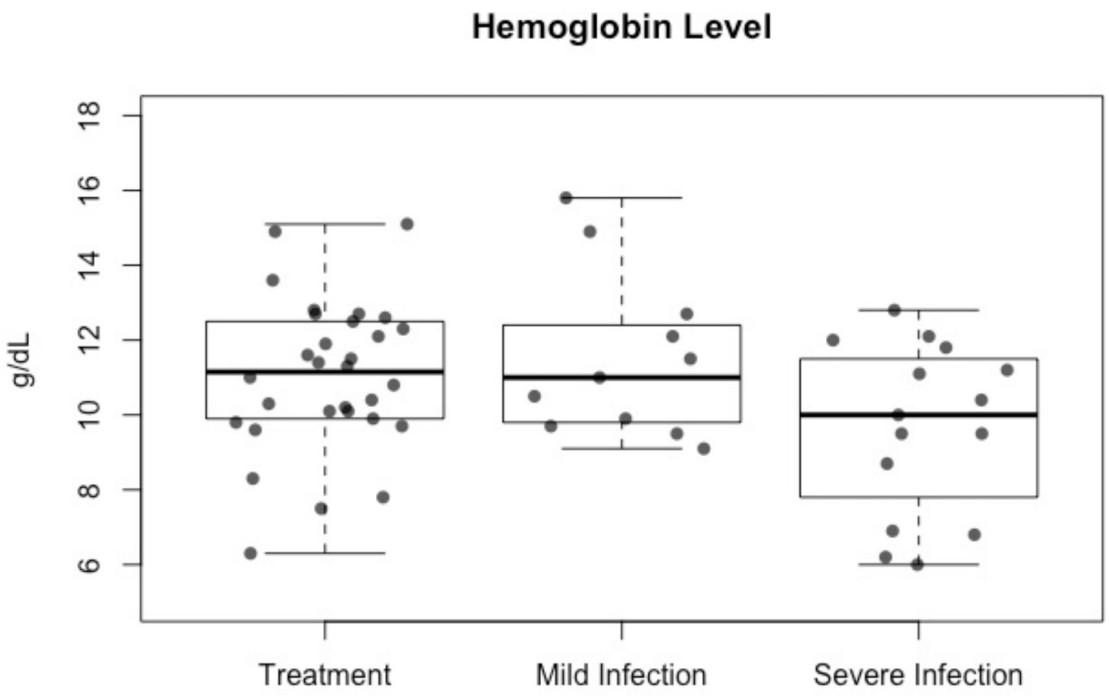
Plots with problems!


Oropouche
Yarra
Carroni
Standard Metabolic Rate
0.015
0.020
0.025
0.030
0.035
Slow
Fast
Slow
Fast
Intro 1981 Slow
El Cedro Fast
Pace of Life

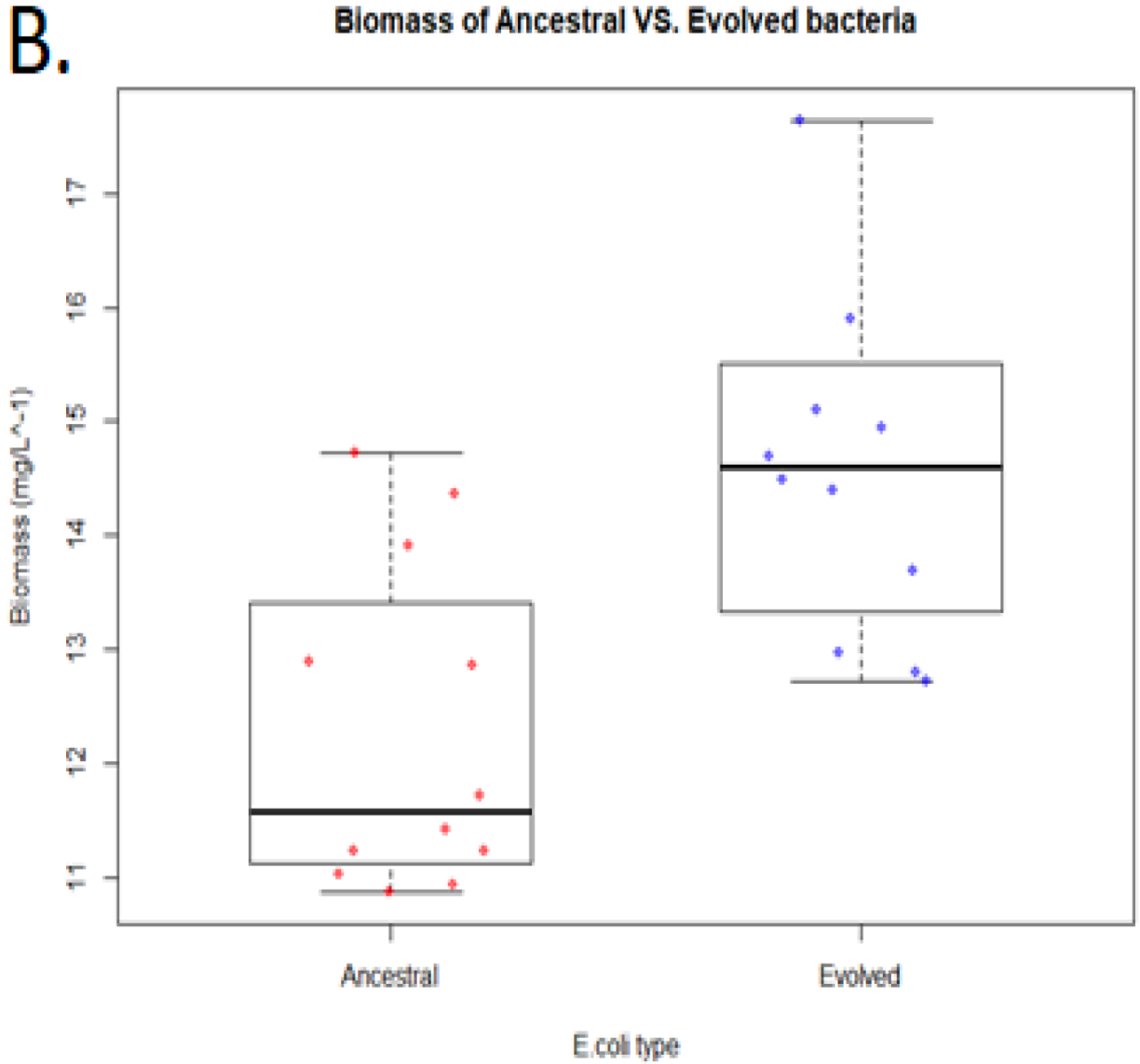

Substrate type of lagartixas sightings by urbanization level
Highly Urbanized
n=8
Mildly Urbanized
n=29
Peri-Urban
n=60
Rural
n=9
Natural Substrate
Artificial Substrate
% sightings by substrate type
(scale markings for artificial substrate %)
100%
75%
50%
25%
0%


CarChl_ratio
2.5
2.0
1.5
1.0
lit control
lit exclusion
unlit control
unlit exclusion
Light
Light
lit control
lit exclusion
unlit control
unlit exclusion

