Assignment 1: Improve a Graph

Penny Kahn

I have chosen a graph from a paper I found on Patrick Keeling's website. There are over 70 authors, so I won't provide a complete citation, but here is an abbreviated one:

Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, et al. (2012) Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* 492: 59Ð65.

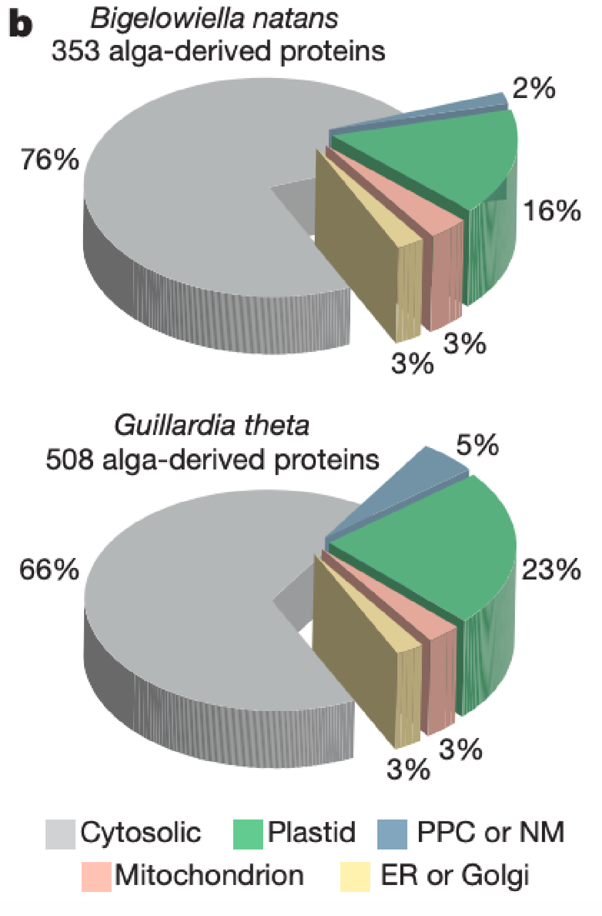
Link to paper on nature.com

Describe the paper

This paper discusses two protists: *Bigellowiella natans* and *Guillardia theta*, both unicellular eukaryotes who contain a secondary plastid endosymbiont. These plastid endosymbionts are unique in that they retain a relict nucleus (termed the nucleomorph). Therefore each of these uni-cells contains four distinct genomes within the plasma membrane (host nucleus, host mitochondria, symbiont nucleus, and symbiont plastid). The goal of the study was to describe and explain the biological significance of the genomic and cellular complexity in these two protists, as well as to document the number and kinds of genes from each of the four origin genomes within the cell.

The purpose of my figure of interest (3b) was to show the functional locations of the proteins encoded by secondary endosymbiont genes. The takehome message is that many of the proteins whose genetic origin is with the secondary endosymbiont function in the cytosol and other areas of the host - showing that during the course of host-endosymbiont integration proteins often acquire new functions and/or new locations in which to function.

Analyze the bad graph



The authors chose to use 3D pie graphs to show this data. In my opinion pie graphs are an ineffective way to show relative abundances because the human brain is not as good at understanding triangular area as it is with relative lengths of bars. Additionally, they have added an unnecessary third dimension to this graph, confusing the eye even more. It's very difficult to compare the distributions of the two graphs because they just look too similar at a glance. Moreover, the graph only presents percents without showing the raw counts. I think a well designed bar graph would be more effective at presenting this data, so that is the direction I have chosen to go

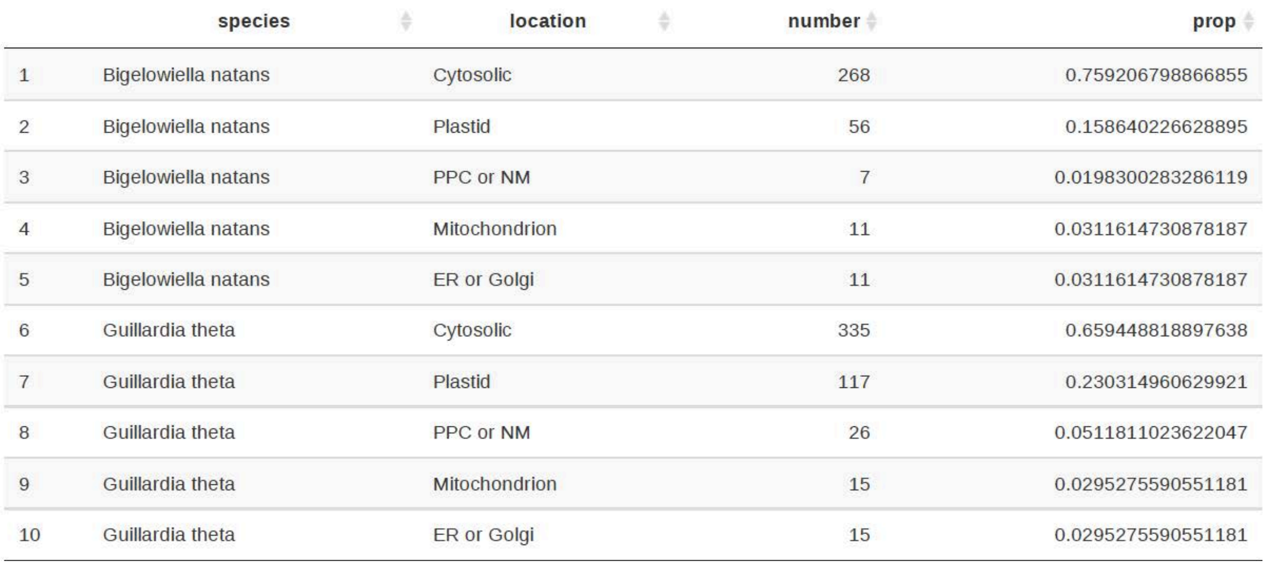
Preparing the data

```
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(wesanderson))
suppressPackageStartupMessages(library(ggthemes))
suppressPackageStartupMessages(library(grid))
suppressPackageStartupMessages(library(DT))
```

```
pie<-read.csv("501_graph_attempt2.csv")
```