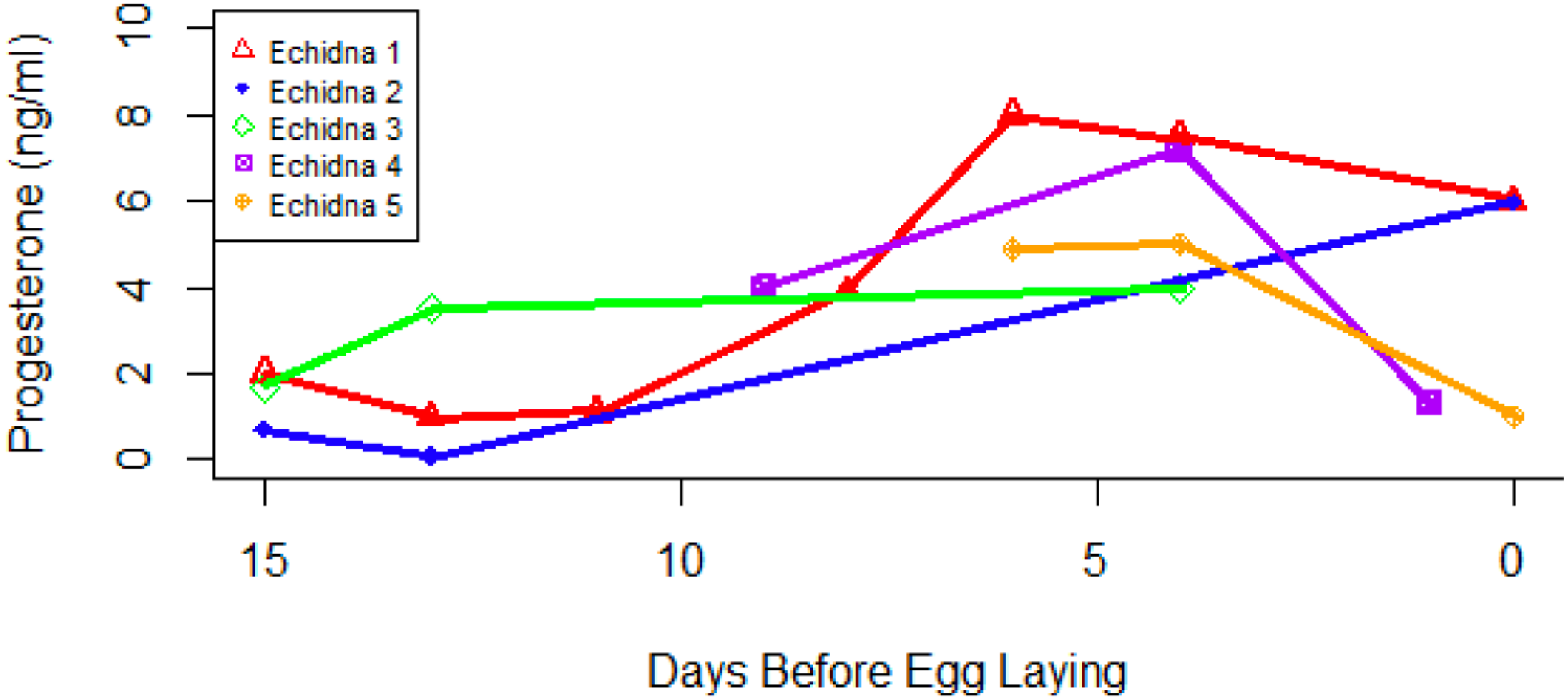

Primer Set 1
(Inside dsRNA)
Lysozyme Normalized mRNA copies
0
2
4
6
8
10
a
b
gfp
c-type lysozyme
dsRNA Treatment

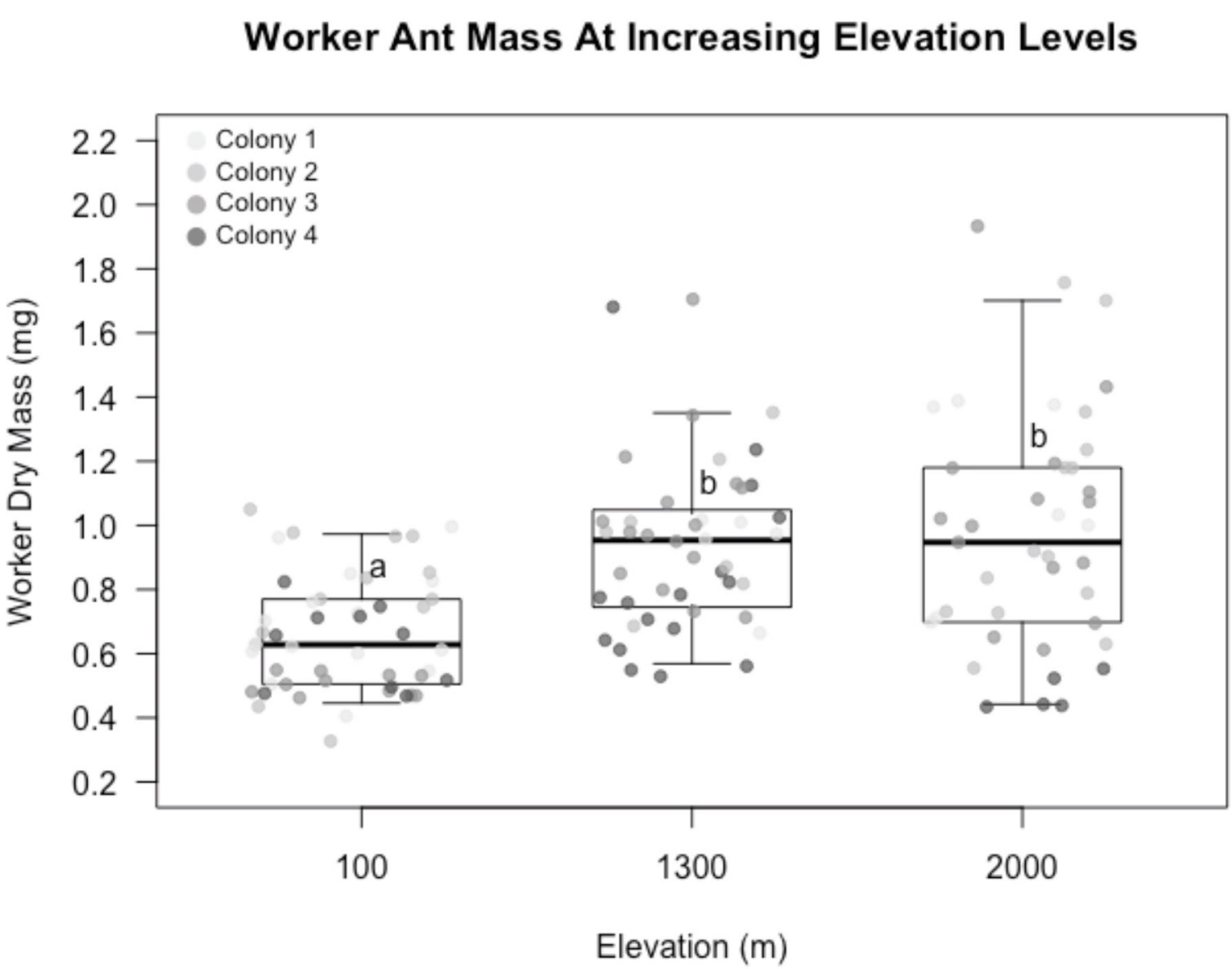

Hemoglobin Level
g/dL
6
8
10
12
14
16
18
Treatment
Mild Infection
Severe Infection

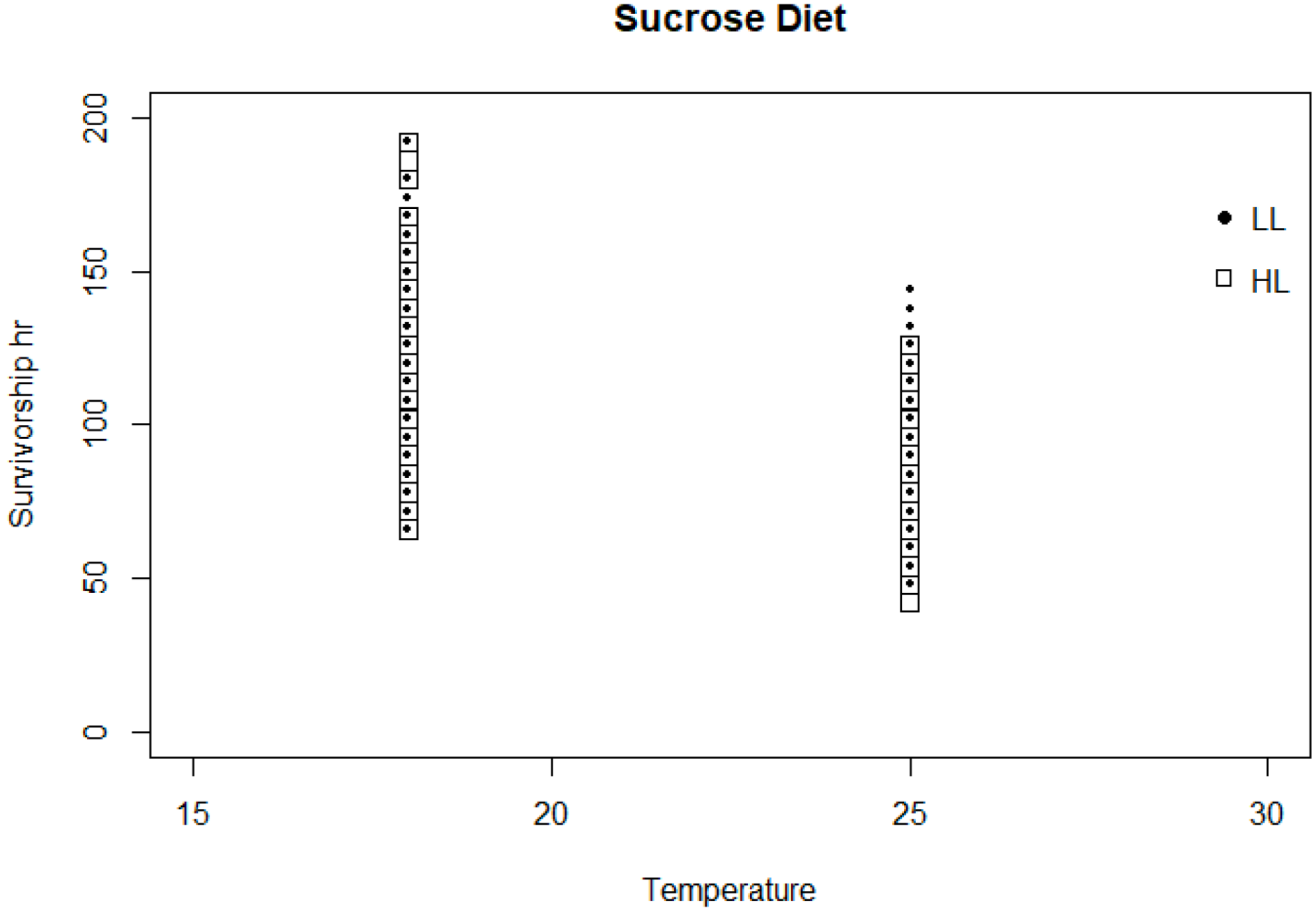
B.

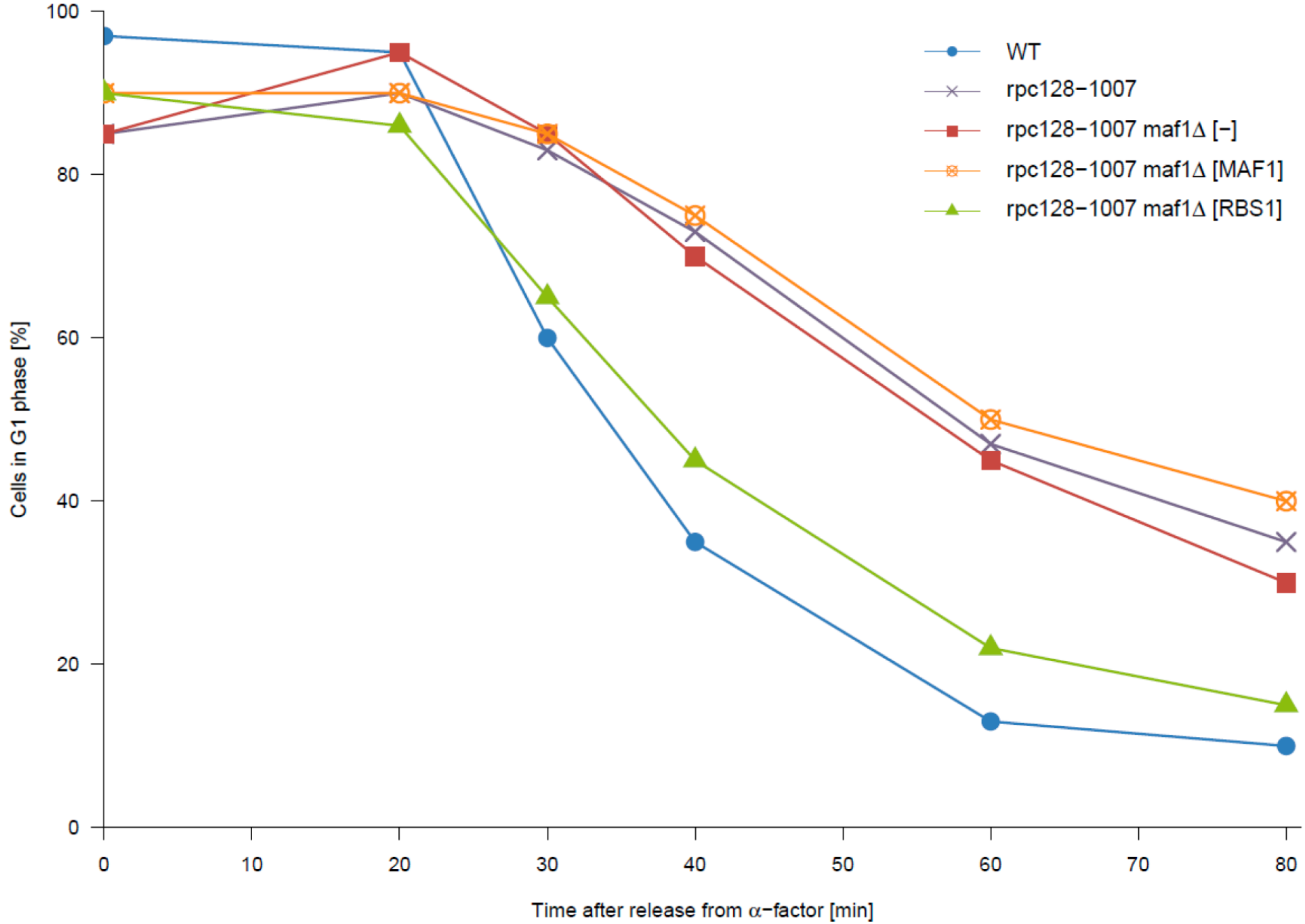

Biomass of Ancestral VS. Evolved bacteria
Biomass (mg/L^-1)
11
12
13
14
15
16
17
Ancestral
Evolved
E.coli type