Here is a table with the raw counts of proteins by location for each species. I have also added a column for location proportions within each species:

```
pie %>%
  select(species, location, number) %>%
  group_by(species) %>%
  mutate(prop = number/sum(number)) %>%
  datatable()
```

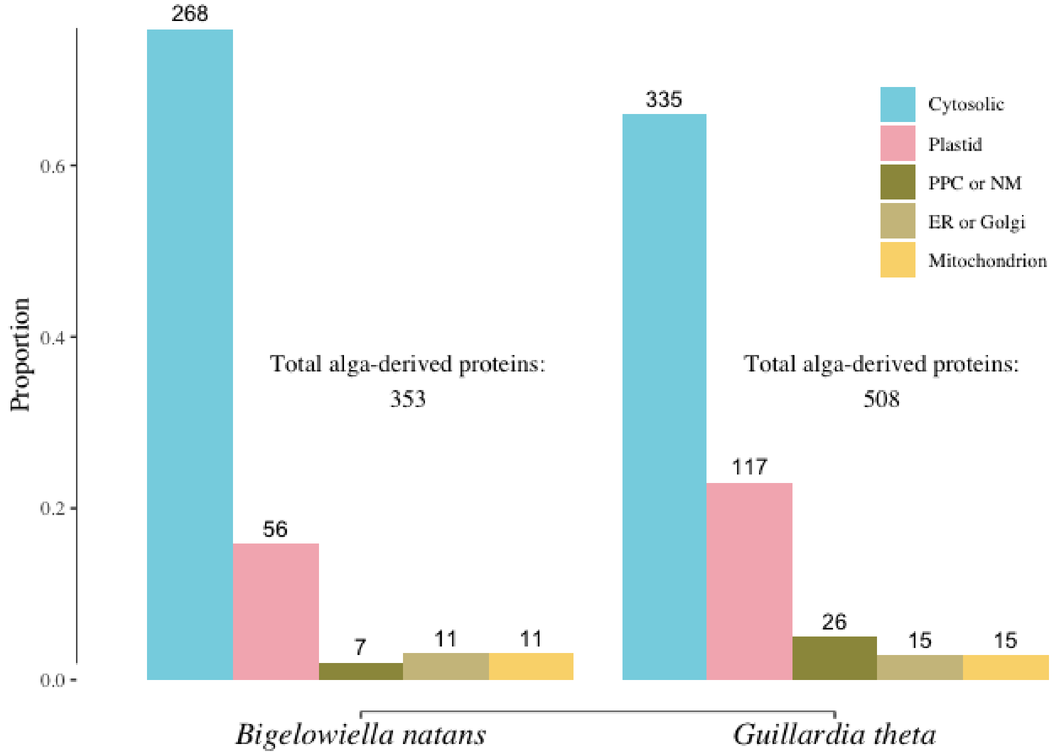


	species	location	number	prop
1	Bigelowiella natans	Cytosolic	268	0.759206798866855
2	Bigelowiella natans	Plastid	56	0.158640226628895
3	Bigelowiella natans	PPC or NM	7	0.0198300283286119
4	Bigelowiella natans	Mitochondrion	11	0.0311614730878187
5	Bigelowiella natans	ER or Golgi	11	0.0311614730878187
6	Guillardia theta	Cytosolic	335	0.659448818897638
7	Guillardia theta	Plastid	117	0.230314960629921
8	Guillardia theta	PPC or NM	26	0.0511811023622047
9	Guillardia theta	Mitochondrion	15	0.0295275590551181
10	Guillardia theta	ER or Golgi	15	0.0295275590551181

Analyze the good graph

Here is the code for producing my bar graph:

```
pie %>%
  select(species, location, number) %>%
  group_by(species) %>%
  mutate(prop = number/sum(number)) %>%
  ggplot(aes(x=species, y=prop, fill=reorder(location, -prop)))+
    geom_col(stat="identity", position=position_dodge())+
    geom_text(aes(label=number), vjust=-.5, position = position_dodge(0.9), size=3.5)+
    labs(x="", y="Proportion")+
    scale_fill_manual(values=wes_palette(n=5, name="Moonrise3"))+
    geom_rangeframe()+
    theme_tufte()+
    theme(legend.position = c(0.85, 0.75), legend.title = element_blank(), axis.text.x=element_text(size=14, face="italic", color="black"), axis.title = element_text(size=12))+
    annotate("text", x = 1.1, y = 0.35,
              label = "Total alga-derived proteins:\n353",
              family = "serif")+
    annotate("text", x = 2.1, y = 0.35,
              label = "Total alga-derived proteins:\n508",
              family = "serif")
```



Seeing the distributions side by side in this way allows for much easier comparisons of proportions between the two species. A pattern is still evident that a cytosolic destination is by far the most common, but you can now also see the relative amounts of the other locations. I also present the raw counts above the bars, which makes the amount of data more transparent than percents alone. My color scheme is also easier for people with red–green and monochromacy color–blindness to detect. Finally there is less perceptual distortion because I have not added extra unnecessary dimensions to the graph.