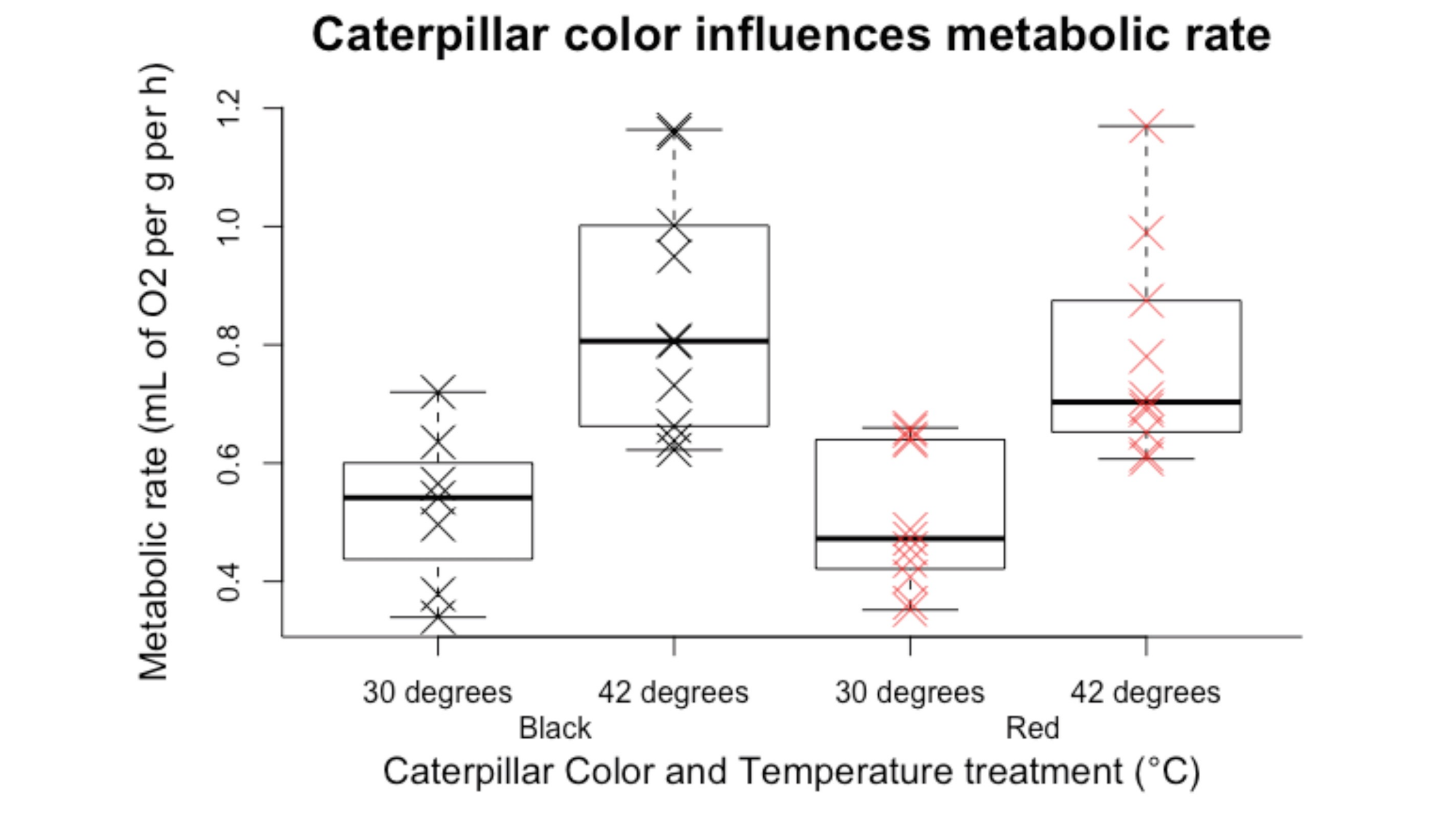
Echidna Progesterone Level VS. Days Before Egg Laying

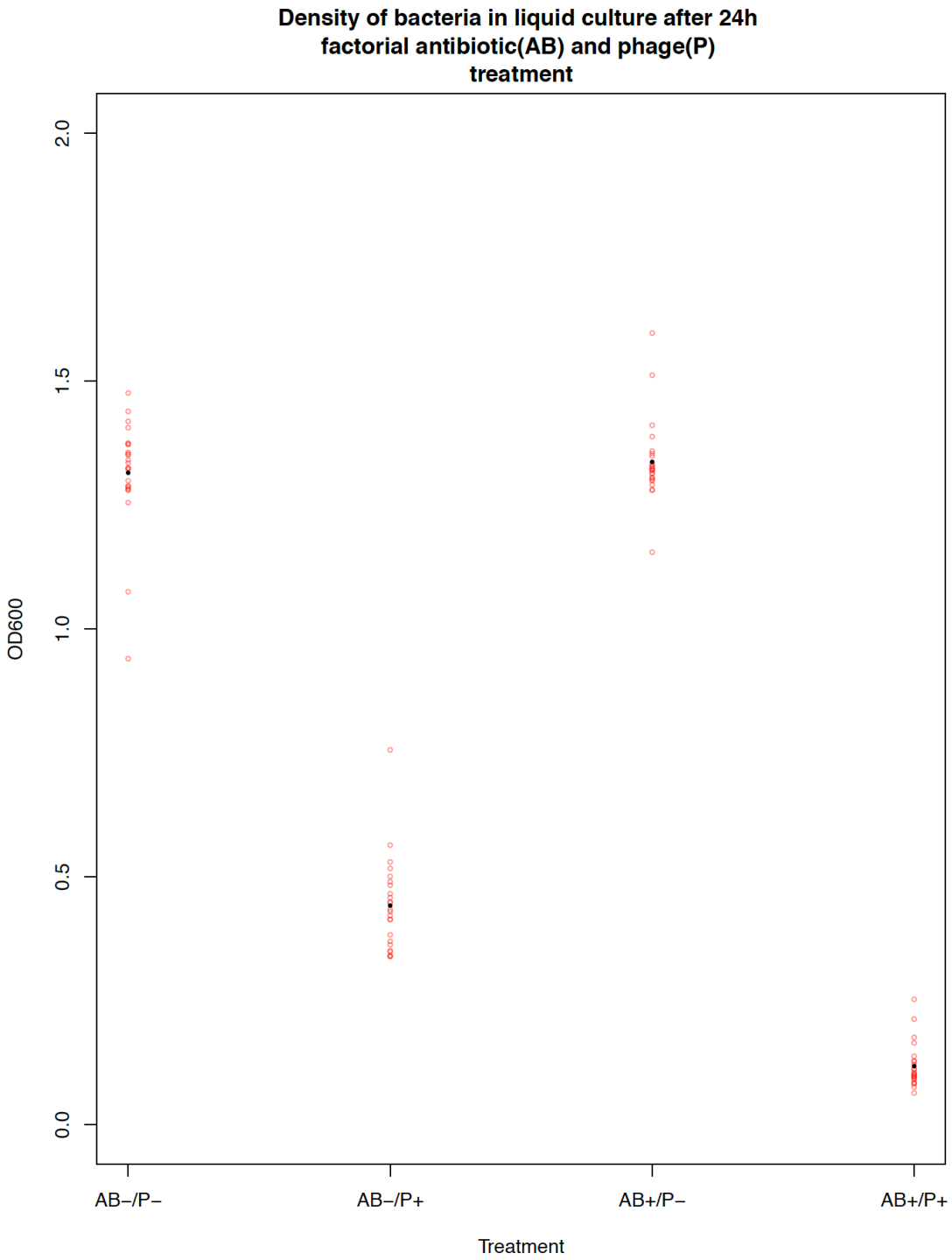


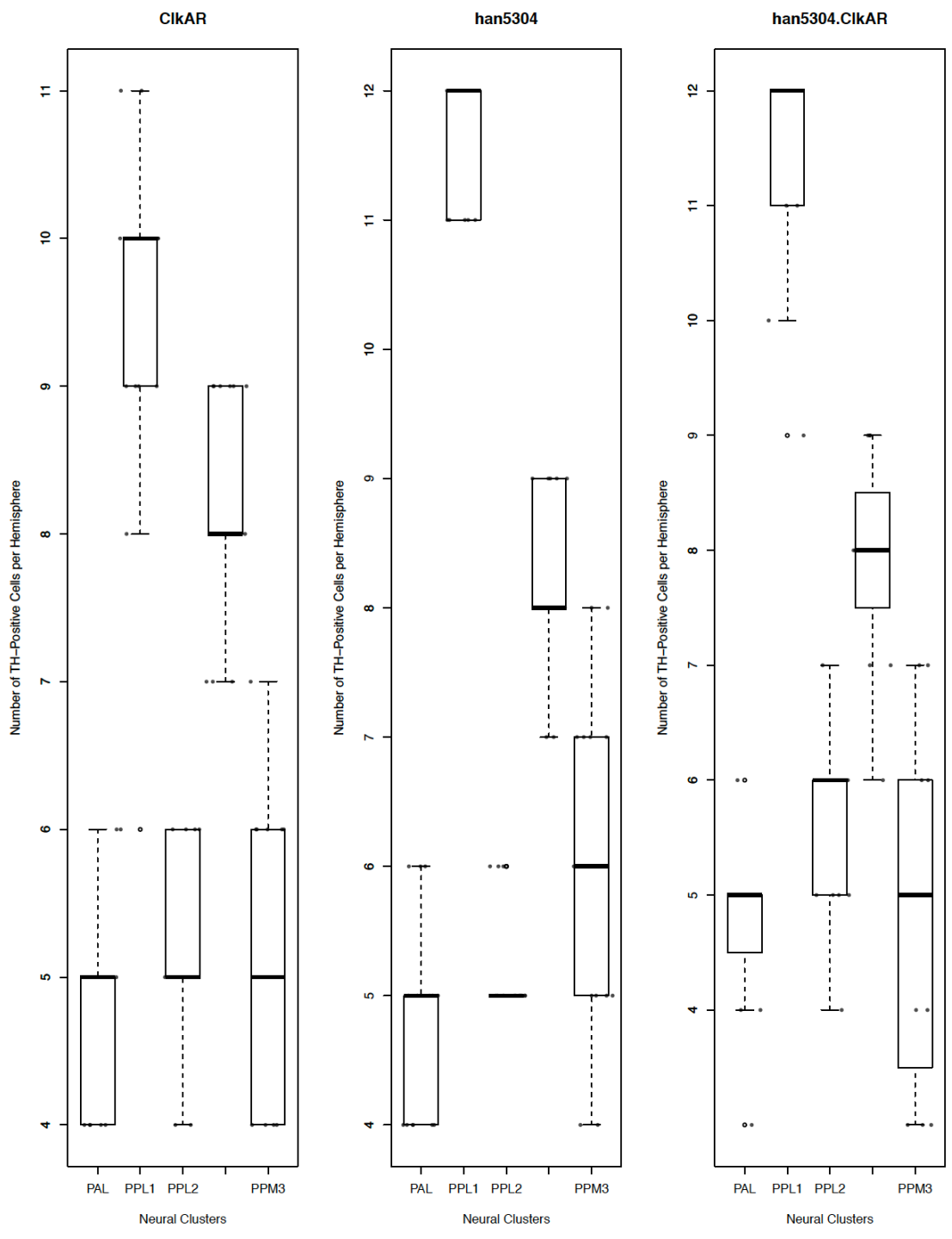

Worker Ant Mass At Increasing Elevation Levels
Colony 1
Colony 2
Colony 3
Colony 4
Worker Dry Mass (mg)
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
a
b
b
100
1300
2000
Elevation (m)

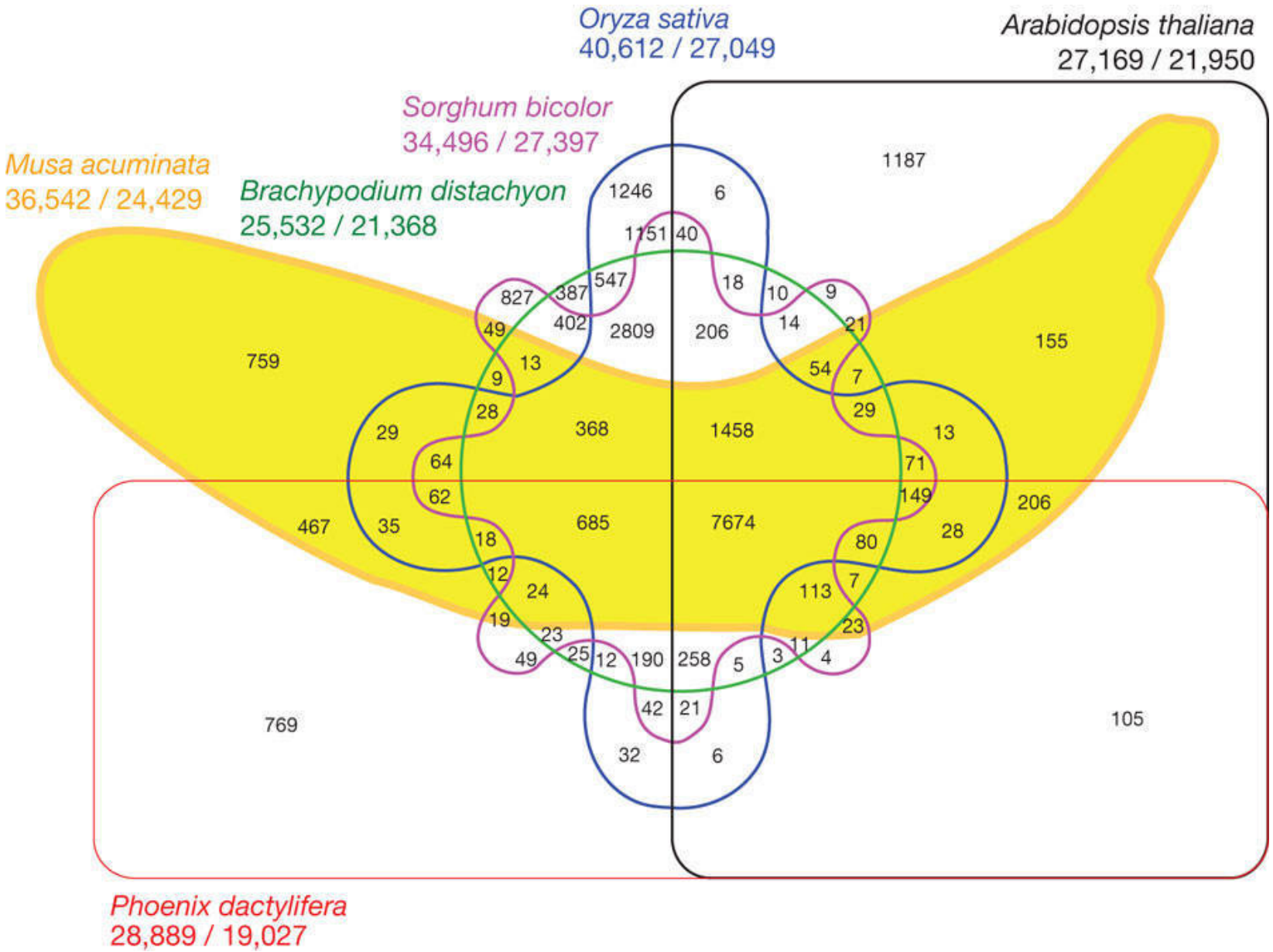

Sucrose Diet
Survivorship hr
0
50
100
150
200
15
20
25
30
Temperature
LL
HL


WT
rpc128−1007
rpc128−1007 maf1Δ [−]
rpc128−1007 maf1Δ [MAF1]
rpc128−1007 maf1Δ [RBS1]
Cells in G1 phase [%]
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
Time after release from α−factor [min]


Caterpillar color influences metabolic rate
Metabolic rate (mL of O2 per g per h)
0.4
0.6
0.8
1.0
1.2
30 degrees
42 degrees
30 degrees
42 degrees
Black
Red
Caterpillar Color and Temperature treatment (°C)


Density of bacteria in liquid culture after 24h
factorial antibiotic(AB) and phage(P)
treatment
OD600
0.0
0.5
1.0
1.5
2.0
AB−/P−
AB−/P+
AB+/P−
AB+/P+
Treatment


ClkAR
han5304
han5304.ClkAR
Number of TH−Positive Cells per Hemisphere
PAL
PPL1
PPL2
PPM3
Neural Clusters


Oryza sativa
40,612 / 27,049
Arabidopsis thaliana
27,169 / 21,950
Sorghum bicolor
34,496 / 27,397
Musa acuminata
36,542 / 24,429
Brachypodium distachyon
25,532 / 21,368
Phoenix dactylifera
28,889 / 19,027
1187
1246
6
1151
40
827
387
547
18
10
9
49
402
2809
206
14
21
155
759
13
54
7
9
28
29
368
1458
29
13
64
71
62
149
206
467
35
685
7674
28
18
80
12
24
113
7
19
23
23
49
25
12
190
258
5
3
11
4
42
21
105
769
32
6

Use the right tool

Plotting continuous or discrete data

Matplotlib: is a python plotting library,

R: endless packages for plotting though base and ggplot2 are the most popular.

Inkscape: free vector graphics editor. Can read PDF and let you tweak things you couldn't quite fix in your script generated version.

Circos: created plots that compare genomes but can show connections between any large datasets.

Excel: it is possible but you have very limited options and control

Adobe Illustrator: Similar to inkscape particularly useful for building complex multipart figures

Image Manipulation

GIMP: Think of it as free photoshop. Steep learning curve but can do just about whatever you want

ImageJ: Similar to GIMP but with many packages for analyses of specific image types.

Photoshop: Image manipulation steep learning curve very powerful and definitely worth learning if your research includes figures of FISH results.

General Editing

Adobe acrobat: often the easiest way to tweak something small you couldn't get exactly right in your script.

Powerpoint: not elegant but often the easiest way to combine different elements into a single plot.

Bold = Free

Resources

[R graph gallery]

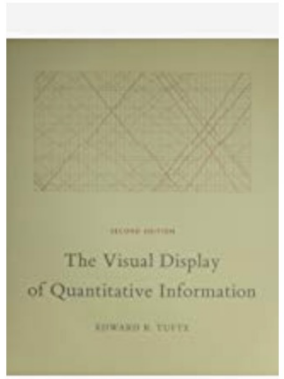
[Python graph gallery]

[Color Brewer]

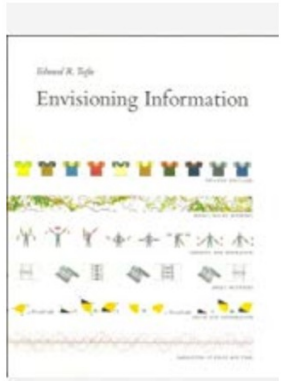
[Plotly]

[Data Visualization Catalog]

Philosophy of Data Visualization

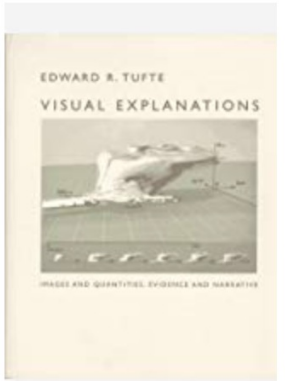


Hardcover
$34.58



Hardcover
$36.13

Hardcover
$41.60



Hardcover
$35.11

Steps in making a great figure

1) Figure out the purpose of the figure. Usually you will have a sentence in the paper or a point you want to make in a talk.

Steps in making a great figure

1) Figure out the purpose of the figure. Usually you will have a sentence in the paper or a point you want to make in a talk.

Haploid chromosome number ranged from 7-50 across sample Polyneoptera species and an XO sex chromosome system was reconstructed as the most probable ancestral state for most orders.

Steps in making a great figure

1) Figure out the purpose of the figure. Usually you will have a sentence in the paper or a point you want to make in a talk.

2) Make a list of the data that will need to be included. Is it continuous, discrete, or more complex.

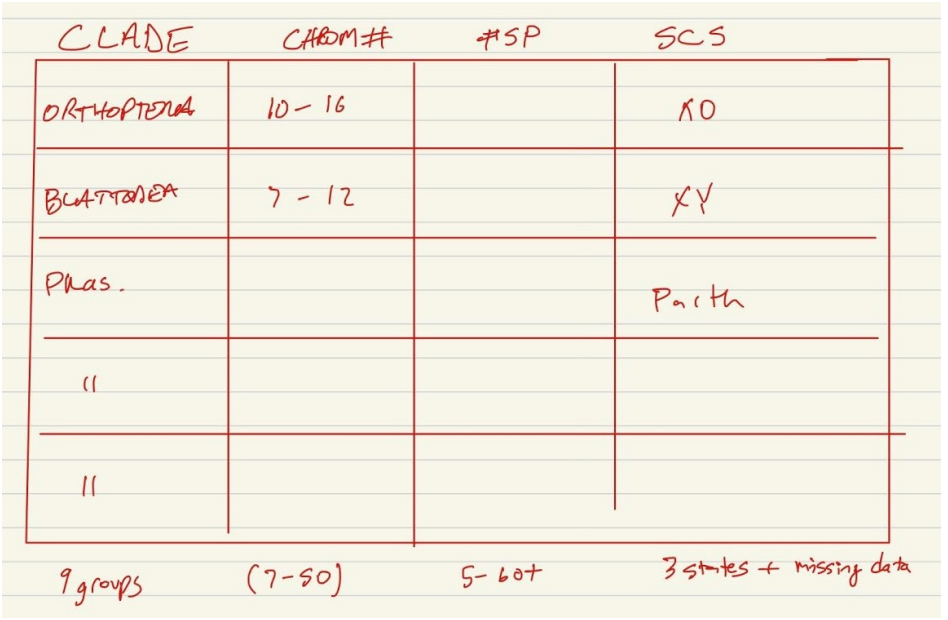
Haploid chromosome number ranged from 7-50 across sample Polyneoptera species and an XO sex chromosome system was reconstructed as the most probable ancestral state for most orders.

Steps in making a great figure

1) Figure out the purpose of the figure. Usually you will have a sentence in the paper or a point you want to make in a talk.

2) Make a list of the data that will need to be included. Is it continuous, discrete, or more complex.

Haploid chromosome number ranged from 7-50 across sample Polyneoptera species and an XO sex chromosome system was reconstructed as the most probable ancestral state for most orders.



Steps in making a great figure

1) Figure out the purpose of the figure. Usually you will have a sentence in the paper or a point you want to make in a talk.

2) Make a list of the data that will need to be included. Is it continuous, discrete, or more complex.

3) Sketch out what you think it will look like. Don't waste time figuring out how to make the perfect plot in your tool until you have settled on a best approach. Check out other papers and the graph gallery for ideas.

Haploid chromosome number ranged from 7-50 across sample Polyneoptera species and an XO sex chromosome system was reconstructed as the most probable ancestral state for most orders.

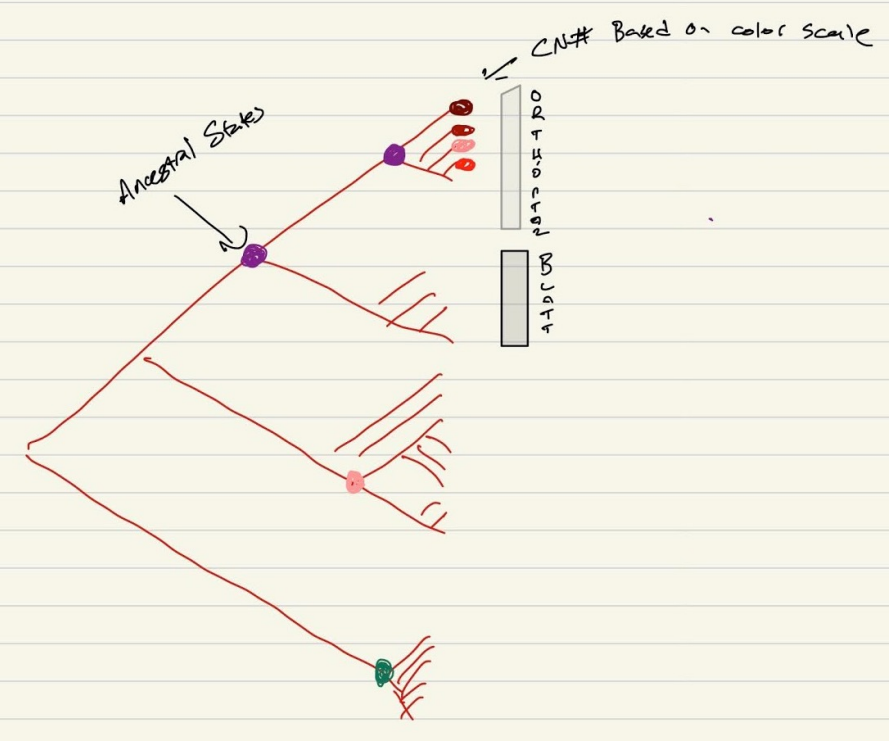
Steps in making a great figure

1) Figure out the purpose of the figure. Usually you will have a sentence in the paper or a point you want to make in a talk.

2) Make a list of the data that will need to be included. Is it continuous, discrete, or more complex.

3) Sketch out what you think it will look like. Don't waste time figuring out how to make the perfect plot in your tool until you have settled on a best approach. Check out other papers and the graph gallery for ideas.

Haploid chromosome number ranged from 7-50 across sample Polyneoptera species and an XO sex chromosome system was reconstructed as the most probable ancestral state for most orders.



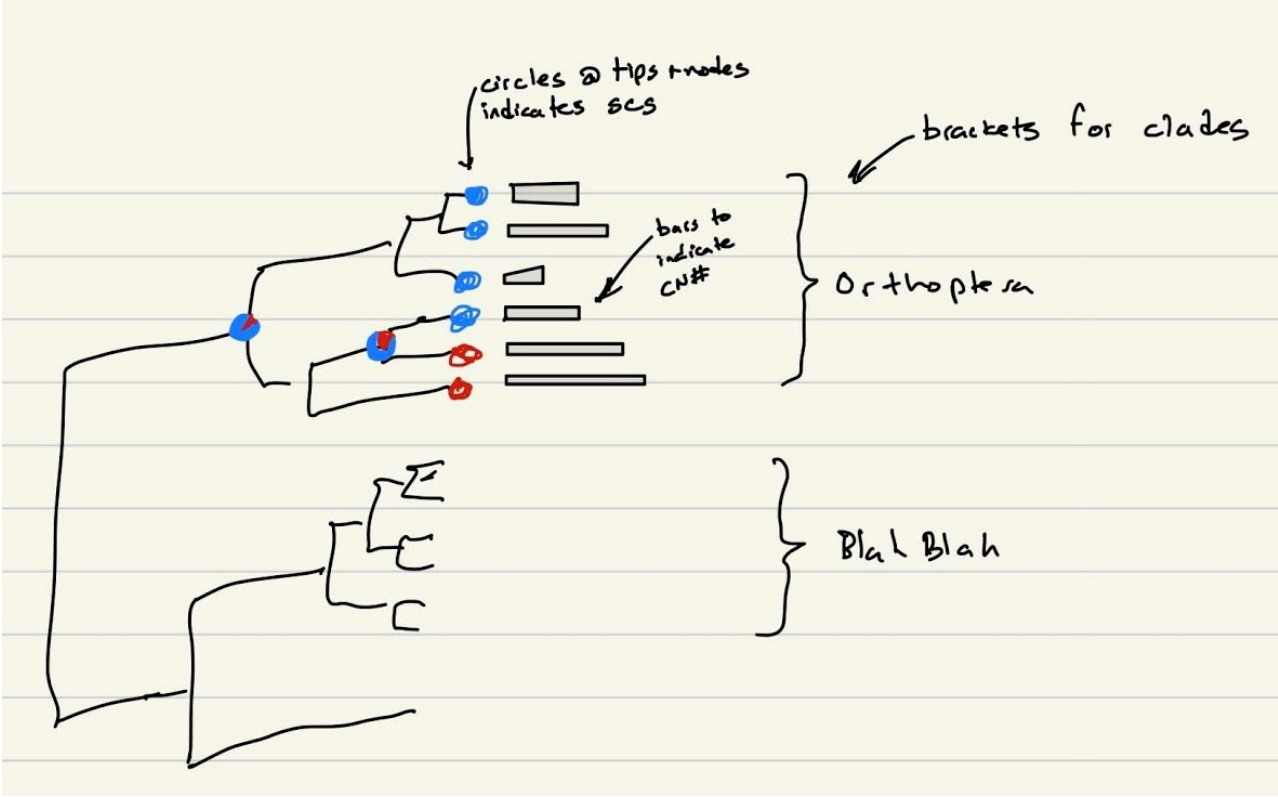
Steps in making a great figure

1) Figure out the purpose of the figure. Usually you will have a sentence in the paper or a point you want to make in a talk.

2) Make a list of the data that will need to be included. Is it continuous, discrete, or more complex.

3) Sketch out what you think it will look like. Don't waste time figuring out how to make the perfect plot in your tool until you have settled on a best approach. Check out other papers and the graph gallery for ideas.

Haploid chromosome number ranged from 7-50 across sample Polyneoptera species and an XO sex chromosome system was reconstructed as the most probable ancestral state for most orders.



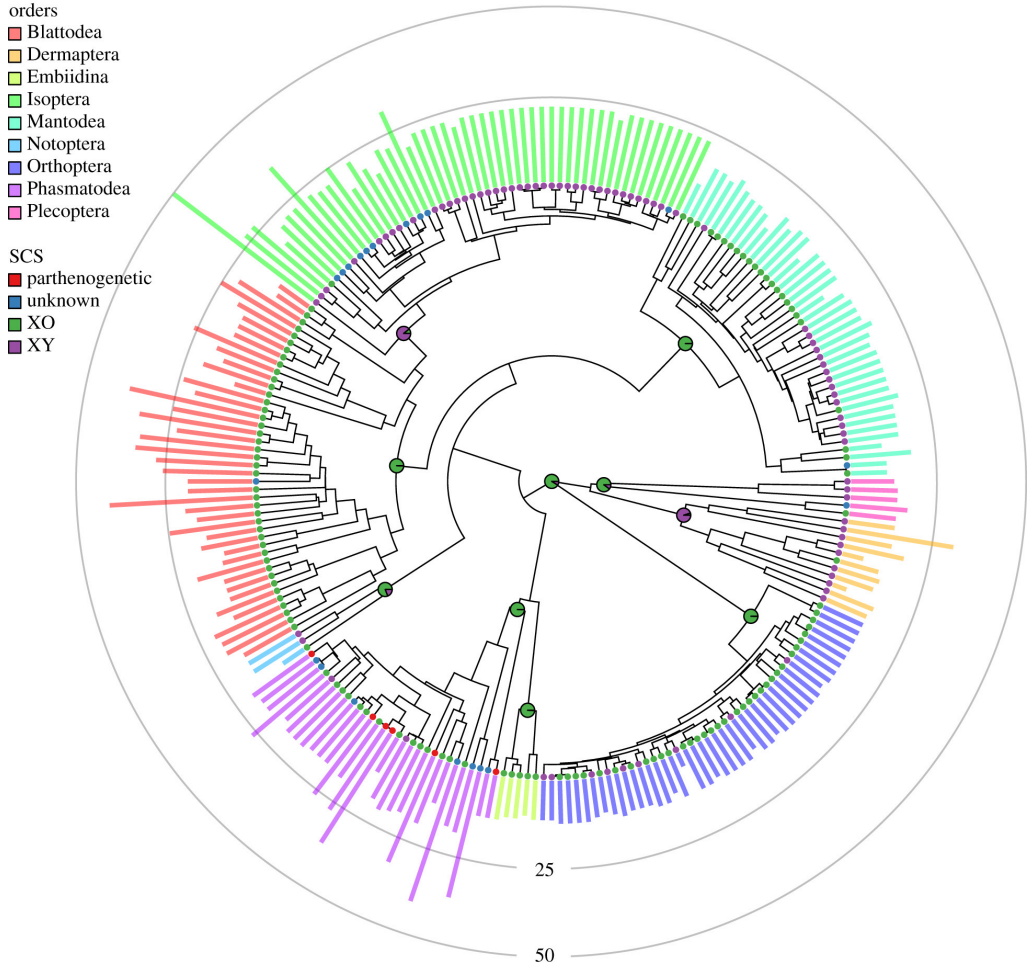
Steps in making a great figure

1) Figure out the purpose of the figure. Usually you will have a sentence in the paper or a point you want to make in a talk.

2) Make a list of the data that will need to be included. Is it continuous, discrete, or more complex.

3) Sketch out what you think it will look like. Don't waste time figuring out how to make the perfect plot in your tool until you have settled on a best approach. Check out other papers and the graph gallery for ideas.

Haploid chromosome number ranged from 7-50 across sample Polyneoptera species and an XO sex chromosome system was reconstructed as the most probable ancestral state for most orders.



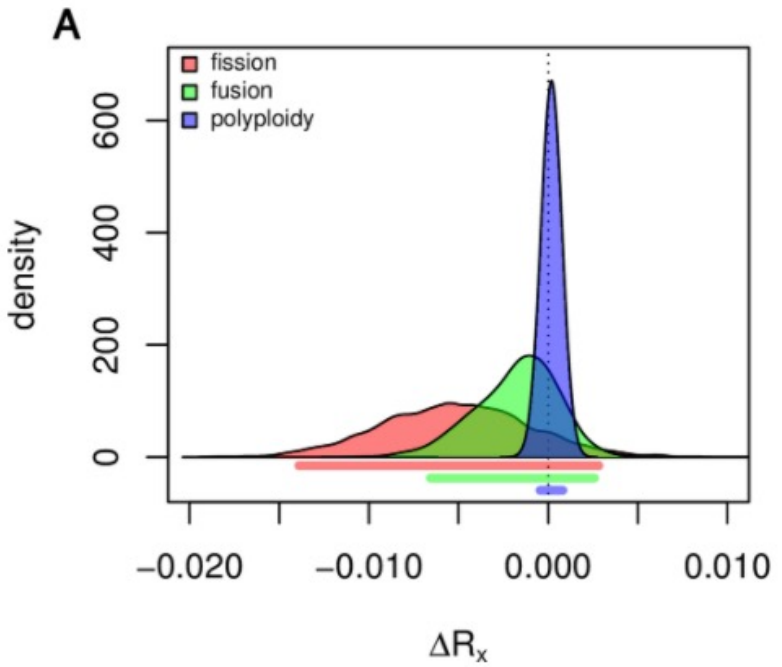
How I make a plot

1) Decide why you are making the plot. Often times you will have a sentence in your paper that demands a plot. On almost every project I've ever been involved in we have much more data than is ever included in a figure in the final paper.

The ΔR_x statistics for fusions, fissions and polyploidy had credible intervals that overlapped zero (Fig 2A).

	fission	fusion	polyploidy
1	1.924933e-02	6.691240e-03	1.194787e-02
2	8.055805e-03	-2.395435e-03	1.453424e-02
3	3.010314e-02	2.857327e-02	9.351597e-03
4	-7.035151e-03	2.745761e-02	8.345796e-03
5	5.011324e-04	-1.635201e-02	-1.280441e-02
6	2.711203e-02	-2.817951e-02	5.322630e-02
7	-4.188582e-03	2.202932e-02	-1.179885e-04
8	3.483755e-02	1.055559e-02	-1.843438e-02
9	3.550188e-02	-3.272404e-04	5.312943e-02
10	-1.145644e-02	2.068717e-03	2.317548e-02
11	1.620784e-02	1.637362e-02	1.869329e-03
12	-4.366205e-02	2.012690e-02	-2.899442e-02
13	3.398371e-03	2.127199e-03	-5.653567e-03
14	2.778317e-02	1.356907e-02	4.306500e-03
15	-6.903131e-03	5.692979e-02	3.769614e-03
16	2.218237e-02	8.972304e-05	1.469798e-02

10,000 rows



Each curve represents the posterior distribution of the ΔR_x statistic, where *x* is either fission, fusion, or polyploidy which is indicated by the color of the fill. Below the curves the lines indicate the 95% credible interval of each statistic.

Plotting in this class

Over the course of the semester we will plot using several approaches (base, ggplot, custom packages)

Our next topic is intro to R and then we will come back to plotting once you have some tools in your toolbox.

1. Show the data
2. Avoid distorting data
3. Avoid chart junk
4. Maximize data:ink ratio maximize information
5. Make it accessible to all (5% color blind + BW prints).
6. Axes and legends that are informative and useful
7. Use vector art when possible

I will evaluate all the plots you make this semester on these criteria. Can you break these? Yes but only for really good reasons!